*Transactions in GIS*, 1999, 3(2): 119–136

Research Article

# Analytical Modelling of Positional and Thematic Uncertainties in the Integration of Remote Sensing and Geographical Information Systems

W Z Shi
*The Hong Kong Polytechnic University, Hong Kong*

M Ehlers
*University of Vechta Vechta, Germany*

K Tempfli
*International Institute for Aerospace Survey and Earth Sciences (ITC) Enschede, The Netherlands*

**Abstract**
This paper describes three aspects of uncertainty in geographical information systems (GIS) and remote sensing. First, the positional uncertainty of an area object in a GIS is discussed as a function of positional uncertainties of line segments and boundary line features. Second, the thematic uncertainty of a classified remote sensing image is described using the probability vectors from a maximum likelihood classification. Third, the "S-band" model is used to quantify uncertainties after combining GIS and remote sensing data.

## 1 Introduction

Remote sensing has been used to obtain information about the Earth while GIS reflects the need to combine land attribute information with its geometric representation in order to carry out spatial analyses. These two disciplines were developed separately in the past but, in fact, they are complementary. Remote sensing data can be used for GIS-

**Address for correspondence**: W Z Shi, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. Email: lswzshi@polyu.edu.hk

based spatial analysis and data from a GIS can guide the analysis of remotely sensed images. Therefore, integration of these techniques for data acquisition and analysis has become more and more important. A number of researchers have investigated their integration in the last few years (e.g. Ehlers et al 1989, Mace 1991, Star 1991, Dobson 1993, Shi 1994, Shi and Tempfli 1994).

In 1990, the error analysis research group of the National Center for Geographic Information and Analysis (NCGIA) Initiative 12 identified six problems for research on uncertainty in the integration of remote sensing and GIS (Lunetta et al 1991). Of the six problems, three are directly related to the spatial distribution of uncertainty originating from a combination of positional and thematic errors. This uncertainty, specifically in the integration of remote sensing and GIS, is the focus of this research. Three critical issues need to be addressed: (1) the spatial structure of positional uncertainty in GIS originating from independent random error of points; (2) the spatial distribution of thematic uncertainties in classified remotely sensed images originating from the classification procedure; and (3) the combination of positional and thematic uncertainties in the integration of remote sensing and GIS.

The error distribution of points has been well investigated in disciplines such as geodesy and surveying (e.g. Mikhail and Ackermann 1976). The methods for analysing the uncertainties of line segments and area objects have been investigated by several researchers. For example, Perkal developed the epsilon band model for positional errors of a line (Perkal 1956, 1966) that was later utilized by Chrisman (1982) and Blakemore (1984). The epsilon band is constructed as a simple buffer of constant width (epsilon) on either side of a measured line, and the true location of the line is assumed to be contained within the epsilon band. However, there is no provision to describe the distribution of a measured line segment around its true location. Zhang and Tulip (1990) and Caspary and Scheuring (1992) derived the variances in the X and Y directions for an arbitrary point on the line segment based on the law of error propagation. Dutton (1992) and Caspary and Scheuring (1992) used Monte Carlo simulation methods to model the distribution of line segments and other geometric features. The simulation approach, however, cannot describe the results in the form of formulae and thus cannot directly be used in a GIS. Leung (1997) described the distribution of area objects in GIS based on the probability theory.

A prerequisite to combining positional and thematic uncertainties is statistical derivation of the spatial structures and the error distribution for geometric features (Shi and Tempfli 1994). There is no well-developed model available that can be readily applied to combine positional and thematic classification uncertainties. Haemers (1990) investigated the accuracy assessment problems that arise from the integration of GIS and remote sensing; however, the spatial structure of uncertainties following integration is still an open question.

In this paper, we present a rigorous statistical approach for describing the characteristics of the positional uncertainty of the geometric features in GIS. These include line segments, line features, boundary lines and area objects. The properties of the latter features are determined by the constituent former ones. For example, the uncertainty properties of a polygon are determined by those of the component boundary lines.

## 2 Modelling Uncertainties

### 2.1 Positional and Thematic Uncertainties

The analysis of uncertainties is an important practical problem encountered in many GIS applications. For instance, one wants to create an inventory of the land cover areas over a certain region or set of regions, e.g. within a county. The boundary of this area has been digitized in point mode from a map and is available in GIS format. The land cover types are obtained from a classified remote sensing image using a maximum likelihood (ML) classifier (Figure 1). A typical question is: What is the area of each land cover class in this county? A complete answer would not only include the respective size of the areas (e.g. in $km^2$ or number of pixels) but also their respective spatial uncertainties (e.g. the spatial distribution of a certainty parameter).

In this example, two types of spatial data are involved: original GIS data and classified remote sensing data brought into GIS. We assume that the original GIS data only have positional uncertainties while the classified remote sensing data only exhibit thematic uncertainties, and the thematic uncertainty originates from the classification process. We will not discuss possible thematic uncertainties of the GIS data nor include positional uncertainties of remote sensing data when registered to a GIS data layer. Geometric procedures for the rectification and registration of remote sensing imagery are well developed and have been discussed elsewhere (e.g. Ehlers 1997).

