High-accuracy surface modelling and its application to DEM generation

TIAN-XIANG YUE*†, DUN-JIANG SONG‡, ZHENG-PING DU† and WEI WANG§

†Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, 100101 Beijing, China
‡Institute of Policy and Management, Chinese Academy of Sciences, 100190 Beijing, China
§Remote Sensing Department, National Disaster Reduction Center of China, Room 1410, Zhong Min Plaza, No.7 Baiguang Road, Xuanwu District, 100053 Beijing, China

(Received 15 September 2006; in final form 25 July 2007)

An innovative method, high-accuracy surface modelling (HASM), is presented, which is based on the fundamental theorem of surfaces. The fundamental theorem of surfaces makes sure that a surface is uniquely defined by the first and second fundamental coefficients. Numerical tests of a Gaussian synthetic surface show that the Root Mean Square Error of the HASM method is much less than the ones of classical methods, such as the Triangulated Irregular Network, Spline, Inverse Distance Weight and Kriging methods. The HASM method gives a solution to the error problem that has long troubled generations of digital elevation models. All the methods for surface modelling are used to simulate a Digital Elevation Map (DEM) of Qian-Yan-Zhou Experimental Station for Red Soil and Hilly Land, which has a great topographical variety. Shaded relief maps of the simulated DEMs are developed to represent the terrain relief of Qian-Yan-Zhou Experimental Station, which shows that simulation results of the HASM method are much better than the ones of classical methods.

1. Introduction

Digital Elevation Maps (DEMs) are typically represented in two formats: contour maps, where the surface is represented by lines of constant elevation at even intervals, or point heights, where the elevation surface is sampled on either regular or irregular bases (Raaflaub and Collins 2006). DEMs are playing a more and more important role in the area of remote sensing image-based data capture, especially with the development of high-spatial-resolution satellite images (Zhu et al. 2005). A DEM has become an indispensable quantitative environmental variable in most research topics in remote sensing (San and Suezen 2005). A DEM is very commonly used for rectifying satellite images. The accuracy of rectified satellite images is highly dependent on the accuracy of the DEM. Hence, it is very important to develop highly accurate DEMs.

Earlier, DEMs were generated using stereo-pair datasets from the same sensors. Due to the high spatial resolution of recent satellite sensors in the visible spectrum, a large number of researchers around the world have investigated the extraction of elevation by means of clinometry and stereoscopy (Wolf 1974, Buchroithner 1989, Petrie et al. 1997). Elevation data have been derived from sets of Synthetic Aperture Radar (SAR) images...
Li et al. (2003) and night-time thermal infrared (NTI) images (Saraf et al. 2005). DEMs are considered to be the most permanent and reusable geo-related datasets over time and the most important data structures used for geo-spatial analysis. Unfortunately, DEMs of usable details are still not available for much of the Earth, and, when available, they are not always of sufficient accuracy (Toutin 2001).

A DEM, even one of the highest quality, is an approximation to the real-world continuous surface. DEM errors include sampling errors, the errors introduced from data-capturing equipments, errors introduced in the transformation of control points, errors from the mathematical model for constructing the surface, errors propagated from the data source, representation errors and errors caused by grid resolution and orientation (Zhou and Liu 2002). DEM errors can be propagated through the simulation process and become manifested in the final products (Huang and Lee 2005). Although there are many types and sources of error and uncertainty in geographical data and their processing, the problem is not simply technical (Unwin 1995) and it arises from an evident inability of a Geographic Information System (GIS) that was without complete theoretical foundation for surface modelling. Integration of data from different sources and in different formats, at different original scales, plus inherent errors, can yield a product of questionable accuracy. Manipulation of thematic overlays within the GIS to derive model variables are susceptible to inherent and operational errors, from which results may have such error margins as to be useless for specific applications (Burrough 1986). Any decision based on such products would thus be flawed (Walsh et al. 1987, Goodchild 1995, Unwin and Fisher 2005). To find a solution for the DEM error problem, the high-accuracy surface modelling (HASM) method is developed so that a surface is uniquely defined by the first and second fundamental coefficients.

2. HASM

2.1 Theoretical formulation of the HASM method

In terms of the fundamental theorem of surfaces, a surface is uniquely defined by the first and second fundamental coefficients (Henderson 1998). The first fundamental coefficients are used to express how the surface inherits the natural inner product of $\mathbb{R}^3$, in which $\mathbb{R}^3$ is the set of triples $(x, y, z)$ of real numbers (Carmo 2006). The first fundamental coefficients of a surface yield information about some geometric properties of the surface, by which we can calculate lengths of curves, angles of tangent vectors, areas of regions and geodesics on the surface. Those geometric properties and objects that can be determined only in terms of the first fundamental coefficients of a surface are called the intrinsic geometric properties. The collection of these geometric properties and objects forms the subject of the intrinsic geometry of a surface, which studies those properties that do not depend on the shape of the surface, but depend only on measurements that we can carry out while staying on a surface itself (Toponogov 2006). The second fundamental coefficients reflect the local warping of the surface, namely its deviation from a tangent plane at the point under consideration (Liseikin 2004).

If a surface is a graph of a function $z = f(x, y)$ or $r = (x, y, f(x, y))$, the first fundamental coefficients, $E, F$ and $G$, can be formulated as

\[
\begin{align*}
E &= 1 + f_x^2, \\
F &= f_x f_y, \\
G &= 1 + f_y^2.
\end{align*}
\]
The second fundamental coefficients, \( L, M \) and \( N \), can be formulated as

\[
\begin{align*}
L &= \frac{f_{xx}}{1+f_x^2+f_y^2}, \\
M &= \frac{f_{xy}}{1+f_x^2+f_y^2}, \\
N &= \frac{f_{yy}}{1+f_x^2+f_y^2},
\end{align*}
\]

where \( f_x \) is partial derivative of \( f(x, y) \) with respect to \( x \); \( f_y \) is partial derivative of \( f(x, y) \) with respect to \( y \); \( f_{xy} \) and \( f_{xx} \) are second-order partial derivatives of \( f(x, y) \) respectively with respect to \( x \) and \( y \); \( f_{xy} \) is second-order mixed derivative; \( E_x, F_x \) and \( G_x \) are partial derivatives of the first fundamental coefficients with respect to \( x \); \( E_y, F_y \) and \( G_y \) are partial derivatives of the first fundamental coefficients with respect to \( y \).

