Estimativa da Incerteza de um Modelo Digital de Elevação por Meio de uma Análise de Agrupamento

Laércio Massaru Namikawa

Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
laercio@dpi.inpe.br

Abstract. This paper presents a method to define the distribution of clusters of statistically high values of uncertainty in a DEM using Cluster Analysis. The method requires that more than one DEM be available for the area. If there are more than one value for the same position, a statistical analysis can be executed and used to create a map with clusters of high and low uncertainty in elevation values. The main applications of the generated map are the ones where simulations are executed and different scenarios are compared to support decision-making. The simulation results can be analyzed with the uncertainty map to indicate if the results are reliable or if a better DEM should be used. In addition, the uncertainty map may indicate that the critical areas are related to areas of low uncertainty; therefore, even if the map quality is not good overall, it is good in the critical areas. The study case is the DEM of a region in Sao Paulo State in Brazil, with heterogeneous terrain features. The uncertainty maps is created using SRTM, ASTER G-DEM and IBGE elevation data.

Palavras-chave: Modelo Digital de Elevação, Incerteza, Análise por Agrupamento, SRTM, ASTER-GDEM.

1. Introdução

Modelos de processos do meio ambiente são úteis para diversas finalidades, como para a criação de cenários e para preencher lacunas nos dados. Estes modelos utilizam informações sobre o meio ambiente para definir os valores das variáveis do modelo. No entanto, qualquer informação tem uma componente de incerteza, ou seja, qualquer valor de uma variável é composto por um valor verdadeiro e um erro de medição. Portanto, os resultados do modelo são afetados pela incerteza nos dados de entrada. Os efeitos da incerteza nos modelos de processos ambientais podem gerar resultados não-confiáveis e a tomada de decisão com base nestes resultados podem ser falhos. Por exemplo, em modelos utilizados para o mapeamento de áreas de risco à inundação, uma determinada região pode ser mapeada como não tendo nenhum ou com um baixo risco para o perigo mapeado, enquanto na realidade a área é altamente vulnerável.

Se informações da incerteza do dado estiver disponível, então o modelo pode gerar informações sobre a confiabilidade do resultado. Com a informação espacial de incerteza, a confiabilidade do mapa em cada local da saída do modelo pode ser utilizada para melhorar o resultado. Para o exemplo de mapeamento de áreas de risco, a indicação de que uma área de alta vulnerabilidade tem incerteza mais elevada do que o restante da região pode ser utilizada para direcionar a aquisição de dados adicionais com uma melhor exatidão. Desta forma, a informação sobre a vulnerabilidade naquele local poderá ser melhorada.

Portanto, deve-se criar um mapa de incertezas espacialmente distribuídas para ser utilizado por modelos de processos ambientais. Como os dados de elevação são usado em muitos desses modelos uma vez que a força gravitacional é determinante em muitos processos físicos, o mapa de incerteza a ser criado é sobre a variação espacial da incerteza em dados de elevação representados em um Modelo Digital de Elevação (MDE).

Tradicionalmente, a exatidão de um conjunto de dados é definida através de amostras posicionadas em poucos pontos, idealmente distribuídos aleatoriamente, onde são obtidos valores com uma exatidão maior do que o conjunto de dados a ser analisado. A exatidão é calculada baseada na diferença entre os valores das amostras e os valores no conjunto de dados para as mesmas localizações das amostras. A principal desvantagem desta abordagem é que um mapa de incertezas não pode ser criado, uma vez que a exatidão do conjunto de dados

será um valor global. A abordagem utilizada aqui é usar dados públicos disponíveis para definir o mapa de incertezas de um determinado conjunto de dados. Como resultado desta metodologia, as regiões do mapa são classificadas como pertencendo a regiões com maior probabilidade de serem de baixa exatidão. A delimitação das regiões usa estatística espacial para encontra as regiões com agrupamento de valores com alta incerteza.