### 2.2 Problem Formulation

On a digitized map in GIS (see Figure 1), we know $P(Z_p(\mathbf{X}) \in O_j)$, i.e. the probability that point $Z_p$ belongs to a certain area $O_j$ (e.g. a county). Here we assume that $Z_p$ has

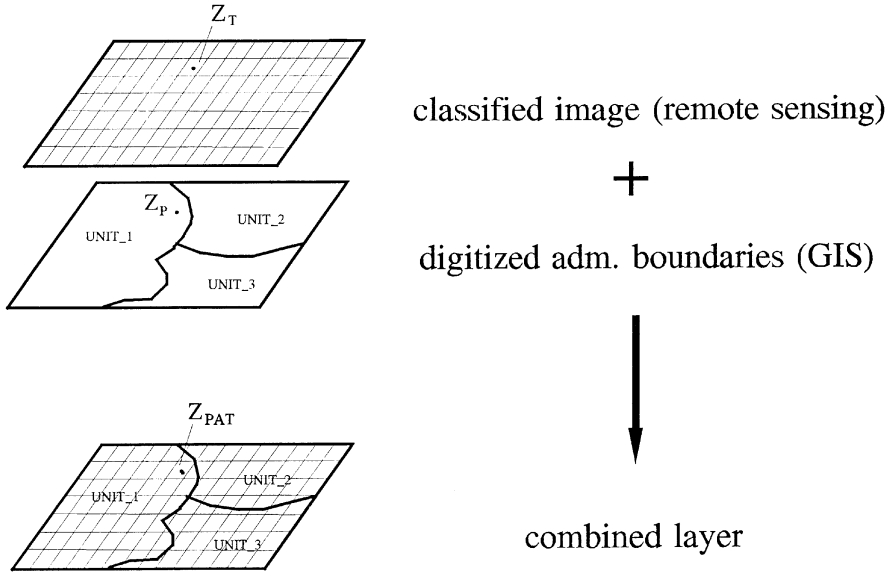


**Figure 1** Land resources inventory using remote sensing and GIS techniques. $Z_T$ is a pixel in the classified image, $Z_P$ is a GIS point which is close to a boundary. $Z_{PAT}$ is a point in the combined layer with the same location as $Z_P$ and $Z_T$.

no thematic error and we refer to this probability as the positional uncertainty indicator. $\mathbf{X}$ is a vector in a two-dimensional Euclidean space, normally $\mathbf{X} = (X, Y)^T$, which determines the geometric location of point $Z_p$.

Typically, a remote sensing image is classified using an ML classification. For a given pixel $Z_T(\mathbf{X})$ which is geometrically located at $\mathbf{X}$, its thematic characteristics are determined by its position in an n-dimensional feature space, where n is the number of spectral bands of the remotely sensed image. Using ML classification techniques, the probability that this pixel belongs to a specific class $C_i$ (i.e. $P(Z_T(\mathbf{X}) \in C_i)$) can be calculated (Richards 1986). $C_i$ is one class type of the whole class category set. This set is usually pre-defined in a supervised ML classification procedure. The probability value per class can be used as a thematic uncertainty (or certainty) indicator.

Using this notation, we can now define our problem as follows: After combining the classified remotely sensed image and the GIS boundary layer, what is the probability that a point which is located at $\mathbf{X}$ belongs to $C_i$ and $O_j$, i.e. $P((Z_T(\mathbf{X}) \in C_i) \wedge (Z_p(\mathbf{X}) \in O_j)) = ?$

To solve this problem, three aspects have to be investigated: (1) the positional uncertainty of an area object, (2) the thematic uncertainty of a classified remote sensing image, and (3) the combination of positional and thematic (PAT) uncertainties.

### 2.3 *PAT Uncertainty, Fuzzy Boundary and Interior Regions*

A PAT uncertainty indicator is defined as an uncertainty indicator which models the integrated positional and thematic uncertainties. In order to describe the uncertainty of a two-dimensional object (e.g. area feature) in a vector-based GIS, we need to distinguish two regions: (1) the fuzzy boundary region, and (2) the interior region (Figure 2).

The difference between interior and boundary regions is based on positional uncertainty. An object in a vector-based GIS is built of line segments. The error of the

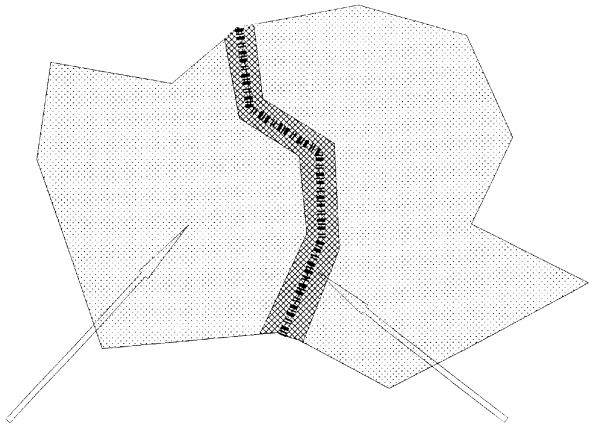


**Figure 2** Fuzzy boundary and interior regions. The positional uncertainties of boundary points affect the fuzzy boundary region and have no effect on the interior region.

endpoints of these line segments (or vertices of the area object) directly affect the positional uncertainty of the object boundary. The region which is affected by boundary errors is called the fuzzy boundary region. The interior region is the one where the effect of the positional uncertainties of its vertices is smaller than a certain threshold (e.g. larger than $3\sigma$ standard deviation for a normal distribution) and can be ignored. The interior region is influenced mainly by thematic uncertainties which originate from errors in the classification process. Therefore, the distribution of spatial uncertainty can be derived by probability methods commonly used in remote sensing classification. In the fuzzy boundary region, however, both positional and thematic uncertainty factors contribute to the overall uncertainty. Consequently, we will focus on the uncertainty in the fuzzy boundary region.

