INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2001, VOL. 15, NO. 8, 721–741

Research Article

Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS

HONGXING LIU* and KENNETH C. JEZEK

The Byrd Polar Research Center and Department of Geography, The Ohio State University, Columbus, OH 43210, USA;
e-mail: jezek@iceberg.mps.ohio-state.edu

and MORTON E. O'KELLY

Department of Geography, The Ohio State University, Columbus, OH 43210, USA; e-mail: okelly.1@osu.edu

(*Received 4 February 2000; accepted 13 February 2001*)

Abstract. In this paper, we propose a new method for detecting outliers in an irregularly distributed spatial data set. Our method has two desirable properties. First, it is functionally effective due to the introduction of sensitive outlier indices and locally adaptive and robust statistical criteria. Second, it is computationally efficient because of the use of super-block based spatial data sorting and searching scheme. Our method has been implemented using the C programming language and integrated with the Arc/Info GIS system. The integration leads to a powerful exploratory data analysis tool for checking and analysing anomalous values in a GIS environment. Local outliers can be automatically labeled with our method, subject to some user-defined parameters. Outliers represent anomalous or suspicious values in a statistical sense, which may not necessarily be erroneous values. Instead of being simply discarded, statistical outliers should be investigated further using prior qualitative knowledge or in association with other GIS data layers.

1. Introduction

With the rapid development of geographical information system (GIS) technology, various spatial databases have been constructed in recent decades in support of a wide range of geo-scientific and environmental studies. However, errors may arise during the data acquisition process due for a variety of reasons, such as malfunction or improper calibration of instruments, mistaken readings, gross recording, and calculation and execution faults. The existence of erroneous values plagues subsequent spatial data analyses and often distorts the inference process. To ensure the quality of spatial databases and prevent error propagation, it is critical to develop

*Current address: Department of Geography, Texas A & M University, College Station, TX 77843, USA; e-mail: liu@geog.tamu.edu

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810110060442

error detection and correction techniques (Chrisman 1991, Lunetta *et al.* 1991, Lanter and Veregin 1992, Thapa and Bosseler 1992).

Previously, formal statistical methods have been developed to deal with outliers in the context of non-spatial data (Tietjen and Moore 1972, Barnett 1983, Iglewicz and Hoaglin 1993, Barnett and Lewis 1994, Luceño 1998). Also, some techniques have been proposed to detect errors in grid-based spatial data sets. Hannah (1981) developed an algorithm for error detection in a digital elevation model (DEM). He used the constraints on both the slope and the change in slope within a local area. Felicísimo (1994) advanced a parametric statistical method for detecting anomalous values in grid-based data sets. His method is based on the comparison of the observation value at each grid node with the value interpolated from its neighbouring grid nodes. López (1997) used principal component analysis to deal with random errors in digital elevation models. However, little research has been reported on error detection techniques for irregularly distributed spatial data.

In reality observations and measurements of geographical phenomena frequently occur unevenly and irregularly over space. Moreover, many regular spatial data sets are originally interpolated from irregular data sets, namely, scattered and/or traverse type of measurements. Due to a lack of effective automatic methods, some informal procedures are often used in practice for checking and editing irregularly distributed data sets. One intuitive procedure is visual inspection of point values portion by portion in hardcopy printouts or on computer screen. This procedure is prohibitively time consuming and tedious for a large and dense data set. Moreover, it is often difficult to detect outliers through visual inspection without the aid of analytical tools. The other procedure often used is to interpolate an irregularly distributed data set into a regular grid, and then to apply error detection techniques designed for grid-based data, for example, analytical hill-shading methods (Kraus 1994) and the methods developed by Hannah (1981) and Felicísimo (1994). This indirect procedure has two major drawbacks. First, when an irregularly distributed data set is interpolated into a regular grid, errors propagate and spread out into their neighbourhoods. Furthermore, outliers may already be suppressed in the interpolation process. This makes subsequent error detection and removal very difficult or impossible. Second, even if errors are identified in the interpolated regular data set, they have to be traced back to the original irregular data set. Interpolation and backward error-tracing certainly need additional efforts.

In this paper, we present an automatic outlier detection method designed for irregularly distributed data sets. By combining a super-block based searching strategy and locally adaptive and robust statistical analysis, our method provides a strong capability for detecting local outliers at a minimized computational cost. The proposed approach has been implemented using C programming and integrated with the ARC/INFO GIS. By specifying a desired degree of confidence and some other parameters, outliers can be automatically labelled within a massive irregularly distributed data set. The specification of these parameters involves user's subjective judgment. Outliers are anomalous or suspicious values in a statistical sense; that is, they are strongly inconsistent with their neighbour points and deviate markedly from a statistical model based on other nearby values. It should be noted that statistical outliers are not necessarily erroneous data though they are highly likely to be so. Instead of being simply discarded, statistical outliers should be further investigated using prior qualitative knowledge or in association with other geographical data sets. The justified errors can then be eliminated, remeasured, or replaced by the interpolated value, depending on the actual application context.

In the following sections, we first outline basic assumptions underlying our method, and then examine the algorithms involved in our method. Afterwards, we illustrate how to verify and analyse the detected anomalous values in a GIS environment. This is followed by an application example. In the final section we present conclusions.

2. Basic assumptions

In this paper, we assume that irregularly distributed spatial data under investigation are observations or measurements on a spatially continuous geospatial variable, such as topographical elevation, geomagnetic field, gravitational potential, atmospheric temperature and pressure, salinity of ocean water, or soil acidity. These types of geospatial variables exhibit gradual and continuous variation over space, and can be graphically represented as the shape of a surface. Consequently, spatial data sampled on continuous geographical phenomena are often referred to as surface-type data (Robinson *et al.* 1984, Carter 1988) or terrain-type data (Clarke 1995).

A continuous geospatial variable Z can be represented as a function of planimetric coordinates (X,Y), namely, $Z = f(X,Y)$. Mathematically, it is often simplified as a continuous and single-valued function. Surface-type data are a sample of discrete measurements or observations on the continuous surface, consisting of a series of XYZ triplets. If we have adequate information about the nature of the observations, we might be able to give a tangible explanation for most of the anomalous values and hence work out a way to treat them. In many cases, however, we, especially as data users, may not know about unusual circumstances affecting the observations. Therefore, we can only rely on the internal spatial relationship of the data in judging which observations are outliers.

The fundamental assumption of our method is the spatial continuity and autocorrelation of surface-type data. The presence of serious errors tends to destroy the local continuity. Gross errors would appear as erratic peaks, pits, or unnatural abrupt features if the data were graphically represented as a surface. Checking local consistency and continuity for each data point in the context of its nearby points can provide an important outlier detection clue. If a point is strongly inconsistent with its neighbour points and statistically unreasonable, we consider it as a local outlier. The key components constituting our method include identification of neighbour points, construction of outlier indices, and locally adaptive and robust discordance analysis.

3. The outlier detection method

3.1. *Definition of neighbour points*

To check local continuity and consistency, we need to define a neighbourhood for each point by selecting a subset of surrounding points. In the case of grid-based spatial data, the adjacency and proximal relationships among data points are visually obvious and computationally implicit in the order of row and column indices. However, for an irregularly distributed data set, either scattered or traverse data, the identification of neighbour points is not straightforward because the number and locations of nearby points are varied and irregular. To avoid the directional bias and keep the number of neighbouring points from being excessively large or insufficiently small, we choose the eight nearest points, each from one octant as illustrated in figure 1.