The Gauss equation set can be formulated as

\[
\begin{align*}
f_{xx} &= \Gamma_{11}^1 f_x + \Gamma_{11}^2 f_y + L(EG - F^2)^{-\frac{1}{2}}, \\
f_{xy} &= \Gamma_{12}^1 f_x + \Gamma_{12}^2 f_y + M(EG - F^2)^{-\frac{1}{2}}, \\
f_{yy} &= \Gamma_{22}^1 f_x + \Gamma_{22}^2 f_y + N(EG - F^2)^{-\frac{1}{2}},
\end{align*}
\]

where

\[
\begin{align*}
\Gamma_{11}^1 &= \frac{1}{2} (GE_x - 2FF_x + FE_y)(EG - F^2)^{-1}, \\
\Gamma_{12}^1 &= \frac{1}{2} (GE_y - FG_x)(EG - F^2)^{-1}, \\
\Gamma_{22}^1 &= \frac{1}{2} (2GF_y - GG_x - FG_y)(EG - F^2)^{-1}, \\
\Gamma_{11}^2 &= \frac{1}{2} (2EF_x - EE_y - FE_x)(EG - F^2)^{-1}, \\
\Gamma_{12}^2 &= \frac{1}{2} (EG_x - FE_y)(EG - F^2)^{-1}
\end{align*}
\]

and

\[
\begin{align*}
\Gamma_{22}^2 &= \frac{1}{2} (EG_y - 2FF_y + FG_x)(EG - F^2)^{-1}.
\end{align*}
\]

For the surface, \( z = f(x, y) \), in terms of numerical mathematics (Quarteroni \textit{et al.} 2000), the iterative formulation of the finite difference of the Gauss equation set can be formulated as

\[
\begin{align*}
f_{xx}^{n+1} &= \left( \Gamma_{11}^1 \right)_{i,j} f_x^n + \left( \Gamma_{11}^2 \right)_{i,j} f_y^n + \frac{L^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}}, \\
f_{xy}^{n+1} &= \left( \Gamma_{12}^1 \right)_{i,j} f_x^n + \left( \Gamma_{12}^2 \right)_{i,j} f_y^n + \frac{M^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}}, \\
f_{yy}^{n+1} &= \left( \Gamma_{22}^1 \right)_{i,j} f_x^n + \left( \Gamma_{22}^2 \right)_{i,j} f_y^n + \frac{N^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}},
\end{align*}
\]

\[
\begin{align*}
\frac{f_{xx}^{n+1} - 2f_{xx}^{n+1} + f_{xx}^{n+1}}{h^2} &= \left( \Gamma_{11}^1 \right)_{i,j} f_x^n + \left( \Gamma_{11}^2 \right)_{i,j} f_y^n + \frac{L^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}}, \\
\frac{f_{xy}^{n+1} - 2f_{xy}^{n+1} + f_{xy}^{n+1}}{h^2} &= \left( \Gamma_{12}^1 \right)_{i,j} f_x^n + \left( \Gamma_{12}^2 \right)_{i,j} f_y^n + \frac{M^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}}, \\
\frac{f_{yy}^{n+1} - 2f_{yy}^{n+1} + f_{yy}^{n+1}}{h^2} &= \left( \Gamma_{22}^1 \right)_{i,j} f_x^n + \left( \Gamma_{22}^2 \right)_{i,j} f_y^n + \frac{N^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}},
\end{align*}
\]

\[
\begin{align*}
\frac{f_{xx}^{n+1} - 2f_{xx}^{n+1} + f_{xx}^{n+1}}{4h^2} &= \left( \Gamma_{11}^1 \right)_{i,j} f_x^n + \left( \Gamma_{11}^2 \right)_{i,j} f_y^n + \frac{L^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}}, \\
\frac{f_{xy}^{n+1} - 2f_{xy}^{n+1} + f_{xy}^{n+1}}{4h^2} &= \left( \Gamma_{12}^1 \right)_{i,j} f_x^n + \left( \Gamma_{12}^2 \right)_{i,j} f_y^n + \frac{M^n_{i,j}}{\sqrt{E^n_{i,j} + G^n_{i,j}}},
\end{align*}
\]
after equation (4), an added sentence should be, “where \( f_{ij}^n \) represents the iterant of \( f(x, y) \) at lattice \((x_i, y_j)\) in the \( n \)th iterative step; \( E_{ij}^n \), \( F_{ij}^n \) and \( G_{ij}^n \) are respectively iterants of the first fundamental coefficients of \( E \), \( F \) and \( G \) at lattice \((x_i, y_j)\) in the \( n \)th iterative step; \( L_{ij}^n \), \( M_{ij}^n \) and \( N_{ij}^n \) are respectively iterants of the second fundamental coefficients of \( L \), \( M \) and \( N \) at lattice \((x_i, y_j)\) in the \( n \)th iterative step; \((\Gamma_{11})_{ij}^n \), \((\Gamma_{12})_{ij}^n \), \((\Gamma_{21})_{ij}^n \), \((\Gamma_{22})_{ij}^n \), \((\Gamma_{11})_{ij}^n \), \((\Gamma_{22})_{ij}^n \) are respectively iterants of the Christoffel symbols of the second kind, \( \Gamma_{11}^1 \), \( \Gamma_{12}^1 \), \( \Gamma_{12}^2 \), \( \Gamma_{22}^1 \) and \( \Gamma_{22}^2 \), at lattice \((x_i, y_j)\) in the \( n \)th iterative step”. Suppose that \( F^{n+1} = (f_{1,1}^{n+1}, \ldots, f_{1,N}^{n+1}, f_{2,1}^{n+1}, \ldots, f_{2,N}^{n+1}, \ldots, f_{N-1,1}^{n+1}, \ldots, f_{N-1,N}^{n+1}) \), \( n \geq 0 \); the normalized computational domain is \([0, 1] \times [0, 1] \); the simulation step length is \( h = \frac{1}{N \times T} \); and \( F^0 \) are the interpolations based on the sampling points. Then, the first equation of equation set (4) can be formulated as

\[
AF^{n+1} = D^n,
\]

where \( D^n = [D_1^n, D_2^n, \ldots, D_{N-1}^n, D_N^n]_T \) represents the vector of the right-hand term of equation (5) and \( A \) represents the coefficient matrix of equation (5).