Os modelos de processos ambientais poderão gerar cenários mais confiáveis utilizando a informação de exatidão espacialmente distribuída ao invés da informação de exatidão global. Logo, objetivo deste trabalho é descrever uma metodologia para gerar um mapa de exatidão com o suporte de dados disponíveis gratuitamente e com o uso de análises de agrupamento para definir as regiões em que os valores de exatidão são estatisticamente menores do que no resto da mapa. Neste trabalho, o estudo de caso é sobre dados de elevação, usando dados do Shuttle Radar Topographic Mission (SRTM) e dados gerados a partir das imagens do satélite ASTER, conhecidos como Global DEM (G-DEM). A distribuição espacial da exatidão no estudo de caso será para os dados de elevação extraídos de cartas topográficas na escala 1:50000 de São José dos Campos, Brasil.

2. Incertezas em Modelos Ambientais

Modelos de processos ambientais são modelos dos processos físicos (não considera-se os humanos neste trabalho) relacionados ao meio ambiente da Terra que ocorrem em escalas geográficas. Estes modelos requerem dados sobre a distribuição espacial das variáveis geográficas. A exatidão do modelo de saída está relacionada com a exatidão da lógica ou das equações do modelo e da exatidão dos dados de entrada. A incerteza é uma característica que se sabe existir para todos os dados geográficos. Portanto, deve-se procura utilizar os dados mais exatos possíveis no modelo. No entanto, a informação sobre a exatidão não está disponível facilmente ou quando está disponível, não está num formato que seja sobre a sua distribuição espacial. Neste trabalho, os dados sobre a elevação do terreno são utilizados como exemplos de dados de entrada dos modelos de processos ambientais. Os Modelos Digitais de Elevação (MDE) contêm incertezas (Hunter e Goodchild, 1997; Canters et al. 2002), uma vez que as fontes de dados de elevação são medidas realizadas in-situ por meio de observações diretas ou realizadas por meio de sensoriamento remoto.

Um MDE representa a distribuição espacial da elevação, que é um valor numérico e, portanto, valores de exatidão podem ser facilmente adicionados. Infelizmente, os padrões atuais de qualidade para MDEs não definem um requisito para a exatidão espacialmente distribuída e exigem apenas um valor global. Por exemplo, a norma do United States Geological Survey (USGS) requer que a exatidão desejada para o MDE Nível-1 seja melhor do que os 7 metros calculados com erro quadrático médio (Root Mean Squared Error - RMSE) em alguns poucos locais (USGS, 2003). Os dados do Instituto Brasileiro de geografia e Estatística (IBGE) devem estar de acordo com os padrões da Comissão Nacional de Cartografia (CONCAR), que definem o melhor padrão como requerendo que o erro máximo seja de metade do valor do intervalo entre curvas de nível (Brasil, 1984).

3. Análise por Agrupamento

Para melhorar a informação sobre as incertezas em MDEs, este trabalho propõe que a geração de um mapa que mostra onde o MDE tem agrupamentos de áreas onde a exatidão são mais baixos do que a média. Ou seja, uma vez que um MDE deve ter um valor de exatidão declarada, que é uma média de toda a região, o mapa proposto mostra as áreas onde os valores de elevação são menos confiáveis. A análise por agrupamento proposta utiliza estatística espacial e os resultados indicam onde há grupos estatisticamente significativos de exatidão inferiores a exatidão global do MDE.

A elevação é para este estudo uma variável aleatória e, portanto, cada MDE é na realidade uma das realizações do MDE real e a função de densidade de probabilidade da elevação pode ser definida a partir de um conjunto de diferentes MDEs. Quando um valor de exatidão global é definido para um MDE, assume-se que os erros são aleatórios, ou seja, todas as posições dentro do MDE tem a mesma probabilidade de estarem dentro da exatidão global declarada. Por exemplo, se o RMSE de 90% é de 10 metros, então o valor real de elevação num local tem 90% de probabilidade de estar dentro do intervalo de 10 metros do valor indicado.

Este trabalho adiciona uma nova dimensão para a análise de incerteza por assumir que os erros não são distribuídos aleatoriamente e que devem existir regiões onde ocorrem agrupamentos de valores de baixa exatidão. Estes agrupamentos são detectados usando o método proposto por Rogerson (2001). Este método localiza os picos estatisticamente significativos em uma superfície que representa uma medida padronizada, que por sua vez foi suavizada por um kernel gaussiano. O valor crítico para localizar os picos é definido para uma dada probabilidade e os agrupamentos são definidos dentro dos picos onde os valores são superiores a este valor crítico.