The neighbour points identified by octant search are similar to the concept of

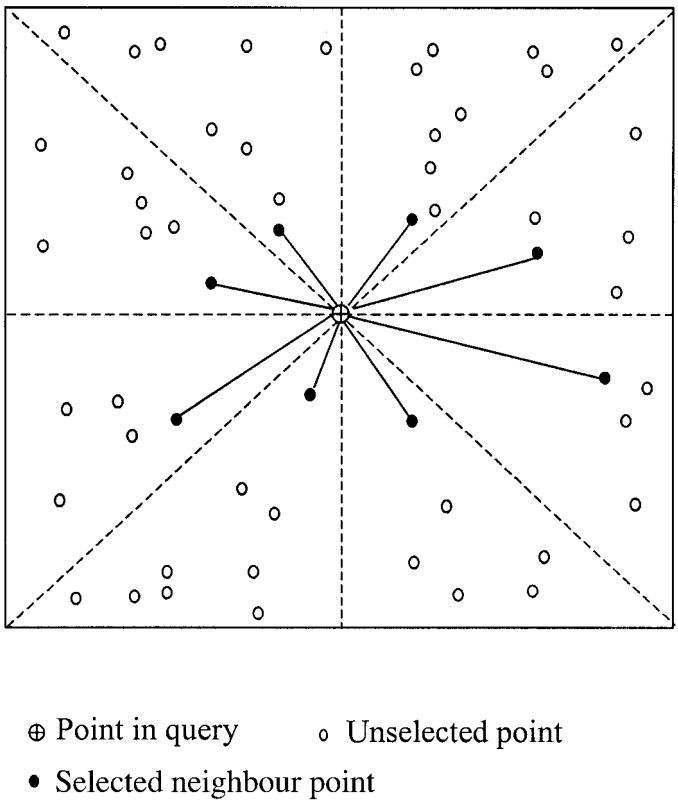

Figure 1. Definition of octant neighbour points.

natural neighbour points advocated by Watson (1985, 1992). The natural neighbour points can be identified through Delaunay triangulation (McCullagh *et al.* 1980, Macedonio and Pareschi 1991, Tsai 1993, Shewchuk 1996). The Delaunay triangulation is the geometric dual of the Voronoi (or Thiessen) polygons. Delaunay triangulation not only generates an elegant tessellation of space, but also establishes full-fledged topological relationships between data points. Natural neighbours of a query point are the surrounding points that are directly linked to the query point by the edges of Delaunay triangles. On average, each data point has six first-order natural neighbours (Watson 1992, pp. 57–85). Our purpose for identifying neighbour points is to construct a local interpolation residual index and a surface gradient index, rather than to create an elegant tessellation or derivation of the topological relationships. In this context, octant neighbour points are adequate to serve our purpose as they automatically compensate for directional bias and local variation in the density of data points like the natural neighbour points. Compared with the Delaunay triangulation based natural neighbour method, our octant search method has less computational cost and relies on a much simpler data structure.

3.2. *Super-block based spatial sorting and searching*

For a data point, its octant neighbours might be identified by exhaustively searching the entire data set. However, this brute-force method is often prohibitively time-consuming and even operationally impossible when the data set under

investigation is large. An efficient approach to the nearest neighbour problem is to establish the proximal order by spatially sorting and organizing data points before the search. Al-Daoud and Roberts (1996) compared three alternative sorting techniques, namely, K-d trees, quadtrees, and super-block based methods. They found that the super-block based method is the most efficient for extracting the proximal relationships among irregularly distributed points. We utilize this method for identifying octant neighbour points and defining a local area for estimation of local robust summary statistics.

Super-block based method, also known as the cell or bin-based method, is an optimal expected-time algorithm (Bentley *et al.* 1980, Hodgson 1989, Deutsch and Journel 1992, pp. 30–34, Al-Daoud and Roberts 1996). The idea is to partition the bounding rectangular area defined by the minimum and maximum X and Y coordinates of the data set into an array of super blocks (cells or bins), then assign each point to a block according to its location (figure 2). Experiments show that the optimal number of data points per block is about three (Bentley *et al.* 1980, Al-Daoud and Roberts 1996), which can be approximately achieved by varying the number of super-blocks along the X or Y direction:

$$l = \sqrt{\frac{m}{c}} \tag{1}$$

where l is the number of blocks in the X or Y direction, m is the total number of data points, and c is the optimal number of points per block. An array is constructed to store the cumulative number of data points for indexed block i:

$$cum(i) = \sum_{j=1}^{i} p(j) \; (i = 1, 2, ..., k) \tag{2}$$

where $cum(i)$ is the cumulative number of points in super blocks 1, 2, ..., i, $p(j)$ is the number of data points in block j, and k is the total number of super blocks. The

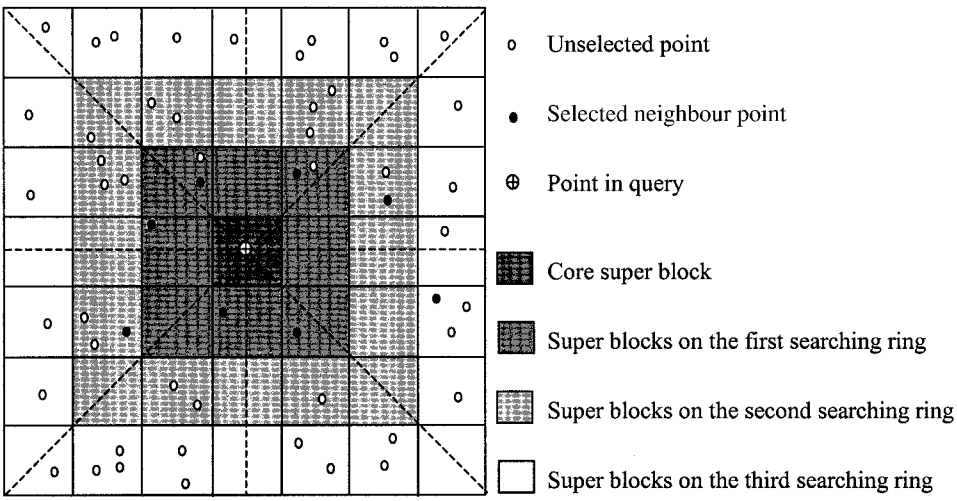

Figure 2. Super-block based sorting and expansion searching.

number of data points falling within any block i can be derived by the equation:

$$p(i) = cum(i) - cum(i-1) \; (i = 1, 2, ..., k)$$
$$cum(0) = 0 \tag{3}$$

Therefore, a single array $cum(i)$ contains information on both the number of data points in each block and their location in memory for retrieval. This is an important computational strategy used in the super-block method to sort data points in two-dimensional space (Deutsch and Journel 1992, pp. 30–34).

With the data points sorted into super-blocks, nearest neighbour points can be quickly identified by searching only a limited number of nearby blocks in a relatively small neighbourhood around the query point **q**, instead of searching the entire data set. As shown in figure 2, the searching is performed in an expansion fashion. Namely, we start with the core super block to which point **q** is located, and then proceed to the surrounding blocks outward ring by ring until at least one point is found in each octant, or until we have searched to a specified maximum radius. As the geometric shape of the searching ring is a square instead of a circle, a candidate nearest point found in the super blocks on the n*th* searching ring in a diagonal direction might be farther than points in the super blocks on the $(n+1)$*th* searching ring in a north-south or east–west direction (figure 2). To ensure a true nearest point for each octant, the searching must continue for several additional rings after candidate nearest points are found for all octants. In the case of square super blocks, the number of additional rings that need to be searched is defined by **integer(*0.414n* + *0.5*)**, if the last candidate nearest point is identified in a super block on the n*th* searching ring.