The second equation of equation set (4) can be formulated as

\[
BF^{n+1} = E^n,
\]

where \( B \) is the coefficient matrix of equation (6) and \( E^n = [E_1^n, E_2^n, \ldots, E_{N-1}^n, E_N^n]_T \) is the vector of the right-hand term of equation (6).

The third equation of equation set (4) can be expressed as

\[
CF^{n+1} = H^n,
\]

where \( C \) is the coefficient matrix of equation (7) and \( H^n = [H_1^n, H_2^n, \ldots, H_{N-1}^n, H_N^n]_T \) is the vector of the right-hand term of equation (7).

2.2 The optimum formulation of the HASM method

The surface model based on equation (3a) of the Gauss equation set, HASM1a, can be expressed as a constraint of the least-squares approximation

\[
\begin{cases}
\min \| AF^{n+1} - D^n \|_2, \\
s.t. \quad JF^{n+1} = K,
\end{cases}
\]

where \( J \) and \( K \) are sampling points and corresponding values of \( z = f(x, y) \) at the sampling points, respectively; \( \min \) means minimize; and s.t. is the abbreviation of subject to, which means that \( \min \| AF^{n+1} - D^n \|_2 \) is subject to \( JF^{n+1} = K \). If \((x_i, y_j, f_{ij})\) is the coordinate and value of \( z = f(x, y) \) at the \( p \)th sampling point, \( J(p, (i - 1) \times N + j) = 1, K(p) = f_{ij} \).

For sufficiently large \( \lambda \), HASM1a can be formulated as an unconstrained least-squares approximation,

\[
F^{n+1} = (A^T A + \lambda^2 J^T J)^{-1} (A^T D^n + \lambda^2 J^T K).
\]
Similarly, the surface model based on equation (3b), HASM1b, can be expressed as
\[ F^{n+1} = (B^T B + \lambda^2 J^T J)^{-1}(B^T E^n + \lambda^2 J^T K). \] (10)
The surface model based on equation (3c), HASM1c, can be expressed as
\[ F^{n+1} = (C^T C + \lambda^2 J^T J)^{-1}(C^T H^n + \lambda^2 J^T K). \] (11)
The surface model based on equations (3a) and (3b), HASM2ab, can be expressed as
\[ F^{n+1} = (A^T A + B^T B + \lambda^2 J^T J)^{-1}(A^T D^n + B^T E^n + \lambda^2 J^T K). \] (12)
The surface model based on equations (3a) and (3c), HASM2ac, can be expressed as
\[ F^{n+1} = (A^T A + C^T C + \lambda^2 J^T J)^{-1}(A^T D^n + C^T H^n + \lambda^2 J^T K). \] (13)
The surface model based on equations (3b) and (3c), HASM2bc, can be expressed as
\[ F^{n+1} = (B^T B + C^T C + \lambda^2 J^T J)^{-1}(B^T E^n + C^T H^n + \lambda^2 J^T K). \] (14)
The surface model based on the three equations in equation (3), HASM3abc, can be formulated as
\[ F^{n+1} = (A^T A + B^T B + C^T C + \lambda^2 J^T J)^{-1}(A^T D^n + B^T E^n + C^T H^n + \lambda^2 J^T K). \] (15)

In order to analyse errors and computer time of the seven HASM expressions, a test surface,
\[ f(x, y) = 3 + 2[\sin(2\pi x)] [\sin(2\pi y)] + \exp[-15(x - 1)^2 - 15(y - 1)^2] + \exp[-10x^2 - 15(y - 1)^2] \]
is selected (figure 1), for which the computational domain is \([0, 1] \times [0, 1]\). The mean absolute error (MAE) and mean relative error (MRE) can be formulated as
\[ \text{MAE} = \frac{1}{IJ} \sum_j \sum_i |f_{i,j} - S_{f_{i,j}}| \] (16)
and
\[ \text{MRE} = \frac{1}{IJ} \sum_j \sum_i \left| \frac{f_{i,j} - S_{f_{i,j}}}{f_{i,j}} \right| \times 100\%, \] (17)
where \(f_{i,j}\) is the true value of \(f(x, y)\), \(S_{f_{i,j}}\) is the simulated value of \(f(x, y)\) at \((x_i, y_j)\); and \(I\) and \(J\) are the numbers of grid cells in directions \(x\) and \(y\), respectively.

The test results (table 1, figures 2 and 3) show that both HASM1a and HASM1b have almost the same absolute and relative errors and computing time. HASM1a and HASM1b coefficient-matrix structures are quite good, which means they cost less computing time. However, the errors of HASM1a and HASM1b slowly become bigger with an increase in the iteration steps and cannot reach convergence.

The coefficient matrix of HASM1c is singular, which leads to a defect of the HASMc simulation. HASM2ac and HASM2bc take the longest computing time and create overflow results because equation (3c) has a very bad structure of the coefficient matrix, and both equations (3a) and (3b) are unable to counter the impact from equation (3c). HASM3abc errors are smaller than the ones of HASM1a and HASM1b, but HASM3abc has much bigger absolute and relative errors and longer computing time cost compared with HASM2ab because HASM3abc includes equation (3c). In other
words, equation (3c) not only destroys the structure of the HASM3abc coefficient matrix, but also increases both simulation errors and computing time.