3.1 Calculando a Medida Padronizada

A medida normalizada é a *z-score*, que indica o número de desvios padrão σ que um valor é diferente da média μ e é calculado para uma distribuição normal por:

$$zscore = \frac{x - \mu}{\sigma} \tag{1}$$

O *z-score* é calculado a cada posição do MDE, utilizando a média global e o desvio padrão global. A medida que é usado para fornecer o valor em cada ponto é o coeficiente de variação, que é a medida relativa da qualidade da dispersão, isto é, como o desvio-padrão relativo σ é em relação à média μ. O coeficiente de variação (*CV*) é calculado a cada posição do MDE, utilizando a média local μ_{rc} e o desvio padrão local σ_{rc}. Portanto, para uma grade retangular representando o MDE e com posições definidas em termos de coordenadas de linha *r* e coluna *c*, o coeficiente de variação (CV_{rc}) é calculado por:

$$CV_{rc} = \frac{\sigma_{rc}}{\mu_{rc}} 100\% \tag{2}$$

3.2 Suavização por Kernel Gaussiano

Os valores de *z-score* devem ser suavizados uma vez que se busca agrupamentos ao invés de localizações individuais. A escolha do valor de desvio padrão para a suavização por kernel gaussiano é baseada no valor que melhor filtre as diferenças aleatórias e reforce onde existem agrupamentos. Um desvio padrão pequeno mantém algumas variações aleatórias e valores altos vão criar agrupamentos muito extensos. Por dados representados por grades regulares, o kernel gaussiano é obtido através de uma soma ponderada dos valores nos vizinhos na grade da célula analisada. Os pesos da soma ponderada são calculados com base na distância a cada célula vizinha usando:

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

onde σ é o desvio padrão do kernel gaussiano e d_{ij} é a distância do centro da célula *i* para a célula vizinha *j*. Os pesos são aplicados para a célula *i* por:

$$y_i = \frac{\sum_j w_{ij} z_j}{\sqrt{\sum_j w_{ij}^2}} \tag{4}$$

onde y_i é o valor do *z-score* suavizado no centro da célula *i*, w_{ij} é o peso para a célula vizinha *j* e z_j é o *z-score* da célula *j*. Deve-se observar que a unidade de distância é em número de células, por exemplo, se duas células são vizinhas na mesma linha ou coluna da grade, a distância entre eles é um.

3.3 Valor Crítico

Os agrupamentos de valores altos de *z-score* estatisticamente significativos que indicam regiões onde o MDE tem exatidão menor do que a média, são definidos com base num valor crítico *M**, que é selecionado com base na probabilidade de se encontrar um valor maior do que *M** a um nível da significância *α* escolhido. O *M** é calculado por (Rogerson 2001):

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

onde *A* é o tamanho da região do MDE em unidades de tamanho de célula, *D* é a metade da soma da altura e da largura do retângulo envolvente da grade), *σ* é o desvio padrão do kernel gaussiano, e *φ* e *Φ* são a função de densidade de probabilidade e a função de distribuição cumulativo da distribuição normal, respectivamente.

O terceiro termo da equação (5) é suficientemente pequeno para ser eliminado (Rogerson, 2001), por conseguinte, a equação (5) é simplificada e aproximada por:

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

A equação (6) é válida apenas se *A* é menor do que 10000 ou se *σ* não é menor do que um (Rogerson, 2001). Se *A* é maior do que 10000, a aproximação pode ser usado somente se:

$$\sigma_t \Big/ \sqrt{A} > 0.01 \tag{7}$$

onde σ_t é a suavização total dada por $\sigma_t = \sqrt{\sigma_0^{\,2} + \sigma^2}$, com σ_0 igual a 10/9 para uma grade quadrada.