A typical geometric configuration of the neighbourhood formed by the selected octant neighbour points is illustrated in figure 3. Though the size and shape changes according to density and spatial arrangement of the data points, the octant neighbourhood defined in this way has the qualities of compactness and equi-angularity. Namely, the area of the neighbourhood is small, and there is a neighbour point within every 45° angular range. This effectively compensates for directional bias and

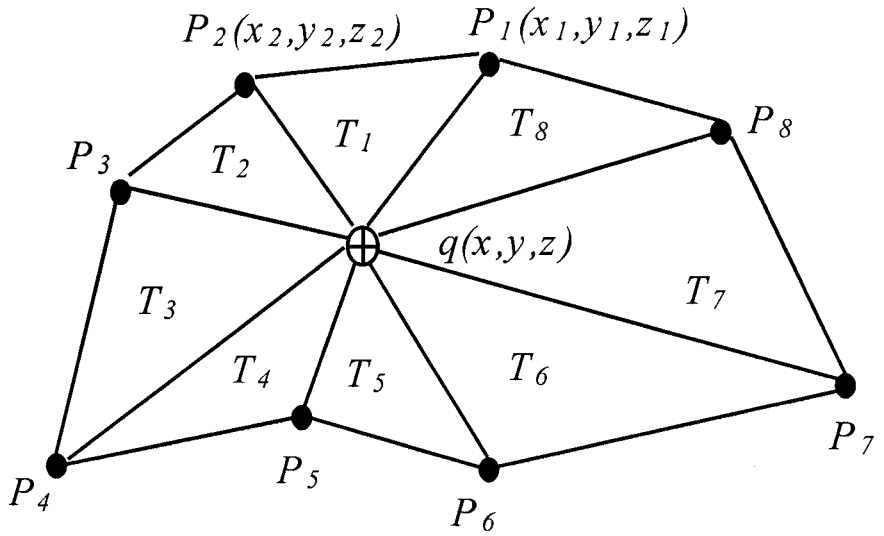

Figure 3. A typical geometric configuration of octant neighbourhood.

local density variation of observation points and hence makes the following local consistency and continuity test more reliable and effective.

3.3. *Local outlier indices*

Based on octant neighbour points, we designed two local outlier indices: interpolation residual based index and surface gradient based index.

3.3.1. *Interpolation residual based index*

With continuity and autocorrelation assumptions of the underlying surface, the Z value of a data point can be well predicted through interpolation based on its surrounding neighbouring points. The difference between the predicted value and the observed value indicates the extent to which the data point under examination is inconsistent with its neighbours. The idea of checking the observation value against the predicted value from its neighbour points is referred to as cross-validation (Burt and Barber 1996). Others have used this technique in different contexts, for example, for selecting among alternative possible covariance models (Davis, 1987, Isaaks and Srivastava 1989, pp. 351–364) and for accommodating noisy scattered data in surface fitting (Wahba and Wendelberger 1980, Wahba 1984).

To obtain a reliable and outlier-resistant prediction, we designed a robust octant inverse distance weighting (IDW) algorithm, in which the predicted value is a linearly weighted function of its octant neighbours:

$$\hat{z}_q^* = \sum_{i=1}^{8} w_i z_i \tag{4}$$

$$w_i = \frac{d_i^{-b}}{\sum_{j=1}^{8} d_j^{-b}} \tag{5}$$

where $\hat{z}_q^*$ is the predicted value for the query point q, z_i is the observation value of neighbour point p_i, d_i is the distance between the query point q and its neighbour point p_i, w_i is the weight of neighbour point p_i, and b is the global distance friction factor.

One obvious problem with equation (4) and (5) is that if outliers exist in the selected octant neighbour points, the predicted result will be contaminated and biased. For example, the interpolated elevation value for point A in figure 4 is 216, which is greatly biased by its immediate neighbour B (outlier). Since the point A has a large prediction residual of 62 (table 1), it tends to be misidentified as an outlier.

To make the prediction robust in the face of outliers in the neighbour points, we use a Jackknife technique to identify the most influential neighbour points (Burt and Barber 1996). Namely, we drop one neighbour point at a time and use the remaining seven neighbour points to make our prediction. Repeating this procedure for every neighbour point, we obtain eight additional predictions:

$$\hat{z}_q^{(k)} = \sum_{i \neq k} w_i z_i \qquad (k = 1, 2, ..., 8) \tag{6}$$

$$w_i = \frac{d_i^{-b}}{\sum_{j=1}^{8} d_j^{-b} - d_k^{-b}} \qquad (k = 1, 2, ..., 8) \tag{7}$$

where $\hat{z}_q^{(k)}$ is the predicted elevation value with the neighbour point p_k omitted, and

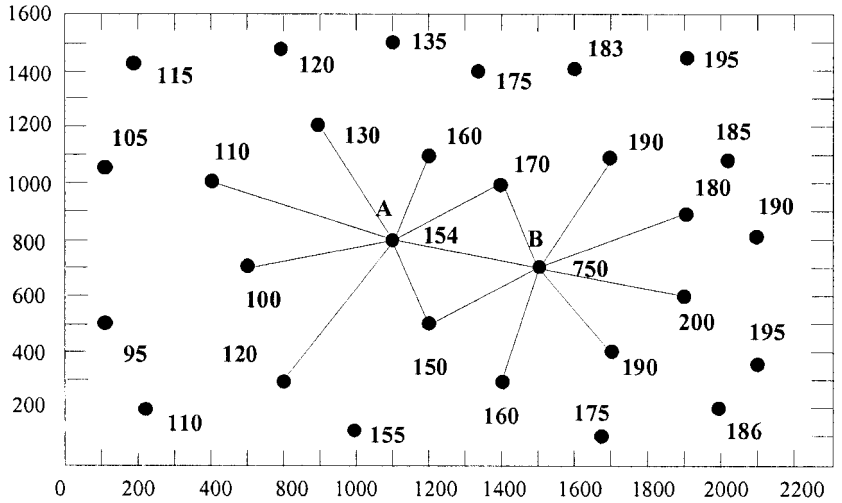

Figure 4. A hypothetical data set for illustration of residual and gradient index calculation. Assume that the data values are surface elevation in metres.

Table 1. Calculation of residual and gradient indices.

Data point	Observed value	Interpolated value		Interpolated residual		Gradient (°)	
		Non-robust	Robust	Non-robust	Robust	Non-robust	Robust
Point A (normal)	154	216	135	62	−9	36	5.7
Point B (outlier)	750	172	181	578	569	59	57

Note: calculated from a hypothetical data set shown in figure 4.

other parameters are the same as in equations (4) and (5). The absolute difference $|\hat{z}_q^{(k)} - \hat{z}_q^*|$ indicates the influence of the neighbour point p_k. We drop the two most influential neighbour points p_{k1} and p_{k2}, those having the largest absolute difference $|\hat{z}_q^{(k)} - \hat{z}_q^*|$, and then take the weighted average of the remaining neighbour points as the robust estimate of the surface value at point q:

$$\hat{z}_q = \sum_{i \neq k1,k2} w_i z_i \tag{8}$$

$$w_i = \frac{d_i^{-b}}{\sum_{j=1}^{8} d_j^{-b} - d_{k1}^{-b} - d_{k2}^{-b}} \tag{9}$$

where $\hat{z}_q$ is the robust prediction value calculated with the two most influential neighbour points p_{k1} and p_{k2} omitted. The decision to remove the two most influential neighbouring points is subjective, where one could instead have chosen to drop one or three influential points, depending upon one's perception about the percentage of outliers contained in the data set.