HASM2ab has a coefficient matrix structure that is easily computable and simulates a surface in both the x and y directions simultaneously. Although HASM2ab computing time is a little longer than HASM1a and HASM1b, which only simulate a surface in one direction x or y, HASM2ab errors are three orders of magnitude lower compared with the ones of HASM1a and HASM1b. HASM2ab is the optimum formulation of high-accuracy surface modelling, which is termed HASM.

Numerical tests show that seven different surface models can be developed in terms of the different combinations of the equations in the equation set (3). Errors and computing time of these surface models differ from each other; the one with the least error and lowest computing time is named HASM and can be formulated as

\[
\begin{align*}
    f_{xx} &= \Gamma_{11}^{1} f_{x} + \Gamma_{11}^{2} f_{y} + L(EG - F^2)^{-\frac{1}{2}}, \\
    f_{yy} &= \Gamma_{22}^{1} f_{x} + \Gamma_{22}^{2} f_{y} + N(EG - F^2)^{-\frac{1}{2}}.
\end{align*}
\]

(18)

2.3 The iterative formulation of HASM

The computational domain under the geographical coordinate system can be transformed into a computational domain Λ under a plane rectangular Cartesian coordinate system by Mercator’s projection, which is formulated as (Maling 1992)
<table>
<thead>
<tr>
<th>Grid number in the computational domain</th>
<th>256</th>
<th>576</th>
<th>1024</th>
<th>1600</th>
<th>2304</th>
<th>3136</th>
</tr>
</thead>
<tbody>
<tr>
<td>HASM1a Absolute error</td>
<td>0.0532</td>
<td>0.0281</td>
<td>0.017</td>
<td>0.0112</td>
<td>0.008</td>
<td>0.0059</td>
</tr>
<tr>
<td>HASM1a Relative error</td>
<td>0.0225</td>
<td>0.0118</td>
<td>0.0071</td>
<td>0.0047</td>
<td>0.0034</td>
<td>0.0025</td>
</tr>
<tr>
<td>HASM1b Absolute error</td>
<td>0.0532</td>
<td>0.0282</td>
<td>0.0169</td>
<td>0.0112</td>
<td>0.0079</td>
<td>0.0059</td>
</tr>
<tr>
<td>HASM1c Absolute error</td>
<td>defect</td>
<td>defect</td>
<td>defect</td>
<td>defect</td>
<td>defect</td>
<td>defect</td>
</tr>
<tr>
<td>HASM2ab Absolute error</td>
<td>5.79 × 10⁻⁵</td>
<td>7.22 × 10⁻⁶</td>
<td>5.32 × 10⁻⁶</td>
<td>4.36 × 10⁻⁶</td>
<td>3.69 × 10⁻⁶</td>
<td>3.24 × 10⁻⁶</td>
</tr>
<tr>
<td>HASM2ac Relative error</td>
<td>2.93 × 10⁻⁵</td>
<td>4.10 × 10⁻⁶</td>
<td>4.02 × 10⁻⁶</td>
<td>3.68 × 10⁻⁶</td>
<td>3.28 × 10⁻⁶</td>
<td>2.85 × 10⁻⁶</td>
</tr>
<tr>
<td>HASM2bc Relative error</td>
<td>overflow</td>
<td>overflow</td>
<td>overflow</td>
<td>overflow</td>
<td>overflow</td>
<td>overflow</td>
</tr>
<tr>
<td>HASM3abc Absolute error</td>
<td>0.0229</td>
<td>0.0098</td>
<td>0.0055</td>
<td>0.0034</td>
<td>0.0023</td>
<td>0.0017</td>
</tr>
<tr>
<td>HASM3abc Relative error</td>
<td>0.0169</td>
<td>0.0082</td>
<td>0.0049</td>
<td>0.0031</td>
<td>0.0022</td>
<td>0.0016</td>
</tr>
</tbody>
</table>
Figure 2. Natural logarithm of absolute errors (MLAE) of all possible surface-model formulations in the iterative process (curve marked by □ represents MLAE of HASM; curves marked by □ and * represent MLAEs of HASM1a and HASM1b respectively, curves marked by □ and △ respectively represent MLAEs of HASM2ac and HASM2bc; curve marked by △ represents MLAE of HASM3abc.

Figure 3. Computing time (PT) of all possible surface-model formulations in the iterative process (curve marked by □ represents PT of HASM; curves marked by □ and * represent PTs of HASM1a and HASM1b, respectively; curves marked by □ and △ respectively represent PTs of HASM2ac and HASM2bc; curve marked by △ represents PT of HASM3abc).
where $x$ is the abscissa of the plane Cartesian coordinate system, of which the origin is located on the equator at its intersection with the Greenwich Meridian; $y$ is the ordinate, which coincides with the Greenwich Meridian; $\phi$ is latitude; $\lambda$ is longitude; $a$ is the major semi-axis of the Earth; $e = \left(\frac{a^2 - b^2}{a^2}\right)^{1/2}$ is the first eccentricity of the Earth defined by $a$ and $b$, in which $b$ is the minor semi-axis of the Earth; and $C$ is a constant of map projection.