Para a maioria dos MDEs, *A* é maior do que 10000, e o *σ* deve se situar entre 1 e 4. Nestas condições, embora a restrição da equação (7) não seja satisfeita, o valor crítico calculado pela aproximação dada pela Equação (6) é apenas ligeiramente menor do que o valor calculado usando a equação completa (6) (Rogerson, 2001). Além disso, a diferença no tamanho da área

dos agrupamentos não é significativa, devido à elevada inclinação da grade suavizada nas regiões com agrupamentos.

4. Estudo de Caso

Neste trabalho, os agrupamentos de valores altos de incerteza em dados de elevação são utilizados para definir um mapa de exatidão espacialmente distribuído. O mapa resultante de exatidão é qualitativo e destaca as regiões do MDE que o usuário do dado deve verificar com mais cuidado, por exemplo, através de um novo levantamento com maior exatidão. A região no entorno de São José dos Campos, Brasil, foi selecionado devido à diversidade de características geomorfológicas, com áreas montanhosas, de várzea, escarpas e cuestas.

O dado de elevação analisado é fornecido pelo IBGE em escala de 1:50000. O mapa topográfico, identificado como sendo a folha São José dos Campos SF-23-YD-II-1, inclui um arquivo vetorial com curvas de nível que representam a altimetria, e abrange a região entre as coordenadas latitude, longitude 23° sul, 46° oeste e 23° 15' sul, 45° 45' oeste. As curvas de nível foram utilizadas para criar dois MDEs com representação por grade regular retangular com 30 metros de resolução espacial através do sistema de geoprocessamento SPRING (disponível em www.dpi.inpe.br/spring). O primeiro MDE foi criado usando o interpolador vizinho mais próximo e o segundo MDE com o uso da Rede Irregular Triangular (Triangulated Irregular Network - TIN).

Os conjuntos de dados de comparação são os MDEs originados do SRTM e os G-DEM gerados a partir de imagens de sensoriamento remoto do sensor ASTER. Uma vez que a elevação é considerada uma variável aleatória neste método, cada um destes três conjuntos de dados representa realizações independentes da função de densidade de probabilidade de elevação. Logo, a precisão do MDE não é relevante para definir os agrupamentos.

Os dados de elevação do SRTM foram gerados a partir de dados capturados por meio de interferometria com radar de abertura sintética e estão disponíveis globalmente em resolução de 3 arco-segundos de ftp://edcsgs9.cr.usgs.gov/pub/data/srtm. O padrão de exatidão declarada para os dados do SRTM é de 16 metros de erro linear vertical 90%. Os dados de elevação do SRTM para a região de estudo de caso foram recortadas a partir do arquivo original e reduzidas a uma grade retangular com 300 linhas e 300 colunas.

Os dados do G-DEM do ASTER também estão disponíveis globalmente, mas com resolução espacial de um arco-segundo, a partir de http://www.gdem.aster.ersdac.or.jp. A exatidão do G-DEM é estimada como sendo melhor do que a dos dados do SRTM; no entanto, uma vez que a elevação é extraída de pares-estéreo de imagens, a exatidão é variável e depende da qualidade dos pontos de controle utilizados para cada imagem. O arquivo original foi utilizado para criar uma grade retangular com 900 linhas por 900 colunas.

4.1 Análise por Agrupamento dos Dados de Elevação do IBGE

Os três MDEs (IBGE, SRTM e G-DEM) foram utilizados para definir a média e o desvio padrão da função de densidade de probabilidade de elevação em cada ponto da célula da grade SRTM, ou seja, a cada 3 arco-segundos. Deve-se observar que uma vez que o SRTM tem a resolução espacial mais baixa, os outros dois MDEs foram reamostrados e reprojetados usando o interpolador de vizinho mais próximo.

O coeficiente de variação da elevação do IBGE foi calculado a cada posição, usando a Equação (2). Em seguida, a média global e o desvio padrão do coeficiente de variação foram calculados de forma a obter o *z-score* usando a Equação (1).

O kernel gaussiano foi aplicado à grade com o coeficiente de variação usando dois desvios padrão, 1 e 2. O objetivo da utilização de dois diferentes fatores de suavização é o de avaliar a sua influência na detecção dos agrupamentos. Os valores para as somas ponderadas para os desvios padrão 1 e 2 foram calculados utilizando a equação (3).