The residual between the observed value (recorded in the data set) and the

predicted value at point q is used as a local outlier index:

$$\Delta z_q = z_q - \hat{z}_q \tag{10}$$

where Δz_q is the signed residual, and z_q is the observed elevation value at the query point q. Since the predicted value from equation (8) approximates the 'true' surface value of the point being checked, the residuals calculated in (10) mainly account for the error term. If the point being checked is contaminated by serious error, the absolute value of Δz_q must be significantly larger and markedly deviate from the central tendency of the residual index.

With this computational approach, we obtain a robust prediction of 135 for point A (table 1), even though outlier point B is located nearby. A small prediction residual of -9 for point A and a large prediction residual of 569 for point B (table 1) enable us to keep point A as a normal point and identify point B as an outlier.

3.3.2. *Surface gradient-based index*

The surface gradient, indicating the surface shape and the local variation, is a good measure of local surface continuity. The surface gradient around an individual data point can be estimated from its adjacent neighbour points. As shown in figure 3, the point q in conjunction with its surrounding neighbour points form eight triangles. These triangles have various orientations in three-dimensional space, depending on the relative positions and heights of each pair of neighbours. The gradient for each triangle can be calculated by taking the cross product of any pair of sides from each triangle (Watson 1992, p. 95). For the triangle T_1 in figure 3, the cross product vector (X_1, Y_1, Z_1) is given by the Cartesian coordinates of the vertices $p_1(x_1, y_1, z_1)$, $p_2(x_2, y_2, z_2)$ and $q(x, y, z)$:

$$\begin{aligned} X_1 &= (y_2 - y_1)(z - z_1) - (y - y_1)(z_2 - z_1) \\ Y_1 &= (z_2 - z_1)(x - x_1) - (z - z_1)(x_2 - x_1) \\ Z_1 &= (x_2 - x_1)(y - y_1) - (x - x_1)(y_2 - y_1) \end{aligned} \tag{11}$$

This three-dimensional vector is perpendicular to the triangle plane T_1, and its length is twice the area of the triangle (Watson 1992: 95–95). Therefore, the gradient (G_1) and the area (a_1) of the triangle plane T_1 can be calculated as:

$$G_1 = \sqrt{\left(\frac{X_1}{Z_1}\right)^2 + \left(\frac{Y_1}{Z_1}\right)^2} \tag{12}$$

$$a_1 = \frac{\sqrt{X_1^2 + Y_1^2 + Z_1^2}}{2} \tag{13}$$

The weighted average of the gradients of the surrounding triangles reflects the overall gradient around the query data point q. If a triangle is large, its two neighbour points must be far away from the query point, and the calculated gradient would be less reliable for representing the surface shape around the query point. Therefore, the larger the triangle, the smaller the weight it receives.

$$G_q^* = \sum_{i=1}^{8} G_i w_i \tag{14}$$

$$w_i = \frac{a_i^{-1}}{\sum_{j=1}^{8} a_j^{-1}} \tag{15}$$

where G_q^* is the weighted average gradient, G_i is the gradient of triangle T_i, a_i is the area of triangle T_i, and w_i is the weight of triangle T_i.

The surface gradient calculated in this way is subject to error, if outliers exist in the neighbour points. To make the gradient estimate robust to the presence of outliers in the neighbouring points, we drop the two most influential triangles T_{k1} and T_{k2} that have the largest gradient values. The weighted average of surface gradients of the remaining six triangles is taken as the robust estimate for the surface gradient around the point q:

$$G_q = \sum_{i \neq k1,k2} w_i G_i \tag{16}$$

$$w_i = \frac{a_i^{-1}}{\sum_{j=1}^{8} a_j^{-1} - a_{k1}^{-1} - a_{k2}^{-1}} \tag{17}$$

where G_q is the robust weighted average gradient with the triangle T_{k1} and T_{k2} dropped out. The decision to drop the two triangles is again arbitrary, where one could omit a different number of influential triangles.

The calculation of surface gradient is very sensitive to data noise, and unreasonable and erratic values are often associated with serious data errors, for example, point B in figure 4. By checking gradient estimates against their local trend, we can identify anomalous data points too. As shown in table 1, the robust estimates of surface gradients for point A and B are 5.7° and 57°, compared with non-robust values of 36° and 59° computed from equations (14)–(15). Again, a small robust gradient value for point A and a large robust gradient value for point B make it possible to keep point A as a normal point and detect point B as an outlier.

3.4. *Locally adaptive and robust discordancy analysis*

Although spatial autocorrelation is expected for direct surface-type data such as topographic elevation, we can safely assume the randomness and independence of the residual index Δz_q and gradient index G_q in a local area. Since the robustly interpolated values approximate the 'true' surface values of the points being checked, the residual between the observed values and the interpolated values mainly account for the random errors, which are expected to have a negligibly small autocorrelation and be stationary within a small local area. Since these two indices are calculated in a local context, they make subtle anomalous points more pronounced. Based on the central tendency and dispersion of the outlier indices, we can establish objective criteria to differentiate outliers from the rest of normal points.

In the context of grid-based data, Hannah (1981) used subjectively predefined, fixed thresholds by specifying the allowable slope and allowable slope change. Felicísimo (1994) employed a parametric statistical test based on global estimates for the central tendency and dispersion. Distinguishing from their global approaches, we developed locally adaptive and robust statistical criteria, which are more powerful in detecting local outliers.

Most spatial processes are non-stationary, and the resulting surfaces are often complex and heterogeneous. In the case of topography, for example, the magnitude and dispersion of interpolation residuals would be relatively small in a flat or slightly undulated area and large in a rugged and hilly area. Since the stationarity of outlier indices is not guaranteed, the estimates for the central tendency and dispersion must be conducted locally in order to account for the heterogeneity of the spatial data

set. If we use global estimates based on the entire data set, the local outliers tend to be more hidden and less intuitively apparent due to regional heterogeneity. For example, in figure 5 the interpolation residual values at data points N, O and P is typical if subset A (relatively flat area) and subset B (rugged area) are considered together. But, they are surely not typical within its local neighbourhood (within subset A). To detect subtle local outliers, we need to develop locally adaptive criteria in place of global criteria derived from the whole data set.

We use super blocks as the basic unit to construct a local area for estimating the central tendency and dispersion of outlier indices. To obtain the valid and stable local summary estimates, the number of data points in a local area must be sufficiently large. In our application example, we set the minimum number of data points in the local area to be 45. For each non-empty super block, we check the number of data points inside. If the number is smaller than the specified number, we expand the local area in an expansion fashion (figure 2) by including the surrounding super blocks until the number of points is equal to or greater than the specified number. The extent of the local area defined in this way varies according to the density of data points. Within a small local area, the spatial processes underlying the surface-type data can be more safely assumed to be stationary.