If the maximum lengths of the computational domain $\Lambda$ in the $x$ and $y$ directions are $L_x$ and $L_y$, respectively, and $\Lambda$ can be included in the rectangular domain $[0, L_x] \times [0, L_y]$ after a coordinate translation. If the spatial resolution of the DEM is $H \times H$, the lattice number in direction $x$ would be $M_x + 2$ and the one in direction $y$ would be $M_y + 2$, in which $M_x = \text{round}(\frac{L_x}{H}) - 2$ and $M_y = \text{round}(\frac{L_y}{H}) - 2$, where $b = \text{round}(a)$ means that $b$ is the integer closest to real number $a$. If the computational domain $[0, L_x] \times [0, L_y]$ is normalized to $\left[0, \frac{L_x}{\max(L_x, L_y)}\right] \times \left[0, \frac{L_y}{\max(L_x, L_y)}\right]$, then $\max(L_x, L_y)$ means the maximum one of $L_x$ and $L_y$, and the normalized $\Lambda$ is expressed as $\Omega$, it can be found that $\Omega \subseteq \left[0, \frac{L_x}{\max(L_x, L_y)}\right] \times \left[0, \frac{L_y}{\max(L_x, L_y)}\right]$.

If $h$ represents the simulation step length, the relationship between $h$ and spatial resolution $H \times H$ can be formulated as $h = \frac{H}{\max(L_x, L_y)}$. The central point of the lattice $(0.5h + (i - 1)h, 0.5h + (j - 1)h)$ can be expressed as $(x_i, y_j)$, in which $i = 1, 2, \ldots, M_x + 2$ and $j = 1, 2, \ldots, M_y + 2$. $f(x + h, y)$ and $f(x - h, y)$ can be formulated as the following Taylor series expansion:

$$f(x + h, y) = f(x, y) + \frac{h \partial f(x, y)}{\partial x} + \frac{h^2 \partial^2 f(x, y)}{2! \partial x^2} + \frac{h^3 \partial^3 f(x, y)}{3! \partial x^3} + O(h^4)$$

and

$$f(x - h, y) = f(x, y) - h \frac{\partial f(x, y)}{\partial x} + \frac{h^2 \partial^2 f(x, y)}{2! \partial x^2} - \frac{h^3 \partial^3 f(x, y)}{3! \partial x^3} + O(h^4).$$

Equation (20) minus equation (21) gives

$$f(x + h, y) - f(x - h, y) = 2h \frac{\partial f(x, y)}{\partial x} + \frac{2h^3 \partial^3 f(x, y)}{3! \partial x^3} + O(h^5).$$

Therefore,

$$f_x(x, y) = \frac{\partial f(x, y)}{\partial x} = \frac{f(x + h, y) - f(x - h, y)}{2h} - \frac{h^2 \partial^3 f(x, y)}{3! \partial x^3} + O(h^4).$$

For sufficiently small $h$, the finite-difference approximation of $f_x(x, y)$ can be expressed as
Similarly, for sufficiently small $h$, the finite-difference approximation of $f_x(x, y)$ can be expressed as

$$f_x(x, y) \approx \frac{f(x + h, y) - f(x - h, y)}{2h}.$$  \hspace{1cm} (24)

Equation (20) plus equation (21) gives

$$f(x + h, y) + f(x - h, y) = 2f(x, y) + \frac{2h^2 \partial^2 f(x, y)}{2!} + O(h^4).$$  \hspace{1cm} (26)

Therefore,

$$f_{xx}(x, y) = \frac{\partial^2 f(x, y)}{\partial x^2} = \frac{f(x + h, y) - 2f(x, y) + f(x - h, y)}{h^2} + O(h^2).$$  \hspace{1cm} (27)

For sufficiently small $h$, the finite-difference approximation of $f_{xx}(x, y)$ can be expressed as

$$f_{xx}(x, y) \approx \frac{f(x + h, y) - 2f(x, y) + f(x - h, y)}{h^2}.$$  \hspace{1cm} (28)

Similarly,

$$f_{yy}(x, y) \approx \frac{f(x, y + h) - 2f(x, y) + f(x, y - h)}{h^2}.$$  \hspace{1cm} (29)

If $f^n_{ij}$ ($n \geq 0$) represents the iterants of $f(x, y)$ at $(x_i, y_j)$ in the $n$th iterative step, in which $f^0_{ij} = \bar{f}_{ij}$ and $\{\bar{f}_{ij}\}$ are interpolations based on sampling values $\{\bar{f}_{ij}\}$, in terms of numerical mathematics (Quarteroni et al. 2000), the iterative formulation of the finite difference of equation set (18) can be expressed as

$$\begin{align*}
E^n_{ij} &= 1 + \left(\frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h}\right)^2, \\
F^n_{ij} &= \left(\frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h}\right)\left(\frac{f^n_{j+1,i} - f^n_{j-1,i}}{2h}\right), \\
G^n_{ij} &= 1 + \left(\frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h}\right)^2,
\end{align*}$$

$$L^n_{ij} = \frac{f^n_{i+1,j} - 2f^n_{i,j} + f^n_{i-1,j}}{h^2} \sqrt{1 + \left(\frac{f^n_{i,j+1} - f^n_{i,j-1}}{2h}\right)^2 + \left(\frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h}\right)^2},$$

$$N^n_{ij} = \frac{f^n_{i+1,j} - 2f^n_{i,j} + f^n_{i-1,j}}{h^2} \sqrt{1 + \left(\frac{f^n_{i,j+1} - f^n_{i,j-1}}{2h}\right)^2 + \left(\frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h}\right)^2},$$

$$\left\{ \begin{array}{l}
\frac{f^n_{i+1,j} - 2f^n_{i,j} + f^n_{i-1,j}}{h^2} + \frac{L^n_{ij}}{\sqrt{F^n_{ij} + G^n_{ij}}} = E^n_{ij} + \frac{f^n_{i+1,j} - f^n_{i-1,j}}{2h} + \frac{f^n_{ij}}{2h} + \frac{L^n_{ij}}{\sqrt{F^n_{ij} + G^n_{ij}}} \quad (30)
\end{array} \right.$$
High-accuracy surface modelling

\[(F_{11}^1)_{i,j}^n = \frac{G^i_{i,j}(E^n_{i+1,j} - E^n_{i-1,j}) - 2F^i_{i,j}(F^n_{i+1,j} - F^n_{i-1,j}) + F^i_{i,j}(E^n_{i,j+1} - E^n_{i,j-1})}{4(E^n_{i,j}G^n_{i,j} - (F^n_{i,j})^2)h},\]