## 3 Modelling Positional Uncertainties in GIS

To model the PAT uncertainty of an object, its positional uncertainty must be addressed first. The basic geometric element of a vector-based GIS object is the point. Two connected endpoints form a line segment. A line feature is composed of two or more line segments. An area object is defined by three or more boundary line features. Thus, we have a hierarchical procedure to build an area object: points, line segments, line features, boundary line features, and area objects.

To describe the nature of the line segment uncertainties, two problems need to be solved. One is concerned with the width of the fuzzy boundary region of a line segment and its shape, and another is the probability distribution of line segments.

### *3.1 Positional Uncertainty of Points*

A point is geometrically defined by a pair of coordinates, e.g. Z(X,Y) in a two-dimensional space. Coordinate errors constitute one of the positional uncertainties in a GIS. The second component is caused by sampling and approximation of a curved line feature by a sequence of straight line segments. This error is directly associated with the curvature of the line and the sampling spacing used. The first component is the focus of this paper.

The coordinates of a point in GIS are usually the result of various measurement and processing steps. Each operation involved adds to the overall error, through blunders, systematic errors and/or random errors. As existing techniques can largely detect systematic errors and blunders, we will only deal here with random errors. If we can express the final coordinates as a function of the original measurements, we can quantitatively determine the error characteristics of a GIS point by applying the laws of error propagation (i.e. propagation of expectations and variances-covariances). We assume that the coordinate errors follow a normal distribution and the errors of any two points are not correlated. Although theoretically manageable, for reasons of simplicity we will investigate the general case of independent error, based on the assumption that the error of points which are digitized in point mode is independent.

We denote two given points that define a line segment by $Z_1 = (X_1,\ Y_1)^T$ and $Z_2 = (X_2,\ Y_2)^T$. They represent a stochastic vector, following a bi-normal distribution $(N_2)$:

$$Z_1 = \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} \sim N_2\left[\begin{bmatrix} \mu_1 \\ \nu_1 \end{bmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix}\right] \quad (1)$$

$$Z_2 = \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} \sim N_2\left[\begin{bmatrix} \mu_2 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix}\right] \quad (2)$$

where $\mu_1$, $\mu_2$, $\nu_1$, $\nu_2$, $\sigma_{XX}$, $\sigma_{YY}$, $\sigma_{XY}$ and $\sigma_{YX}$ are parameters of the two endpoints. Further, we assume equal covariances at the two endpoints of a segment ($\sigma_{XY} = \sigma_{YX}$). The expectation values or true values of the points $Z_1$ and $Z_2$ are $\zeta_1 = (\mu_1, \nu_1)$ and $\zeta_2 = (\mu_2, \nu_2)$, respectively. In the case where we have more than one meassurement for each point, these parameters can be estimated from the measurements. However, if we (typically) only have one measurement for each point, these parameters have to be estimated from test data – points we believe to have the same nature and magnitude of error.

### 3.2 *Positional Uncertainty of Line Segments*

#### 3.2.1 *Definition of a Line Segment*

A line segment is defined by two endpoints $Z_1$ and $Z_2$. By introducing $r = l_r/l (r \in [0, 1])$ we can describe an arbitrary point $Z_r$ on the straight line between $Z_1$ and $Z_2$ (Figure 3):

$$Z_r = (1 - r)Z_1 + rZ_2 \qquad \text{for } 0 \leq r \leq 1 \quad (3)$$

$Z_r$ is a linear function of $Z_1$ and $Z_2$ and also has a bi-normal distribution with the following properties:

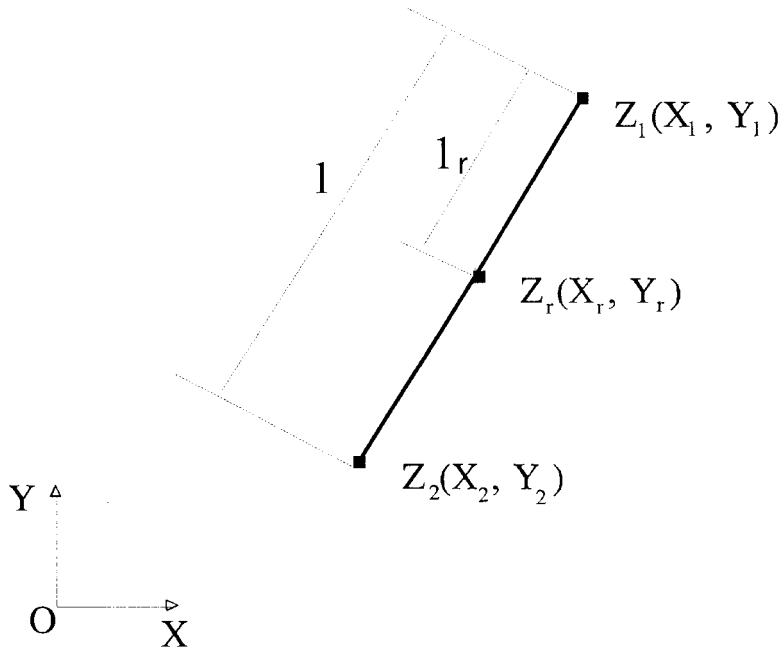


**Figure 3** Definition of a line segment of length l. A line segment $Z_1Z_2$ is defined by two endpoints $Z_1$ and $Z_2$. $Z_r$ is an arbitrary point on the line segment with a distance $l_r$ from $Z_1$.