For a local area, the mean and standard deviation are the best, unbiased measures for the central tendency and the dispersion (spread), on the assumption of a normal distribution of the sample without contamination. In the case of the interpolation residual index, the mean ($\Delta\bar{z}$) and standard deviation (s) are defined as:

$$\Delta\bar{z} = \frac{\sum_{j=1}^{n} \Delta z_j}{n} \tag{18}$$

$$s^2 = \frac{\sum_{j=1}^{n} (\Delta z_j - \Delta\bar{z})^2}{(n-1)} \tag{19}$$

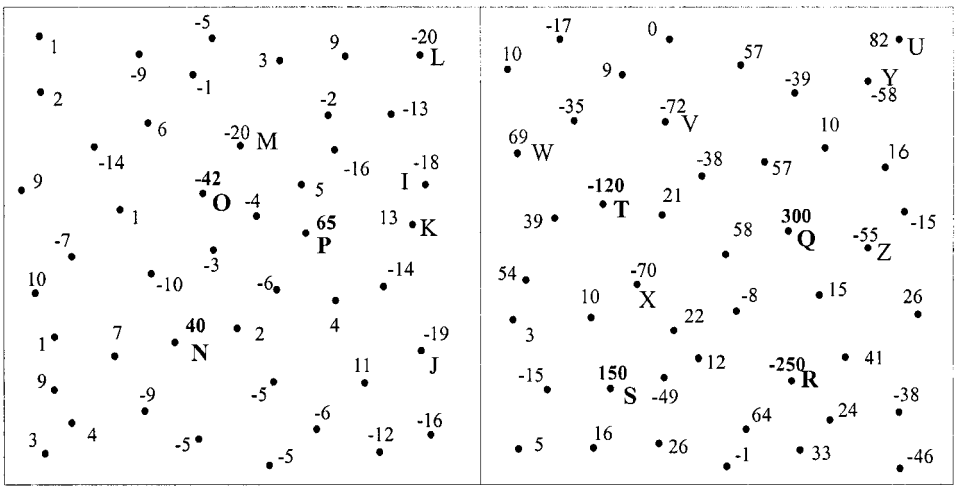

Figure 5. A hypothetical data set for illustration of effectiveness of locally adaptive and robust statistical criteria. Assume that the data values are interpolation residuals.

where n is the number of data points in a local area. The deviation/spread statistic T approximately follows the Student t distribution with $n-1$ degrees of freedom (Burt and Barber 1996, pp. 269–271, Felicísimo 1994, Barnett and Lewis 1994):

$$T=\frac{\Delta z_j-\Delta\bar{z}}{s} \tag{20}$$

Therefore, the probability

$$P(-t_{\alpha/2,n-1}\leqslant\frac{\Delta z_j-\Delta\bar{z}}{s}\leqslant t_{\alpha/2,n-1})=1-\alpha \tag{21}$$

where $1-\alpha$ indicates a confidence level, and $t_{\alpha/2,n-1}$ refers to the t value with an upper-tail probability of $\alpha/2$ and with $n-1$ degrees of the freedom. Based on the local estimates of central tendency and dispersion in this probability model, we can construct a local confidence interval $(\Delta\bar{z}-t_{\alpha/2,n-1}\cdot s, \Delta\bar{z}+t_{\alpha/2,n-1}\cdot s)$. The probability that the random variable Δz_j assumes a value outside this interval is α. Set α sufficiently small, say, 0.05 or 0.01, and we can perform a discordancy test by examining whether the residual value at point q is within the local confidence interval. If the calculated residual Δz_q is beyond this interval, we flag the point under investigation as a local outlier, because the value deviates markedly from the statistical model and appears unreasonable and anomalous relative to the local central tendency and dispersion statistics.

However, the central tendency and dispersion measured by conventional mean and standard deviation in equations (18) and (19) and corresponding confidence interval may be flawed and deficient for two reasons. First, since mean and standard deviation are highly sensitive to skew (asymmetry), kurtosis (peakedness), and unusually short or long tails, they could be poor measures for the central tendency and dispersion if the data departs from the normal distribution. Secondly, the conventional mean and standard deviation are also vulnerable to the presence of outliers in the sample. When multiple outliers exist, the most extreme outlier would yield a strong masking effect on other outliers (Tietjen and Moore 1972, Barnett and Lewis 1994, p. 114) if the conventional mean and standard deviation are used to construct the confidence interval. For example, with local mean and standard deviation based criteria and at the confidence level $a=0.01$ the most extreme point P in subset A masked points N and O, and the most extreme points Q and R in subset B masked points S and T (figure 5 and table 2).

Although other measures of central tendency and dispersion could be used, we adopt trimmed mean and winsorized standard deviation (Barnett and Lewis 1994, pp. 68–69) to provide robust estimates for the central tendency and the dispersion:

$$\Delta\bar{z}_{trim}=\frac{\sum_{i=k+1}^{n-k}\Delta z_{(i)}}{n-2k} \tag{22}$$

$$\Delta\bar{z}_{win}=\frac{k(\Delta z_{(k+1)}+\Delta z_{(n-k)})+\sum_{i=k+1}^{n-k}\Delta z_{(i)}}{n} \tag{23}$$

$$s^2_{win}=\frac{k[(\Delta z_{(k+1)}-\Delta\bar{z}_{win})^2+(\Delta z_{(k+1)}-\Delta\bar{z}_{win})^2]+\sum_{i=k+1}^{n-k}(\Delta z_{(i)}-\Delta\bar{z}_{win})^2}{n-2k-1} \tag{24}$$

Table 2. Effectiveness of locally adaptive and robust statistical criteria.

Methods		Global criteria	Locally adaptive criteria	Locally adaptive and robust criteria
Central tendency		Global mean (2.52)	Local mean Subset A (−1.69) Subset B (6.73)	Local trimmed mean Subset A (−2.70) Subset B (6.21)
Dispersion		Global Std. Dev. (54.53)	Local Std. Dev. Subset A (16.15) Subset B (75.62)	Winsorized Std. Dev. Subset A (7.00) Subset B (29.33)
$\alpha = 0.05$	Confidence interval	(−106.00, 111.03)	Subset A (−34.31, 30.93) Subset B (−146.02, 159.48)	Subset A (−16.98,11.58) Subset B (−53.62, 66.04)
	Detected outliers	Q, R, S, T detected	N, O, P detected in subset A Q, R detected in subset B	I, J, K, L, M N, O P detected in subset A Q, R, S, T, U, V, W, X, Y, Z detected in subset B
$\alpha = 0.01$	Confidence interval	(−141.44, 146.48)	Subset A (−45.13, 41.75) Subset B (−196.69, 210.15)	Subset A (−21.88, 16.48) Subset B (−74.15, 86.57)
	Detected outliers	Q, R, S detected	P detected in subset A Q, R detected in subset B	N, O, P detected in subset A Q, R, S, T detected in subset B

Note: calculated from a hypothetic data set shown in figure 5.
Std. Dev. = Standard Deviation.

where $\Delta\bar{z}_{trim}$ is the trimmed mean, $\Delta\bar{z}_{win}$ is the winsorized mean, s_{win} is the winsorized standard deviation, $\Delta\bar{z}_{(i)}$ denotes the ith ordered interpolation residual, k is the number of observations eliminated or winsorized at each end of the distribution, and n is the total number of observations in the local area. As indicated in equations (22)–(24), winsorization replaces the lower and upper ends of the ordered interpolation residuals by their nearest adjacent values, while the trimming simply discards the lower and upper ends (Barnett and Lewis 1994, pp. 68–69). According to the suggestion of Iglewicz and Hoaglin (1993), we set the number (k) of observations trimmed or winsorized at each end to 15% of the total observations. If 50% of the observations are trimmed or winsorized, then respective means would be reduced to the median of the sample observations. After trimming or winsorizing the extreme sample values at two ends, the remaining data can be regarded as a clean subset that is presumably free of outliers.

With the conventional mean $\Delta\bar{z}$ and standard deviation s replaced by the trimmed mean $\Delta\bar{z}_{trim}$ and winsorized standard deviation s_{win} in equation (20), the corresponding deviation/spread statistic is still distributed essentially as Student's t with $n-2k-1$ degrees of freedom over a range of different possible distributions and in the presence of outliers (Tukey and McLaughlin 1963, Patel *et al.* 1988, Iglewicz and Hoaglin 1993). As a result, we obtain a robust confidence interval $(\Delta\bar{z}_{trim} - t_{\alpha/2,n-2k-1} \cdot s_{win}, \Delta\bar{z}_{trim} + t_{\alpha/2,n-2k-1} \cdot s_{win})$. Since the trimmed mean and winsorized standard deviation are less susceptive both to the form of the distribution and to outliers, they provide robust estimates for the central tendency and dispersion statistics over a spectrum of possible distributional forms, such as contaminated versions of normal or other symmetric distributions.