\[(F_{22}^1)_{i,j}^n = \frac{2G^i_{i,j}(F^n_{i+1,j} - F^n_{i-1,j}) - G^i_{i,j}(G^n_{i+1,j} - G^n_{i-1,j}) - F^n_{i,j}(G^n_{i,j+1} - G^n_{i,j-1})}{4(E^n_{i,j}G^n_{i,j} - (F^n_{i,j})^2)h},\]

\[(F_{11}^2)_{i,j}^n = \frac{2E^n_{i,j}(F^n_{i+1,j} - F^n_{i-1,j}) - E^n_{i,j}(E^n_{i,j+1} - E^n_{i,j-1}) - F^n_{i,j}(E^n_{i,j+1} - E^n_{i,j-1})}{4(E^n_{i,j}G^n_{i,j} - (F^n_{i,j})^2)h},\]

and

\[(F_{22}^2)_{i,j}^n = \frac{E^n_{i,j}(G^n_{i,j+1} - G^n_{i,j-1}) - 2F^i_{i,j}(F^n_{i,j+1} - F^n_{i,j-1}) + F^i_{i,j}(G^n_{i+1,j} - G^n_{i-1,j})}{4(E^n_{i,j}G^n_{i,j} - (F^n_{i,j})^2)h}.\]

The matrix formulation of equation set (30) can be expressed as

\[
\begin{align*}
A_1 F^{n+1} &= b_1^n, \\
A_2 F^{n+1} &= b_2^n,
\end{align*}
\]

where \(F^{n+1} = \{f^{n+1}_{1,1}, f^{n+1}_{1,2}, \ldots, f^{n+1}_{1,M_x}, f^{n+1}_{2,1}, f^{n+1}_{2,2}, \ldots, f^{n+1}_{2,M_x}, \ldots, f^{n+1}_{M_x,1}, f^{n+1}_{M_x,2}, \ldots, f^{n+1}_{M_x,M_y}\}^T\), \(A_1\) and \(A_2\) represent coefficient matrices of the first equation and the second equation in equation set (30), respectively, and \(b_1^n\) and \(b_2^n\) are the right-hand vectors of equation set (30).

If the \(k\)th sampling point is located at the lattice \((x_i, y_j)\) in the computational domain, the simulation value should be equal to or approximate to the sampling value at this lattice so that a constraint equation set is added to the equation set (30). The HASM method can be formulated as

\[
\begin{align*}
\min \|A F^{n+1} - b^n\|_2, \\
\text{s.t.} \quad C F^{n+1} = d,
\end{align*}
\]

where \(C(k, (i-2)M_x+j-1) = 1; \ d(k) = \tilde{f}_{i,j}; \ A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}; \) and \(b^n = \begin{bmatrix} b_1^n \\ b_2^n \end{bmatrix}\).

For sufficiently large \(\lambda\), the HASM method can be transferred into an unconstrained least-squares approximation,

\[
\min_f \| A \frac{\lambda}{\lambda C} F^{n+1} - \begin{bmatrix} b^n \\ \lambda d \end{bmatrix} \|_2
\]

or

\[
A^T \frac{\lambda}{\lambda C} A \left[ \begin{array}{c} A \\ \lambda C \end{array} \right] F^{n+1} = A^T \frac{\lambda}{\lambda C} \begin{bmatrix} b \\ \lambda d \end{bmatrix}.
\]

The parameter \(\lambda\) is the weight of the sampling points and determines the contribution of the sampling points to the simulated surface. \(\lambda\) can be a real number, which means all sampling points have the same weight, or a sector, which means every sampling point has its own weight. The area affected by a sampling point in a complex region is smaller than in a flat region. Therefore, a smaller value of \(\lambda\) is selected in a complex region and a bigger value of \(\lambda\) is selected in a flat region.
According to the theory of differential geometry, a necessary and sufficient condition for a surface to be a plane is that all the second fundamental coefficients equal zero at all points of the surface (Somasundaram 2005). Therefore, we suppose that if $L$, $M$ and $N$ do not vanish at all points of a piece of a surface, the piece of the surface can be simulated by the HASM method. If $L = M = N = 0$ on a piece of a surface, the piece of the surface can be expressed by triangles. If the three nodes of a triangle are $(x_1, y_1, z_1)$, $(x_2, y_2, z_2)$ and $(x_3, y_3, z_3)$, the triangle can be formulated as

$$z = ax + by + c,$$

where

$$a = \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix},$$

$$b = \begin{vmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{vmatrix}$$

and

$$c = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}.$$

3. Comparative error analyses

In addition to the newly developed HASM method, there are many other methods of surface modelling that have been widely used in various GIS applications in recent years, such as the Inverse Distance Weight (IDW), Triangulated Irregular Network (TIN), Kriging and Spline methods. These classical surface models can be used to comparatively analyse HASM errors.

The IDW method uses an inverse distance weighting function to determine the interpolation value for any given point within the calculated area (Lee and Angelier 1994). The IDW method allows an individual spatial analysis of each independent variable (Julià et al. 2004), and it is a good method when all the variables have a similar weight (Sinowski et al. 1997).

The TIN method is a popular model for representing surface models in the GIS because it has a simple data structure and can easily be rendered using common graphics hardware (Yang et al. 2005). The TIN method is composed of nodes, edges, triangles, topology and a hull. It is one of the basic models for representing digital terrain, a piece of land on the Earth’s surface (Tucker et al. 2001, Tse and Gold 2004).

The Kriging method is a generalized linear regression that is used to formulate an optimal estimator in a minimum mean square-error sense (Li et al. 2004). It was introduced to avoid systematic errors in interpolation by the South African mining engineer, D.G. Krige (Kleijnen and van Beers 2004). The Kriging method is referred to as the best linear unbiased estimate because some study cases have shown that the expected value for the estimate error equals zero and whose variance...
is a minimum. The Kriging model has become a fundamental tool in geostatistics (Olea 1999).