$$Z_r \sim N_2\left[\begin{bmatrix}(1-r)\mu_1 + r\mu_2\\ (1-r)\nu_1 + r\nu_2\end{bmatrix}, ((1-r)^2 + r^2)\begin{bmatrix}\sigma_{XX} & \sigma_{XY}\\ \sigma_{XY} & \sigma_{YY}\end{bmatrix}\right] \quad (4)$$

*3.2.2 Perpendicular Distribution*

The probability density function of a point in an arbitrary direction can be derived from the marginal probability density of a bivariate probability density function. Since we are interested in the probability distribution in the direction perpendicular to the line segment, we first rotate the original O-XY coordinate system to the O-$X'Y'$ such that the $X'$ axis is parallel to the line segment $\zeta_1\zeta_2$ (see Figure 4). $Z'_r$, the transformed random vector of $Z_r$, is again normally distributed. The marginal probability density of $Z'_r$ in the $Y'$ direction is used to describe the perpendicular distribution (Shi 1994)

$$\begin{aligned} f'_Y(y') &= \int_{-\infty}^{\infty} f'(x', y')dx' \\ &= \frac{1}{(2\pi)^{1/2}(\sigma'_{yy})^{1/2}} \exp[-(y' - \nu'_r)^2/2\sigma'_{yy}] \end{aligned} \quad (5)$$

where

$$\begin{aligned} v'_r &= -\sin(\theta)[(1-r)\mu_1 + r\mu_2] + \cos(\theta)[(1-r)v_1 + rv_2] \\ \sigma'_{yy} &= [A(-\sin(\theta)) + B(\cos(\theta))]((1-r)^2 + r^2) \\ A &= \cos(\theta)\sigma_{XY} - \sin(\theta)\sigma_{XX} \\ B &= \cos(\theta)\sigma_{YY} - \sin(\theta)\sigma_{XY} \end{aligned}$$

$\theta$ is the rotation angle from the $O - XY$ to the $O - X'Y'$ coordinate system.

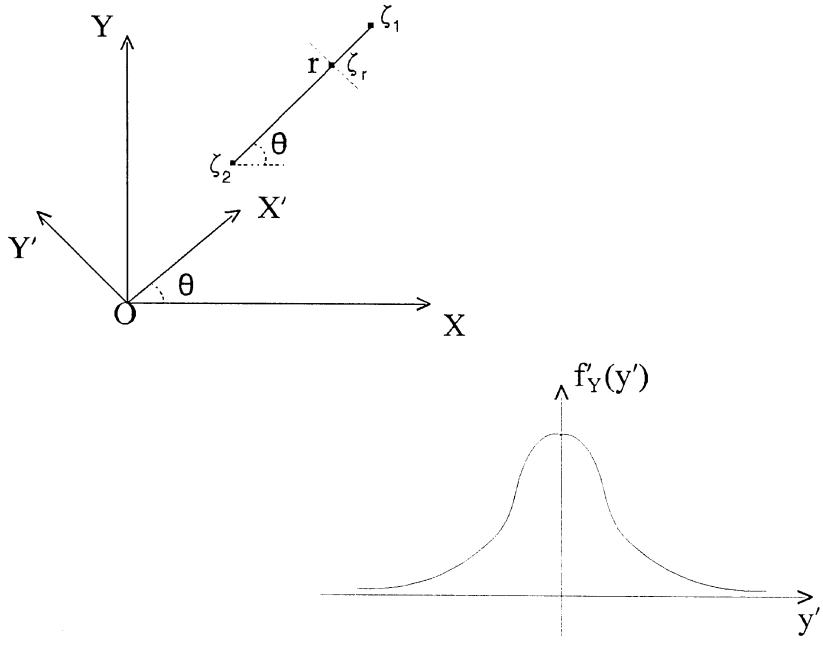


**Figure 4** Perpendicular distribution of an arbitrary point on the line segment.

*3.2.3 Probability Distribution of Line Segments*

The distribution of the line segment $Z_1Z_2$ can be characterized by three density functions: the perpendicular density (see Equation 5) for any $r \in (0, 1)$ and the densities at the two endpoints $Z_1'$ and $Z_2'$. The probability density function of $Z_t'$ ($t = 1$ or 2) can be written as:

$$f_t'(x', y') = \frac{1}{(2\pi)|\Sigma_t'|^{1/2}} \exp\left(-\frac{1}{2}(Z' - E(Z_t'))^T(\Sigma_t')^{-1}(Z' - E(Z_t'))\right) \quad (6)$$

where

$$Z_t' = \begin{bmatrix} X_t' \\ Y_t' \end{bmatrix}, R = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

$$\Sigma_t' = \begin{bmatrix} \sigma_{XX}' & \sigma_{XY}' \\ \sigma_{YX}' & \sigma_{YY}' \end{bmatrix} = R \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{bmatrix} R^T$$

$$E(Z_t') = \begin{bmatrix} \mu_t' \\ \nu_t' \end{bmatrix} = R \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}.$$

The density surface of line segment $Z_1Z_2$ is determined by Equations 5 and 6 and is presented in Figure 5. The probability distribution of a line segment describes how an actual segment $Z_1Z_2$, which is composed of four random variables, normally deviates from the true location $\zeta_1\zeta_2$.