As shown in table 2, if we use the conventional global criteria, only points Q, R and S are detected as outliers at the confidence level $a=0.01$. If we use the locally adaptive and robust criteria based on the trimmed mean and winsorized standard deviation, anomalous data points N, O and P in subset A and points Q, R, S, and T in subset B are successfully detected. The specified confidence level affects the number of detected outliers as well. More outliers will be labeled if we reduce the degree of confidence. For example, at the confidence level $a=0.05$, points I, J, K, L, and M in subset A and U, V, W, X, Y, Z in subset B are also flagged as outliers (figure 5 and table 2).

The distributional form of surface gradient G_q defined for local areas might be more likely to deviate from the normal distribution. Hence, the use of robust estimators is more important for obtaining reliable estimates for the central tendency and dispersion. Consequently, we use the trimmed mean and winsorized standard deviation to construct the confidence interval. Since the surface gradient has positive values and we are only interested in the upper bound outliers, the one-sided discordance test is conducted to see if the surface gradient value G_q at point $\boldsymbol{q}$ is larger than the upper bound of the robust interval, namely, $\bar{G}_{trim} + t_{\alpha,n-2k-1} \cdot s_{win}$, where $\bar{G}_{trim}$ is the trimmed mean, and s_{win} is the winsorized standard deviation of the surface gradient index.

3.5. *Algorithm summary and implementation considerations*

To summarize, the proposed method consists of the following component steps:

1. Partition the irregularly distributed data set into an array of super blocks and sort the data points using a super-block based method;

2. For each data point, identify its octant neighbour points by an expansion searching strategy;
3. For each data point, calculate the interpolation residual between the observation value and the interpolated value from its neighbour points using equations (8), (9), and (10);
4. For each data point, calculate the robust estimate for surface gradient using equations (16) and (17);
5. Repeat the steps 2–4 until all data points are visited.
6. For each super block, define a local area in an expansion way, calculate the trimmed mean and winsorized standard deviation for the residual index by using equations of (22)–(24), and construct the corresponding confidence interval. Similarly, derive the trimmed mean and winsorized standard deviation and corresponding confidence interval for the surface gradient index.
7. Check each data point to see whether the calculated residual index and gradient index are located within their local confidence interval. If either of them is outside the confidence interval, we label the corresponding data point as a local outlier.

The proposed method involves a considerable amount of computation. The reduction of computation cost is essential for practical applications of this method to a large and dense data set. The most time consuming step in our method is the identification of octant neighbour points. As stated earlier, it is optimized by the use of super-block based sorting and searching schemes. The computation cost involved in the Jackknife technique is also minimized by using equations (6) and (7) in which we subtract one point from the sum calculated in the previous step, instead of accumulating the remaining values again. In addition, one-dimensional quicksort algorithm (Press *et al.* 1992, pp. 332–336) is used to sort the data for calculating the trimmed mean and the winsorized standard deviation in equations of (22)–(24). These computational strategies make our error detection method computationally efficient.

In our implementation, almost all parameters were coded as variables instead of fixed constants. These include the maximum search distance, the distance friction factor, the number of the most influential points and triangles to be dropped in the calculation of robust outlier indices, the percentage of sample points to be trimmed or winsorized in the construction of robust summary statistics, minimum number of samples for defining a local area, and the confidence level for the statistical test. This increases the flexibility of the program, and allows for diagnostic exploration of data sets using different specifications of these parameters.

4. Verification and analysis of outliers in a GIS environment

The outliers detected by our locally adaptive and robust statistical analysis represent anomalous and doubtful values in the light of the continuity assumption for the underlying surface. However, statistical outliers are not definitely bad or erroneous points, though with a high probability to be so. For example, in spite of the fact that the terrain surface is generally continuous and smooth, we still can find natural cliffs and escarpments, towers, skyscrapers, and other man-made structures in the urban area. The correct data points along edges of these features might be flagged as outliers as well. If we discard all outliers detected, some good and true observations might be eliminated together with erroneous values. Therefore, we need

to further examine anomalous values and decide whether such statistical outliers should be retained or eliminated.

When the loss in the resolution of the spatial data caused by discarding some good measurements are negligibly small compared to the problems caused by keeping a few bad values, one can simply drop all labelled outliers. However, to gain an understanding of error causes and distinguish good measurements from real erroneous values, further investigation of outliers is surely required.

We implemented our error detection method using the C programming language and integrated it with the ARC/INFO GIS system. By so doing, we are able to look at anomalous values in association with other external data layers in a GIS environment. For example, we can superimpose outlier points detected in an elevation data set on satellite images or digital hydrologic and geological maps for error verification. The outliers in a smooth terrain observed from the satellite imagery must represent errors of some kind. Conversely, if outliers are located around natural cliffs, steep gorges, or human-made structures, they may convey significant information about surface discontinuity and hence should be retained for further investigation. The justified erroneous data points can be removed, remeasured or replaced by the predicted value, depending on the actual application context.

The GIS technique is also useful in analysing the spatial pattern of anomalous points and identifying anomalous regions. The location and spatial pattern of outliers may provide important clues for reasoning potential causes for outlying observations. The random, clustered, or linear pattern of anomalous data often correspond to different error causes. If anomalous data points are randomly distributed, they are most likely caused by random recording errors or white noise of measuring instruments. If they are clustered in some regions, the outliers may be related to some systematic errors, for instance, the malfunctioning of measuring instruments in a special environment, miscalculations, edge effects when assembling smaller data sets into a big data set, or a similar type of circumstance. Identification of each cause surely renders an opportunity for improving the quality of the data-gathering process.

5. An application example

We have successfully applied our method to both scattered and traverse types of spatial data (Liu 1999). Here we only show the outlier detection result for a massive irregular elevation data set derived by an automatic Synthetic Aperture Radar (SAR) stereo technique (Leberl 1990). The stereo pair covers the Terra Nova Bay area of Transantarctic Mountains in Antarctica. It consists of the Radarsat Standard Beam 2 image (figure 6(*a*)) acquired on 9 October 1997 with an incident angle of 28° at the scene centre and Standard Beam 7 image (figure 6(*b*)) acquired on 20 September 1997 with an incident angle of 47°. The terrain in the scene is characterized by mountainous glacial landforms. The rugged mountain slopes and relatively flat valleys of Campbell Glacier and Priestley Glacier can be observed from SAR images (figure 6(*a*) and figure 6(*b*)). After the SAR stereo processing, a dense and irregular raw elevation data set, containing about 1 857 697 x, y, z triplets, was extracted from the stereo image pair, and the average data density is about 186 points per squared kilometre (Liu 1999). Contour representation of the raw elevation data (figure 6(*c*)) shows a considerable number of erroneous and noisy measurements. The most glaring errors include a number of artificial depressions and erratic hills at the locations of A, B, C, D marked on figure 6(*c*). The small closed and jagged contours in the glacial floors highlight the noisy nature of the raw elevation data derived from