The Spline method is a spline-based method, of which the specific types include uniform non-rational basis spline, non-uniform non-rational basis spline and non-uniform rational basis spline. A spline curve is specified by a given set of control points. The Spline method curve is defined, modified and manipulated with operations on the control points (Sarfraz and Siddiqui 2004). Control points are then fitted with piecewise continuous parametric polynomial functions. A designer can modify a local surface area by interactively moving a control point or changing its corresponding weight (Ma and He 1998). The Spline method can be used to create a solid shape from a large number of points in three-dimensional (3D) space. Most 3D CAD systems use a mathematical definition for the surface based on the Spline method (Schoonmaker 2003).

The HASM method uses existing data, as points or contours, to globally fit a surface through several iterative simulation steps. This surface is then used to interpolate a height value at an unknown point. The unknown points are located on a regularly space lattice and the final result is a gridded digital terrain model. The iterative simulation steps are summarized as follows: (1) conducting interpolation on the computational domain \( \Omega \) in terms of sampling data \( (x_i, y_j, f_{ij}) \), from which we can get interpolated approximate values \( \{ \tilde{f}_{ij} \} \) at point \( (x_i, y_j) \); (2) letting \( f^n_{ij} = \tilde{f}_{ij} \) and calculating the first fundamental coefficients \( E^n_{ij}, F^n_{ij} \) and \( G^n_{ij} \) and the second fundamental coefficients \( L^n_{ij} \) and \( N^n_{ij} \) as well as the coefficients of the HASM equations in terms of \( \{ f^n_{ij} \} \); (3) for \( n \geq 0 \), we can get \( \{ f^{n+1}_{ij} \} \) by solving the HASM equations; and (4) the iterative process is repeated until simulation accuracy is satisfied.

### 3.1 A numerical test

The Gaussian synthetic surface, \( f(x, y) = 3(1 - x)^2 e^{-(x^2 - (y + 1)^2)} - 10(\frac{2}{3} - x^2 - y^2) e^{-(x^2 - y^2)} - \frac{1}{3} e^{-(x + 1)^2 - y^2} \), is taken as the test surface (figure 4(a)), so that the ‘true’ value can be pre-determined to avoid uncertainty caused by uncontrollable data errors. Its computational domain is \([-3, 3] \times [-3, 3]\) and \(-6.6 < f(x, y) < 8.1\). In ArcMap’s Visual Basic for Application (VBA) environment, ArcObjects’s functions are used to create data in grid format with a spatial resolution on 0.08 \( \times 0.08 \) in terms of the formulation of the Gaussian synthetic surface. The sampling contour data are produced by employing the Contour command of the Spatial Analyst module.

The HASM method and the classical methods are used to simulate the test surface, in which grid spacing is selected as \( H = 0.08 \) and the sampling interval is \( 5H \). The shaded relief maps are created using the HillShade command in the Surface Analysis menu of the Spatial Analyst module. The simulation error is calculated using the Root Mean Square Error (RMSE), which is expressed as

\[
\text{RMSE} = \sqrt{\frac{1}{76^2} \sum_{i=1}^{76} \sum_{j=1}^{76} (f_{ij} - Sf_{ij})^2},
\]  

where \( f_{ij} \) is the true value of \( f(x, y) \) and \( Sf_{ij} \) is the simulated value of \( f(x, y) \) at \( (x_i, y_j) \).
The simulation results show that surfaces simulated by the classical methods have been considerably deformed, while the HASM method has a perfect simulation result (figure 4(b)). The contour map simulated by the HASM method (figure 5(b)) is almost the same as the one of the original Gaussian synthetic surface (figure 5(a)). The Spline method produced three more contour lines compared to the original one (figure 5(c)). The TIN method made one contour line become two crossing contour lines and one contour line disappear (figure 5(d)). The Kriging and IDW methods made contour lines out of shape (figures 5(e) and 5(f)). The RMSE of the HASM method in the fourth iterative step is 1.4802 times less than the one of the Spline method, 1.7717 times less than the TIN method, 5.0675 times less than the Kriging method and 7.2143 times less than the IDW method (table 2).

Figure 4. Comparison between the HASM simulation surface and surfaces simulated by classical methods: (a) the original Gaussian synthetic surface; (b) the HASM surface; (c) the Spline surface; (d) the TIN surface; (e) the Kriging surface; and (f) the IDW surface.

The simulation results show that surfaces simulated by the classical methods have been considerably deformed, while the HASM method has a perfect simulation result (figure 4(b)). The contour map simulated by the HASM method (figure 5(b)) is almost the same as the one of the original Gaussian synthetic surface (figure 5(a)). The Spline method produced three more contour lines compared to the original one (figure 5(c)). The TIN method made one contour line become two crossing contour lines and one contour line disappear (figure 5(d)). The Kriging and IDW methods made contour lines out of shape (figures 5(e) and 5(f)). The RMSE of the HASM method in the fourth iterative step is 1.4802 times less than the one of the Spline method, 1.7717 times less than the TIN method, 5.0675 times less than the Kriging method and 7.2143 times less than the IDW method (table 2).
Figure 5. Comparison between contour lines of the surface simulated by the HASM method and ones simulated by the classical methods: (a) the original contour lines of Gaussian synthetic surface; (b) HASM contour lines; (c) Spline contour lines; (d) TIN contour lines; (e) Kriging contour lines; and (f) IDW contour lines.
3.2 A test of HASM efficiency with a real surface of the Earth

Qian-Yan-Zhou Experimental Station for Red Soil and Hilly Land of the Chinese Academy of Sciences is selected as the test area. The Qian-Yan-Zhou Experimental Station is located in the middle of Jiang-Xi province, China. The geographic coordinates of its central point are 26° 44' 44" N and 115° 03' 44" E (figure 6). Its area is 2.0416 km², including about 80 hillocks and 9 gullies. Its elevation varies from 67 to 147.5 m. It includes four geomorphologic types: hillock, terrace, higher flood plain and lower flood plain. The hillocks have round summit surfaces; their relative heights vary from 20 to 50 m; and their slopes vary from 10° to 30°. The terraces have even surfaces; their relative heights are about 2 m; and their slopes vary from 1° to 2°. The surface slopes of the higher flood plains vary from 1° to 3° and relative heights are from 0.6 to 1 m. The lower flood plains are slightly undulant and their relative heights vary from 0.5 to 0.6 m. The Qian-Yan-Zhou Experimental Station has a great topographical variety and is suitable for testing the efficiency of all the methods of surface modelling.