Based on a simplified model (equations 5 and 6), the probability distribution of a line segment is illustrated in Figure 6. The width of the distribution regions is enlarged by 400% for demonstration purposes. For reasons of faster computer calculations, a

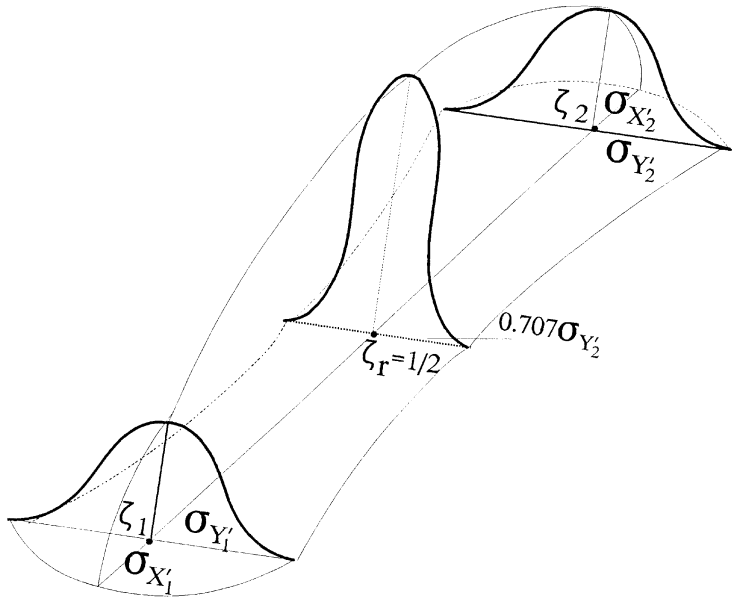


**Figure 5** Probability distribution around the true location $\zeta_1\zeta_2$.

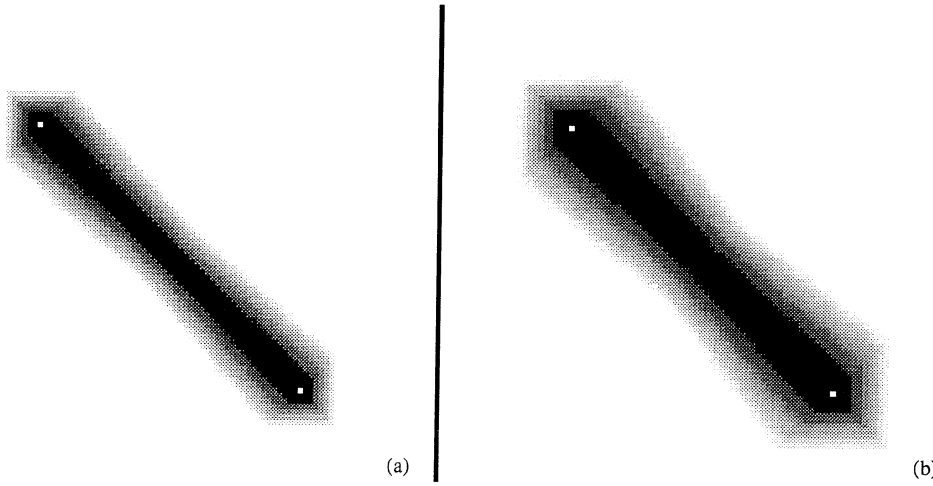


**Figure 6** The probability distribution region of a line segment varies with the variances of the two endpoints and the segment error indicators that are used.

triangular density function instead of a Gaussian one was used. In the diagram, the darker densities represent higher probabilities that a measured line segment is actually located in that zone. The white dots are the "true" locations of the points. Variances of points in (a) and (b) are 2.0 and 3.0, respectively.

*3.2.4 Confidence Region of a Line Segment*

We can now define a region around the measured segment $Z_1Z_2$ so that the true location of this line segment is included with a predefined confidence level. The derivation of the confidence region is based on the distribution of an arbitrary point on the line segment. We construct $J_r$ so that it contains $\zeta_r$ with a predefined confidence level $\gamma$, while all other $\zeta$ of the line segment is contained in their respective confidence regions. This involves an upper bound condition, leading to the inequality:

$$P(\zeta_r \in J_r, r \in [0,1]) > \gamma. \tag{7}$$

The confidence region J of a line segment is the union of the sets $J_r$ for all $r \in [0,1]$. One region $J_r$ is a set of points $(x, y)^T$ satisfying:

$$\begin{aligned} X_r - c \le x \le X_r + c \\ Y_r - d \le y \le Y_r + d, \end{aligned} \tag{8}$$

where

$$\begin{aligned} c &= k^{1/2}[((1-r)^2 + r^2)\sigma_{XX}]^{1/2} \\ d &= k^{1/2}[((1-r)^2 + r^2)\sigma_{YY}]^{1/2} \end{aligned} \tag{9}$$

The parameter k depends on the selected confidence level $\gamma$ and can be obtained from a chi-square table, $k = \kappa^2_{2;(1+\gamma)/2}$. For example, for $\gamma = 0.90$, $(1+\gamma)/2 = 0.95$, $k = 5.99$. A detailed derivation can be found in Shi (1994).