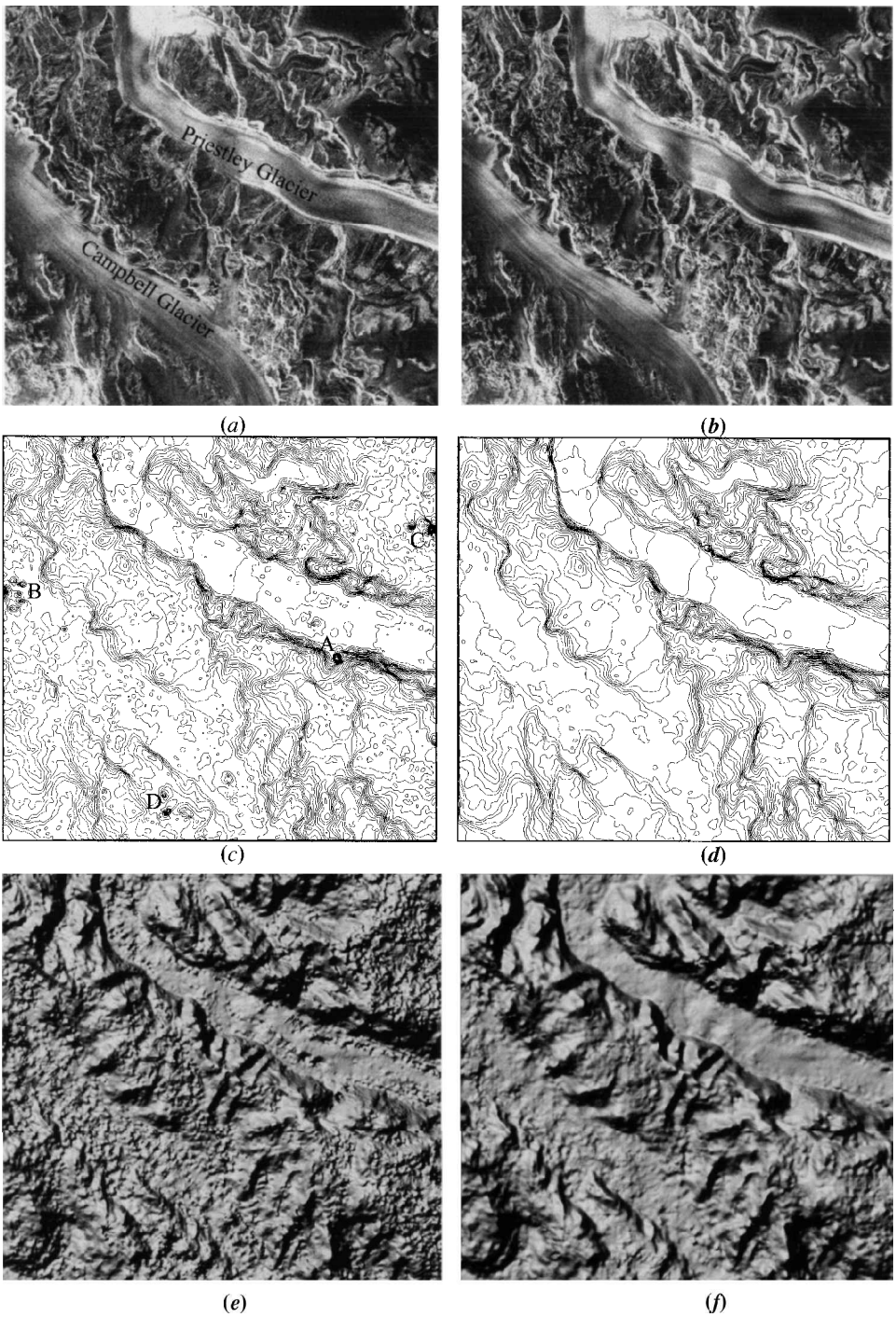

Figure 6. Outlier detection and removal in an irregular elevation data set derived from SAR stereo technique. (*a*) original Standard Beam 2 image of SAR stereo pair; (*b*) original Standard Beam 7 image of SAR stereo pair; (*c*) contour map derived from irregular raw elevation data before outlier removal. Contours are at 100 m intervals; (*d*) contour map of irregular elevation data after outlier removal; (*e*) hill-shaded image of the raw elevation data before outlier removal; and (*f*) hill-shaded image of the elevation data after outlier removal.

the SAR stereo technique. The artificial features and noisy spots are also evident in the relief-shaded image of the raw elevation data (figure 6(*e*)).

Using a Silicon Graphics Indy (SGI) workstation with a 175 MHz IP32 processor, it required 25 minutes 12 seconds to complete the entire computation for our massive data set of 1 857 697 irregular points, including the identification of octant neighbour points, calculation of outlier indices, and locally adaptive and robust discordance testing. Using the brute-force method, 1 minute 58 seconds was required to identify octant nearest neighbours for 100 data points in our massive data set. At this rate, the entire data set would require 608 hours 54 minutes, just for the identification of octant neighbours. Therefore, the brute-force search method is practically impossible for our example application.

Our method detects the overwhelming majority of anomalous values. Figure 6(*d*) and 6(*f*) respectively show the contours and hill-shaded image derived from irregular elevation data points after the removal of detected anomalous values. From figure 6(*d*) and figure 6(*f*), we can see that erroneous values have been successfully eliminated and the artificial features and spiky spots disappear, while the subtle terrain features are still preserved.

When labelled outliers are overlain onto the satellite radar image in a GIS system, we noticed that anomalous values are mainly concentrated either in the low contrast and featureless regions of flat glacial floors or in highly sloped regions of rugged mountains. Because of the poor texture of the terrain or relief-induced geometric distortions of SAR images (layover and excessive foreshortening) in these regions, an automatic image matching algorithm failed with SAR stereo pair, thereby producing gross errors in elevation measurements during the SAR stereo processing (Liu 1999).

6. Discussions and conclusions

Spatial data are often riddled with erroneous values that frustrate subsequent spatial analysis and inference. In this paper, we proposed a new method to detect anomalous values in irregularly distributed spatial data set. Based on octant neighbour points, we constructed two sensitive outlier indices. Different from the conventional global approach, we introduced locally adaptive and robust statistical criteria, namely, letting the criteria varying across the entire data set according to the robust summary statistics. Consequently, subtle local outliers that may not be detectable with conventional methods are revealed. Through synthetic and real data sets, we demonstrated that our method offers a sensitive capability for detecting local anomalous values in irregularly distributed data sets with a known degree of confidence. We believe that our outlier detection method will have a wide range of applications in disciplines that deal with spatially continuous data.

Outliers detected using our method only represent anomalous or doubtful values in a statistical sense, and they are not necessarily erroneous data. Outliers may convey important information about true discontinuity of geographical variables under examination or provide clues for determining potential causes for anomalous observations. In many cases, the location and spatial pattern of anomalous values have serious practical implications. It should be stressed that the analysis of statistical outliers has a wide range of applications other than error detection. For example, one is often most interested in the anomalies in a geostatistical analysis, such as the high-grade vein in a gold deposit or the impermeable layers that condition flow in a petroleum reservoir (Isaaks and Srivastava 1989, pp. 40–65, 351–368). The outlier

detection method can help locate such anomalous regions. In the field of geographical data mining and geographical knowledge discovery, outlier detection is a key method for finding anomalous patterns in large databases (Ng in press), which are not erroneous data but interesting findings worthy of further attention. The danger of losing useful information warrants further investigation of statistical outliers. By combining our method with GIS software, we can conveniently verify and analyse anomalous values in association with other geographic data layers.

Our method is mainly designed for handling large data sets. To minimize the computational cost, we adopted a super-block based sorting scheme and other efficient algorithms. This effectively reduced the computational cost and made the diagnostic check on a massive data set possible. When the data set under examination is small, say, with fewer than 45 observations, the overall advantage of our methods over conventional methods is minimal. This is partly because the super-block searching algorithm is comparatively more efficient with relatively large data sets, and partly because reliable local robust summary statistics can only be estimated with sufficient observations.