O valor crítico M^* foi calculado para a dois desvios-padrão do kernel gaussiano ($\sigma = 1$ e $\sigma = 2$), o tamanho da região A igual a 90000 (tamanho da grade é de 300 linhas por 300 colunas) e o nível de significância α selecionado igual 0,05, usando a equação (6). O valor críticos de M^* são 4,692 para $\sigma = 1$ e 4,529 para $\sigma = 2$.

4.2 Análise dos Resultados

A Figura 1 mostra a análise de agrupamento do MDE do IBGE interpolado pelo método do vizinho mais próximo. A estatística *z-score* é mostrada com a codificação de cores indicando a faixa de valores.

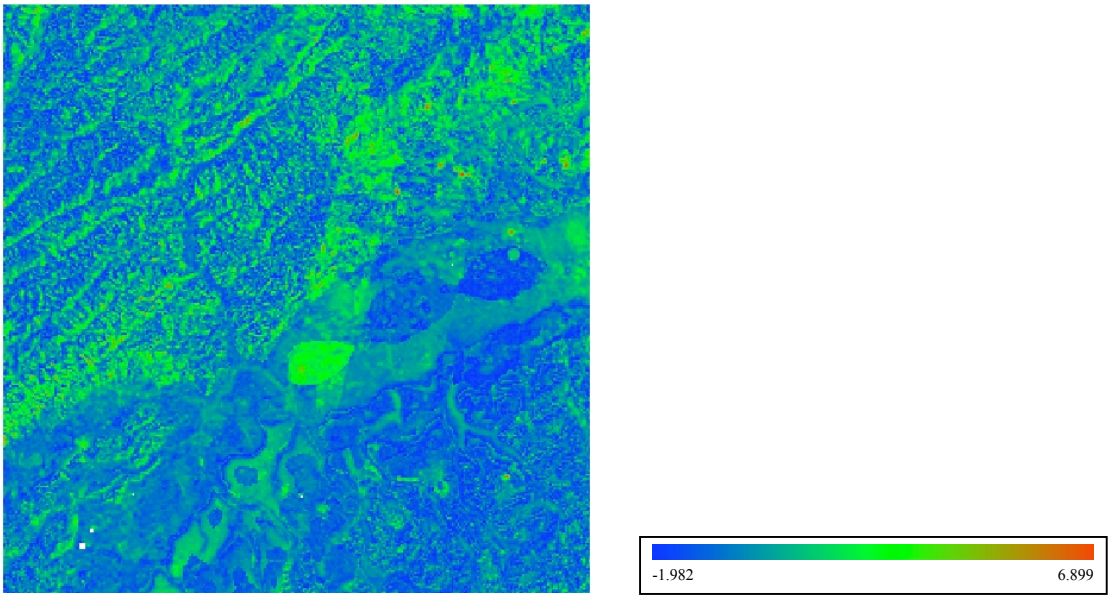

Figura 1. A análise de agrupamento do MDE do IBGE interpolado por vizinho mais próximo, apresentando a distribuição do *z-score*, com valores entre -1,982 (azul) e 6,899 (vermelho).

Usando o valor crítico M^* igual a 4,692 para um desvio padrão, as regiões de baixa exatidão significativas (para $\alpha = 0,05$) são de 1,8176 Km^2 em tamanho, e estão dentro das linhas grossas escuras na Figura 2. As linhas finas da Figura 2 indicam as regiões de baixa exatidão significativas (para $\alpha = 0,05$) para dois desvio padrão (valor crítico $M^* = 4,529$), com 14,158 Km^2 em tamanho. Pode-se observar a geomorfologia particular da região do caso de estudo indicado pelas linhas de cor-de-rosa que representam as curvas de nível do mapa topográfico do IBGE.