It is easy to verify that the maximum value of $[(1-r)^2 + r^2]^{1/2}$ in Equation 9 occurs if $r = 0$ or 1, whereas the minimum value is at $r = 0.5$. This means that the

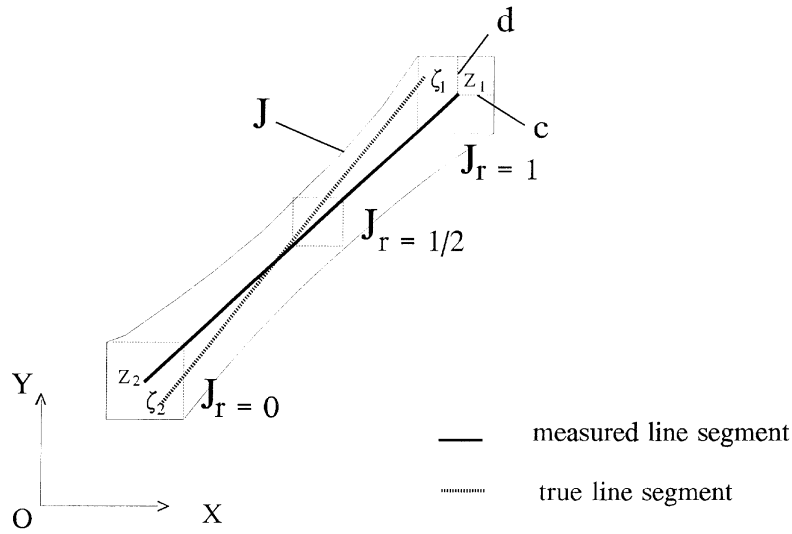


**Figure 7** Confidence region of a line segment.

confidence region is smallest at the centre of the line segment and largest at the endpoints (Figure 7). We can state, therefore, that for a given statistical confidence level the true line segment lies somewhere inside the confidence region. Hence, the confidence region determines the fuzzy boundary region. The area covered by the confidence region depends on the variance-covariance matrices of the two endpoints and the pre-defined confidence level.

### *3.3 Positional Uncertainty of Line Features and Boundary Line Features*

A line feature is usually composed of several line segments. In describing positional uncertainties of boundary line features, two problems need to be solved: the confidence region of boundary line features and their probability distribution. The confidence region of a line feature can be constructed by the union of the confidence regions of the constituent line segments. It provides an uncertainty zone of the spatial extension of a line feature.

One of the major problems in describing positional error distributions of line features is to understand the nature of the uncertainty in the region where two line segments join (see Figure 8). Within this region, the probabilities that a given point Q belongs to an object A are made up of two parts, the uncertainty distribution of line segment $L_1$ and $L_2$. They are denoted by $P_1(Q \in A)$ and $P_2(Q \in A)$ respectively. To obtain the overall probability, we need the combined uncertainty distribution $P_{1\wedge 2}(Q \in A)$ of the line feature $L_{12}$ which is composed of $L_1$ and $L_2$ (Figure 8).

Fuzzy set theory is used to resolve this problem. To apply fuzzy set theory, we need to treat the probability values as corresponding membership values. For example, the probability that Q belongs to object A is treated as the membership value that element Q belongs to a fuzzy set A. The reason that we can follow this approach is that the subjective interpretation of probability considers probability as a measure of belief. Thus, we can state:

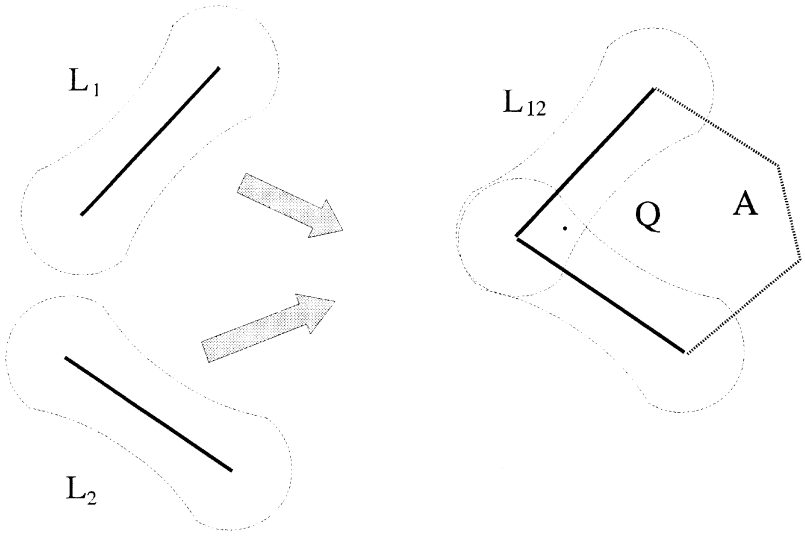


**Figure 8** The uncertainty of a line feature $L_{12}$ is determined by the uncertainties of line segments $L_1$ and $L_2$ and by considering the uncertainty of points in the joint region (e.g. point Q). The line feature $L_{12}$ is part of the boundary of polygon A.

$$P_{1\wedge 2}(Q \in A) = \min[P_1(Q \in A), P_2(QinA)] \tag{10}$$

which means we can use the minimum operation between the probabilities for segments $L_1$ and $L_2$ within the joint region of two line segments. Accordingly, we can generate the uncertainty value for the composed line feature (Figure 9).