The performance of our method for a specific data set can be adjusted by changing the desired confidence level and other parameters. For example, the degree of robustness of our statistical analysis can be strengthened by increasing the number of the most influential points and triangles to be dropped in the calculation of outlier indices or by increasing the percentage of samples to be trimmed or winsorized in the construction of robust summary statistics. However, the trade-off is a risk of reducing representativeness of statistical estimates. A decrease in the maximum searching distance or a decrease in the minimum number of samples for the calculation of local summary statistics can increase the degree of locality of our method, but the reliability of summary statistics may be attenuated due to the reduced number of samples. Apparently, the specification of these parameters involves user's subjective judgment or even intuition. Any qualitative information or prior knowledge about the data set will assist the selection of appropriate parameters. With no prior knowledge, one is still able to select a set of reasonable parameters through trial and error. In this sense, our approach serves as an exploratory spatial data analysis tool, rather than a deterministic and analytical solution to outlier problems.

Our method was validated using both hypothetic and real data sets. As future work, the performance of our method will need to be further examined with varied parameters through a more controlled experiment. In addition, we assumed that the interpolation residual index and gradient index are random and spatially independent in a small local area based on the theoretical reasoning. The actual strength of spatial autocorrelation of these two indices within a local area is also worth further investigation.

Acknowledgments

The authors want to express their thanks to three anonymous reviewers and the editor whose constructive comments have been very helpful in improving this paper.

References

Al-Daoud, M., and Roberts, S., 1996, Applying efficient techniques for finding nearest neighbours in GIS applications. In *Innovations in GIS 3*, edited by D. Parker (London: Taylor & Francis), pp. 95–103.

Barnett, V., 1983, Principles and methods for handling outliers in data sets. In *Statistical*

Methods and Improvement of Data Quality, edited by D. E. Wright (Orlando, Florida: Academic Press), pp. 131–166.

Barnett, V., and Lewis, T., 1994, *Outliers in Statistical Data*, third edition (Chichester: John Wiley & Sons).

Bentley, J. L., Weide, B. W., and Yao, A. C., 1980, Optimal expected-time algorithms for closest point problems. *ACM Transactions on Mathematical Software*, **6**, 563–580.

Burt, J. E., and Barber, G. M., 1996, *Elementary Statistics for Geographers,* second edition, (New York: Guilford Press).

Carter, J. R., 1988, Digital representations of topographic surfaces. *Photogrammetric Engineering & Remote Sensing*, **54**, 1577–1580.

Chrisman, N. R., 1991, The error component in spatial data. In *Geographical Information Systems: Principles and Applications,* edited by D. J. Maguire, M. F. Goodchild and D. W. Rhind (London: Longman), pp. 165–174.

Clarke, K. C., 1995, *Analytical and Computer Cartography,* second edition (Englewood Cliffs, NJ: Prentice Hall).

Davis, B. C., 1987, Uses and abuses of cross-validation in geostatistics. *Mathematical Geology*, **19**, 241–248.

Deutsch, C. V., and Journel, A. G., 1992, *GSLIB-Geostatistical Software Library and User's Guide* (New York: Oxford University Press).

Felicìsimo, A. M., 1994, Parametric statistical method for error detection in digital elevation models. *Photogrammetric Engineering & Remote Sensing*, **49**, 29–33.

Hannah, M. J., 1981, Error detection and correction in digital terrain models. *Photogrammetric Engineering & Remote Sensing*, **47**, 63–69.

Hodgson, M. E., 1989, Searching methods for rapid grid interpolation. *Professional Geographer,* **41**, 51–61.

Iglewicz, B., and Hoaglin, D. C., 1993, *How to Detect and Handle Outliers* (Milwaukee, Wisconsin: ASQC Quality Press).

Isaaks, E., and Srivastava, R., 1989, *An Introduction to Applied Geostatistics* (New York: Oxford University Press).

Kraus, K., 1994, Visualization of the quality of surfaces and their derivatives, *Photogrammetric Engineering & Remote Sensing*, **60**, 457–462.

Lanter, D. P., and Veregin, H., 1992, A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing*, **58**, 825–833.

Leberl. F. G., 1990, *Radargrammetric Image Processing* (Norwood, Massachusetts: Artech House, Inc.).

Liu, H., 1999, Generation and refinement of a continental scale digital elevation model by integrating cartographic and remotely sensed data: a GIS-based approach. Ph.D. Thesis, Department of Geography, The Ohio State University, Columbus, OH.

Lòpez, C., 1997, Locating some types of random errors in digital elevation models. *International Journal of Geographical Information Science*, **11**, 677–698.

Luceño, A., 1998, Multiple outliers detection through reweighted least deviances. *Computational Statistics & Data Analysis*, **26**, 313–326.

Lunetta, R. S., Congalton, R. G., Fenstermaker, L. K., Jensen, J. R., McGwire, K. C., and Tinney, L. R., 1991, Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing*, **57**, 677–687.

Macedonio, G., and Pareschi, M. T., 1991, An algorithm for the triangulation of arbitrarily distributed points: applications to volume estimate and terrain fitting. *Computers & Geosciences*, **17**, 859–874.

McCullagh, M. J., and Ross, C. G., 1980, Delaunay triangulation of a random data set for is arithmic mapping. *The Cartographic Journal*, **17**, 93–99.

Ng, R. T., 2001, Detecting outliers from large data sets. In *Geographic Data Mining and Knowledge Discovery,* edited by H. J. Miller and J. Han (London: Taylor and Francis).

Patel, K. R., Mudholkar, G. S., and Fernando, I. J. L., 1988, Student's t approximations for three simple robust estimators. *Journal of the American Statistical Association*, **83**, 1203–1210.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 1992, *Numerical Recipes in C: The Art of Scientific Computing*, second edition (Cambridge: Cambridge University Press).

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muehrcke, P. C., 1984, *Elements of Cartography*, fifth edition (New York: John Wiley & Sons).

Shewchuk, J. R., 1996, Triangle: engineering a 2D quality mesh generator and Delaunay triangulator. In *Applied Computational Geometry: Towards Geometric Engineering*, edited by M. C. Lin and G. Manocha (Berlin: Springer-Verlag), Volume 1148 of Lecture Notes in Computer Science, pp. 203–222.

Thapa, K., and Bosseler, J., 1992, Accuracy of spatial data used in geographic information systems. *Photogrammetric Engineering and Remote Sensing*, **58**, 835–841.

Tietjen, G. L., and Moore, R. H., 1972, Some Grubbs-type statistics for the detection of several outliers. *Technometrics*, **14**, 583–597.

Tsai, V. J. D., 1993, Delaunay triangulations in TIN creation: an overview and a linear-time algorithm. *International Journal of Geographical Information Systems*, **7**, 501–524.

Tukey, J. W., and McLaughlin, D. H., 1963, Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1. *Sankhya, Series A*, **25**, 331–352.

Wahba, G., 1984, Surface fitting with scattered noisy data on Euclidean d-space and on the sphere. *Rocky Mountain Journal of Mathematics*, **14**, 281–299.

Wahba, G., and Wendelberger, J., 1980, Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, **108**, 1122–1143.

Watson, D. F., 1985, Natural neighbour sorting. *The Australian Computer Journal,* **17**, 189–193.

Watson, D. F., 1992, *Contouring: A Guide to the Analysis and Display of Spatial Data* (Oxford: Pergamon Press).