A análise de agrupamentos do MDE do IBGE interpolado utilizando o TIN é mostrado na Figura 3. A distribuição espacial da estatística *z-score* é mostrada na Figura 3. Utilizando o valor crítico $M^* = 4,692$ para um desvio padrão, as regiões de baixa exatidão significativas (para $\alpha = 0,05$) são de 1,710 Km^2 em tamanho, e estão dentro das linhas grossas escuras na Figura 3. As linhas finas da Figura 3 indicam as regiões de exatidão significativas (para $\alpha = 0,05$) para dois desvios padrão, com 12,227 Km^2 em tamanho. Deve-se observar que a que a diferença no tamanho entre os interpoladores não é grande, e indica que os resultados não são influenciados pelos interpoladores.

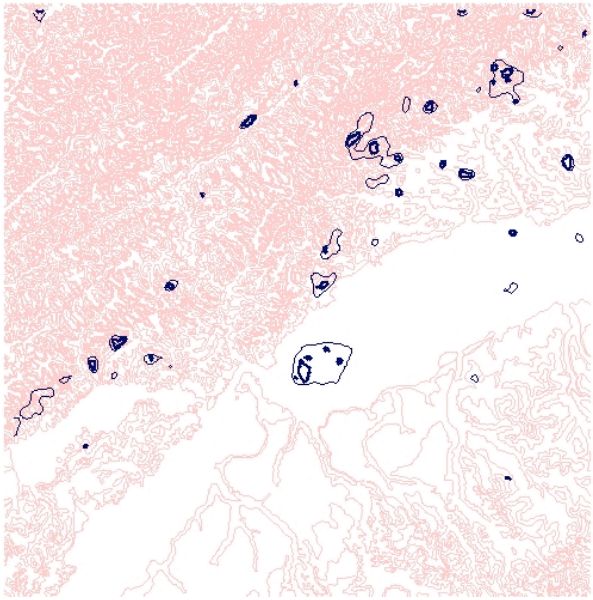

Figura 2. A análise de agrupamento do MDE do IBGE interpolado por vizinho mais próximo, apresentando as curvas de nível em cor de rosa, os agrupamentos de baixa exatidão para um desvio padrão em linhas grossas escuras, e os para dois desvio padrão em linhas finas escuras.

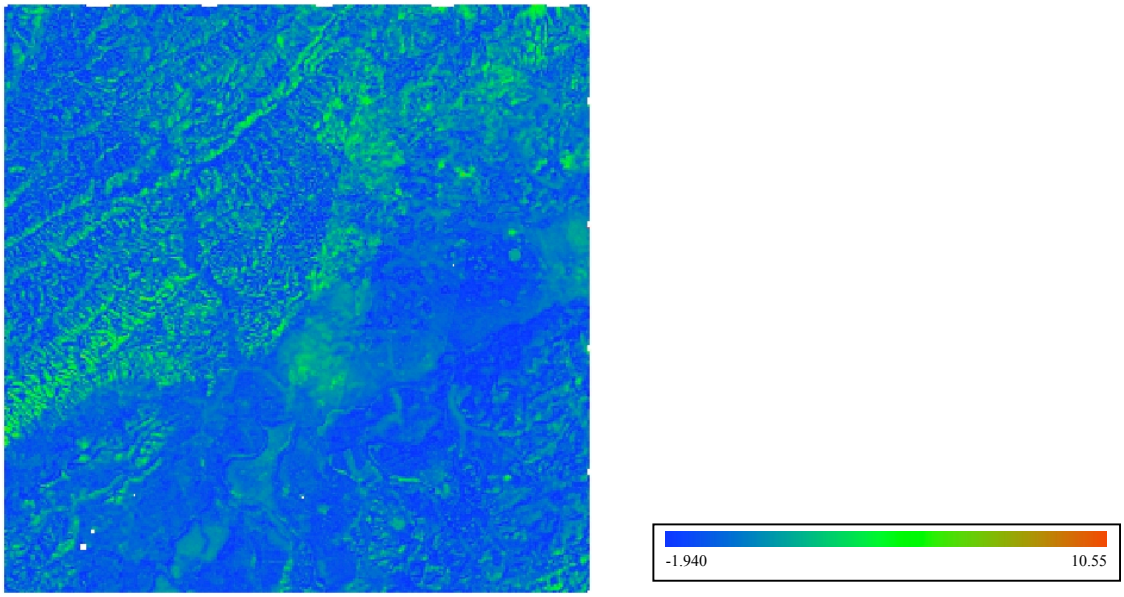

Figura 3. A análise de agrupamento do MDE do IBGE interpolado por vizinho mais próximo, apresentando a distribuição do *z-score*, com valores entre -1,940 (azul) e 10,55 (vermelho).