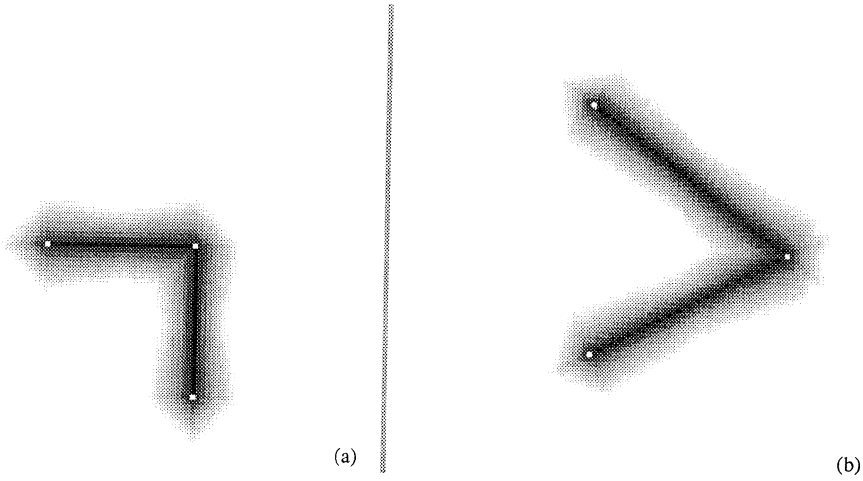


**Figure 9** Uncertainty distribution of lines that are composed of two line segments. For all cases, the probability distribution region is chosen as 95% of the total region; $\sigma_{xy} = \sigma_{yx} = 0$ and $\sigma_{xx} = \sigma_{yy} = (3.0)^{1/2}$ for both endpoints. The white dots are vertices of the lines. The grey values represent the probabilities for the location of a measured line feature. The darker the grey values, the higher the probability values.

### 3.4 Positional Uncertainty of Area Objects

An area object is defined as an area enclosed by a boundary line feature. The positional uncertainty of an area object is determined by that of the boundary. As shown above, the positional uncertainty affects mainly the fuzzy boundary region. The positional uncertainty of an area object is described by the probability that a point (x, y) belongs to the area object (O), i.e. $P((x, y) \in O) \in [0, 1]$. When a point "moves" from the outside to the interior region of the area object, the probability changes from 0 to 1. The probability value of a point in the boundary region is dependent on the probability distribution of the boundary line feature and is determined by the cumulative probability function perpendicular to the boundary. Figure 10a is an example for the uncertainty of an area object. The grey values represent the probability that a point at this location belongs to the object. Darker values indicate higher probabilities. The white dots are vertices of the lines.

#### 3.4.1 Comparison with Epsilon Band-based "Point-in-polygon" Description

Blakemore (1984) used the epsilon band model to describe the "point-in-polygon" problem, i.e. the uncertainty of an area object enclosed by a polygon. He distinguished five relationships between a point and the area object (Figure 10b). These are: 'definitely in' (point 5), 'definitely out' (point 1), 'possibly in' (point 4), 'possibly out' (point 2) and 'ambiguous' (point 3). Using the epsilon band, only five different qualitative relationships between an area object and a point can be provided.

Using the probability distribution of line segments proposed in this paper, we can describe the relationships between a point and an area object by probability values varying continuously within [0, 1] (Figure 10a). This approach provides a quantitative indicator of uncertainty and, moreover, facilitates the combination with thematic uncertainty indicators. We can also characterize the positional uncertainty of the area object by computing a probability frequency distribution. For example, with 10 probability interval classes (i.e. 0–10%, >10–20%, ..., >90–100%), we can calculate the result of Table 1 for the area object in Figure 10a, which is a catchment of the study area. Of a total of 1638 pixels, 649 (i.e. about 40% of the area) have a probability of less than 90% that they belong to the area object. The rate (40%) is dependent on the

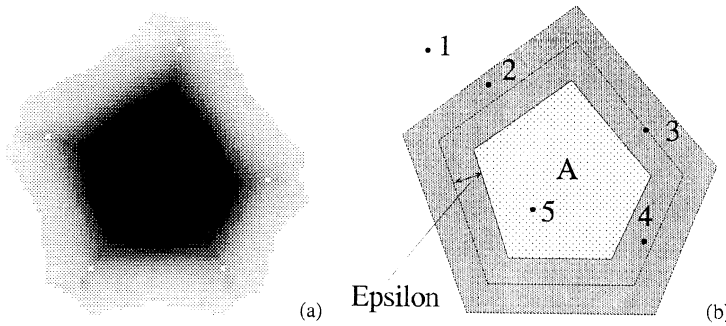


**Figure 10** "Point-in-polygon" description of uncertainty for an area object. For this developed uncertainty model (a), the uncertainty values vary continuously within [0,1]; for the epsilon band model (b), uncertainty is distinguished only in five qualitative levels.

**Table 1** Frequency distribution of probability values of Figure 10a. In this table, the row "Prob." shows the intervals of probability values in percent; "Num." is the number of pixels located within a certain interval (e.g. 20 pixels have a probability value between 40 and 50%); "Sum" is the total number of pixels in the study area. From this table, one can see the positional uncertainty of the area object indicated by the numbers of pixels located within each probability interval.

| Prob. | 0–10 | > 10–20 | > 20–30 | > 30–40 | > 40–50 | > 50–60 | > 60–70 | > 70–80 | > 80–90 | > 90–100 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Num | 0 | 0 | 0 | 0 | 20 | 107 | 154 | 170 | 198 | 989 | 1638 |

error of the vertices and the size of the area object. If the error of the vertices is relatively small compared to the size of the area object, the rate will be significantly lower than 40%.