The DEM of the Qian-Yan-Zhou Experimental Station is simulated by means of the HASM, Spline, TIN, Kriging and IDW methods, for which a scanned topographical map on a spatial scale of 1 : 5000 is vectorized. Shaded relief maps of the simulated DEMs are developed to represent the terrain relief of the Qian-Yan-Zhou Experimental Station. All the shaded relief maps are made in the same azimuth and altitude angle (figures 7 to 11). The azimuth angle of the light source is 180°, which is from south to north. The azimuth is expressed in positive degrees from 0° to 360° and measured clockwise from the north. The altitude angle of the light source above the

<table>
<thead>
<tr>
<th></th>
<th>IDW</th>
<th>Kriging</th>
<th>TIN</th>
<th>Spline</th>
<th>HASM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.3636</td>
<td>0.2554</td>
<td>0.0893</td>
<td>0.0746</td>
<td>0.0504</td>
</tr>
<tr>
<td>The ratio of HASM error to ones of the classic methods</td>
<td>7.2143</td>
<td>5.0675</td>
<td>1.7718</td>
<td>1.4802</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6. Location of Qian-Yan-Zhou Experimental Station for Red Soil and Hilly Land.
Figure 7. The DEM developed by the HASM method.

Figure 8. The DEM developed by the Spline method.
Figure 9. The DEM developed by the TIN method.

Figure 10. The DEM developed by the Kriging method.
horizon is 45°. The slope is expressed in positive degrees, with 0° at the horizon and 90° directly overhead.

The 3D shaded relief maps show that the HASM method (figure 7) has a much better simulation result than the ones of the classical methods. The Spline method has a serious oscillation problem, especially in the right bottom area (figure 8). The TIN method has an obvious peak truncation and pit-fill problem (figure 9). The IDW and Kriging methods make the Earth’s surface of the Qian-Yan-Zhou Experimental Station become very coarse and the hillocks like sand dunes (figures 10 and 11).

4. Discussion and conclusions

In this article, we have presented an innovative method for surface modelling, the HASM method. It has a much higher accuracy than the classical methods such as the TIN, Spline, IDW and Kriging methods.

TIN surfaces are constructed by triangulating a set of vertices (points), for which the vertices are connected with a series of edges to form a network of triangles. Each triangular facet describes the behaviour of a portion of a TIN surface, and two triangular facets become one rectangle facet when the vertices are evenly distributed, which would cause the so-called peak truncation and pit-fill problem (Yue et al. 2007a).

Each facet of the Spline surface is estimated by a mathematical function, $z = \sum_{i=0}^{n} \sum_{j=0}^{m} P_{ij} B_i(x) B_j(y)$, where $P_{ij}$ is a control parameter; $n$ is the number of control
points in the $x$ direction; $m$ is the number of control points in the $y$ direction; and $B_i(x)$ and $B_j(y)$ are primary functions corresponding to $P_{ij}$. However, a few facets of the surfaces can be formulated into this kind of mathematical functions so that the Spline surface might have an oscillation problem.

Cell values of IDW surfaces are estimated by averaging the values of sample data points in the neighbourhood of each processing cell. The closer a point is to the centre of the cell being estimated, the bigger weight it has in the averaging process, which sometimes causes ‘Bull’s eye’ problems.

The Kriging method is an appropriate method for issues that have a spatially correlated distance or directional bias in the data. The Kriging method fits a mathematical function to a specified number of points, or all points within a specified radius, to determine the output value for each location, which might make a Kriging surface fragmental.

The HASM method has theoretically given a solution to error problems that have long troubled DEM compilation. It will greatly contribute to enhancing the accuracy of digital surface modelling (DSM). DSM can be defined as the process of numerically representing a planetary surface by grids with known coordinates in an arbitrary coordinate system, such as surface modelling of ecosystem changes, ecosystem scenarios and population distribution (Yue et al. 2003, 2005a,b,c, 2006, 2007b).

The HASM method can be used to support Earth observation image processing. Earth observations from satellites have been recognized as an excellent tool for providing both global and local views. A great variety of satellites and sensors are commercially or experimentally available. Recent developments show that SARs can provide high-resolution images that are independent from cloud coverage. Conventionally these images consist of a large amount of data. But, without combining the derived space-borne information with suitable ancillary data, especially ground references, which are present in different kinds of information systems, the resulting information is not able to reflect the complex interactions on the Earth’s surface. When a global positioning system is connected with HASM, the satellite images can be used in the field (Zingler et al. 1999).

The biggest limitation to applications of the HASM operation is that it needs a very long computational time to process huge data. Therefore, the most important work in the future is to develop the fastest numerical methods for the HASM method by means of multigrid methods.

Acknowledgements
This work is supported by the Major Directivity Projects of Chinese Academy of Science (kzcx2-yw-429 and kzcx2-yw-308), by the National Key Technologies R&D Program of the Ministry of Science and Technology of the People’s Republic of China (2006BAC08B04), by the One-Hundred Outstanding Scientists Project of the Chinese Academy of Sciences (KZCX0504), by the National High-tech R&D Program of the Ministry of Science and Technology of the People’s Republic of China (2006AA12Z219) and by the International Cooperation Program for Science and Technology (2006DFB919201).
References