5. Conclusão

Este trabalho propõe um método para criar a informação de incertezas espacialmente distribuídas para dados de elevação a partir de qualquer fonte. A principal razão para criar esses tipos de mapa é para complementar a informação sobre a exatidão existentes. Tradicionalmente, a exatidão de um conjunto de dados é definida em termos de uma medida global, tal como o RMSE para os dados de elevação. Uma vez que devem existir áreas com menor exatidão na região, o mapa de incertezas espacialmente distribuídas criado pelo método aqui proposto pode ser utilizado para verificar se o MDE é adequado em aplicação do usuário ou para direcionar a coleta de novos dados para melhorar os existentes de modo que sejam obtidos onde são mais importantes.

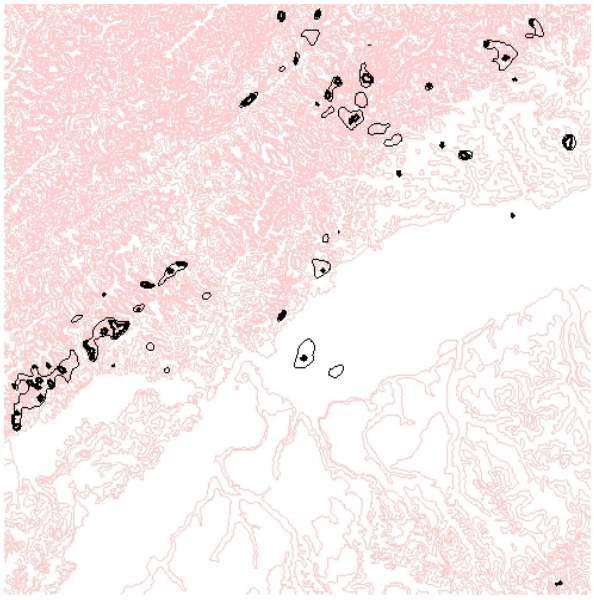

Figura 4. A análise de agrupamento do MDE do IBGE interpolado usando TIN, apresentando as curvas de nível em cor de rosa, os agrupamentos de baixa exatidão para um desvio padrão em linhas grossas escuras e os para dois desvios padrão em linhas finas escuras.

O método utiliza a análise de agrupamento de dados em uma grade regular de células proposto por Rogerson (2001). Os aglomerados de baixa exatidão são detectados em uma grade regular de valores de medidas padronizadas, como a estatística *z-score* gerada a partir do mapa do coeficiente de variação para o MDE a ser analisado. Este mapa do coeficiente de variação é criado a partir das estatísticas locais extraídas utilizando dois outros MDEs. Neste trabalho, os MDE do SRTM e do G-DEM do ASTER foram utilizados para calcular as estatísticas locais. Deve-se observar que estes dois últimos MDEs podem ser obtidos globalmente sem custo.

No estudo de caso, o método demonstrou que pode ser utilizado para criar o mapa de incerteza. Estes mapas podem ser definidas pela suavização dada pelo desvio padrão do kernel gaussiano. Por isso, se o usuário deseja ter agrupamentos menores, os valores do desvio padrão do kernel gaussiano devem ser pequenos. No estudo de caso, o tamanho das regiões diminuiu em 7 vezes quando o desvio padrão foi alterado de dois para um (de 12 para 1,7 Km^2 para os agrupamentos extraídos utilizando a interpolação por TIN). Os resultados numéricos indicam que as regiões que devem ser cuidadosamente analisadas diminuiu de 12412 Km^2 para 12 Km^2 (para o caso de interpolação por TIN com desvio padrão dois); portanto, os custos de coleta de dados adicionais podem ser mil vezes menor (se os custos forem linearmente proporcionais ao tamanho da região).

6. Referências

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.