## 4 Modelling Thematic Uncertainties of a Classified Image

Thematic uncertainty, in this paper, refers to the thematic uncertainty inherent in a classification derived from a remote sensing image. We will make use of the probability vectors in the well developed ML classification technique as a basic thematic uncertainty indicator. The parameters used for classification in this technique are estimated from training samples, and then a probability vector is calculated for each pixel in the image defining the likelihood of specific class membership. The pixel is then assigned to the class with the maximum probability (Richards 1986). For example in an image with four classes (urban, water, forest grass), a pixel with the probability vector:

$$[\mathrm{P}(Z_{\mathrm{P}}(\mathbf{X}) \in \text{urban}) = 0.33, \mathrm{P}(Z_{\mathrm{P}}(\mathbf{X}) \in \text{forest}) = 0.31,$$
$$\mathrm{P}(Z_{\mathrm{P}}(\mathbf{X}) \in \text{grass}) = 0.31, \mathrm{P}(Z_{\mathrm{P}}(\mathbf{X}) \in \text{water}) = 0.05]$$

will be assigned to the class ''urban''. The other probability values are usually ignored. In the above case, however, there is only very weak evidence that this pixel actually belongs to the class urban (the probability is only 33%). If the maximum probability value for each pixel is retained, the certainty of the classification result can be described. If we attach the probability value $\mathrm{P}(Z_{\mathrm{P}}(\mathbf{X}) \in \text{urban}) = 0.33$ to the classification result, it is easy to see that this classification is very uncertain. If the whole probability vector could be attached, a user may further learn that the pixel may just as well be forest or grass (both with probabilities of 31%).

To demonstrate the effects of thematic uncertainties based on ML classification techniques, we will use a test image of Mongolia with four derived classes as described above. We have classified a subset of a Landsat TM scene based on ground truth evidence that was available for this data set using ML. Figure 11a shows the maximum probability value of each pixel as a shade of grey. Darker grey values indicate higher probabilities. Figure 11b shows the classified image based on the maximum probability value for each pixel. The four grey levels (from dark to light) indicate four classes: urban, grass, water and forest. To compute the frequency distribution of probability values, we use ten intervals (0–10%, > 10–20%, ..., > 90–100%). The quantitative results are listed in Table 2.

**Table 2** Maximum likelihood classification and thematic uncertainty expressed as frequency distribution of maximum probability values. In this table, the column "Sum" shows the total area in pixels that were classified as 'urban', 'grass', etc. The table also shows the distribution of each class within the probability intervals, thus describing the uncertainty of the classification.

| Prob. | 0–10 | > 10–20 | > 20–30 | > 30–40 | > 40–50 | > 50–60 | > 60–70 | > 70–80 | > 80–90 | > 90–100 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Urban | 0 | 0 | 56 | 13 | 3 | 9 | 7 | 5 | 11 | 13 | 117 |
| Grass | 0 | 0 | 18 | 17 | 4 | 15 | 7 | 11 | 12 | 138 | 222 |
| Water | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 19 |
| Forest | 0 | 0 | 394 | 57 | 65 | 48 | 65 | 73 | 126 | 452 | 1280 |
| | | | | | | | | | | Total: | 1638 |

Using the visualization technique demonstrated in Figure 11a and the statistical results summarized in Table 2, the user of the classified data not only knows the total area and spatial distribution for each class, but also the certainty (or uncertainty) of this classification. For example, 56 of the total of 117 pixels that are classified as urban have a probability of less than 30%. Thus, we can see the classification result of "urban" is rather uncertain. On the other hand, 138 out of a total of 222 pixels classified as "grassland" have a probability that is higher than 90%, making this result much more certain.

Based on the techniques discussed above, we can now combine positional and thematic uncertainty assessments in the integration of GIS and remote sensing data.

## 5 Modelling the Combined Positional and Thematic Uncertainties

The "S-band" model was developed to combine positional and thematic uncertainties (Shi 1994, Shi and Ehlers 1993). There are two alternatives within the "S-band" model: one is based on the product rule, the other is based on a certainty factor model with probabilistic interpretation. If two data layers are from two different data sources, for example one is from GIS and another is from remote sensing data, they are independent to each other. The product-rule-based approach can thus be used to combine positional and thematic uncertainties. The uncertainty values are within the range [0,1]. For the general case with non-zero correlation between the data layers, we developed a model based on a certainty factor model with probabilistic interpretation. This model is also used in expert system design for uncertainty-based reasoning. With this model, the range of uncertainty expressions is extended from [0, 1] to [−1, 1]. This is particularly important for a reasoning which includes uncertainty indicators covering both positive and negative ranges.

The problem defined in Section 2.2 is about integrating GIS and remote sensing data, and we know that the uncertainties of the two layers' data are independent to each other. We can therefore directly apply the product rule model to calculate the combined positional and thematic (PAT) uncertainty (Figure 12):

$$P((Z_T(\mathbf{X}) \in C_i) \wedge (Z_P(\mathbf{X}) \in O_j)) = P(Z_t(\mathbf{X}) \in C_i)P(Z_P(\mathbf{X}) \in O_j) \qquad (11)$$

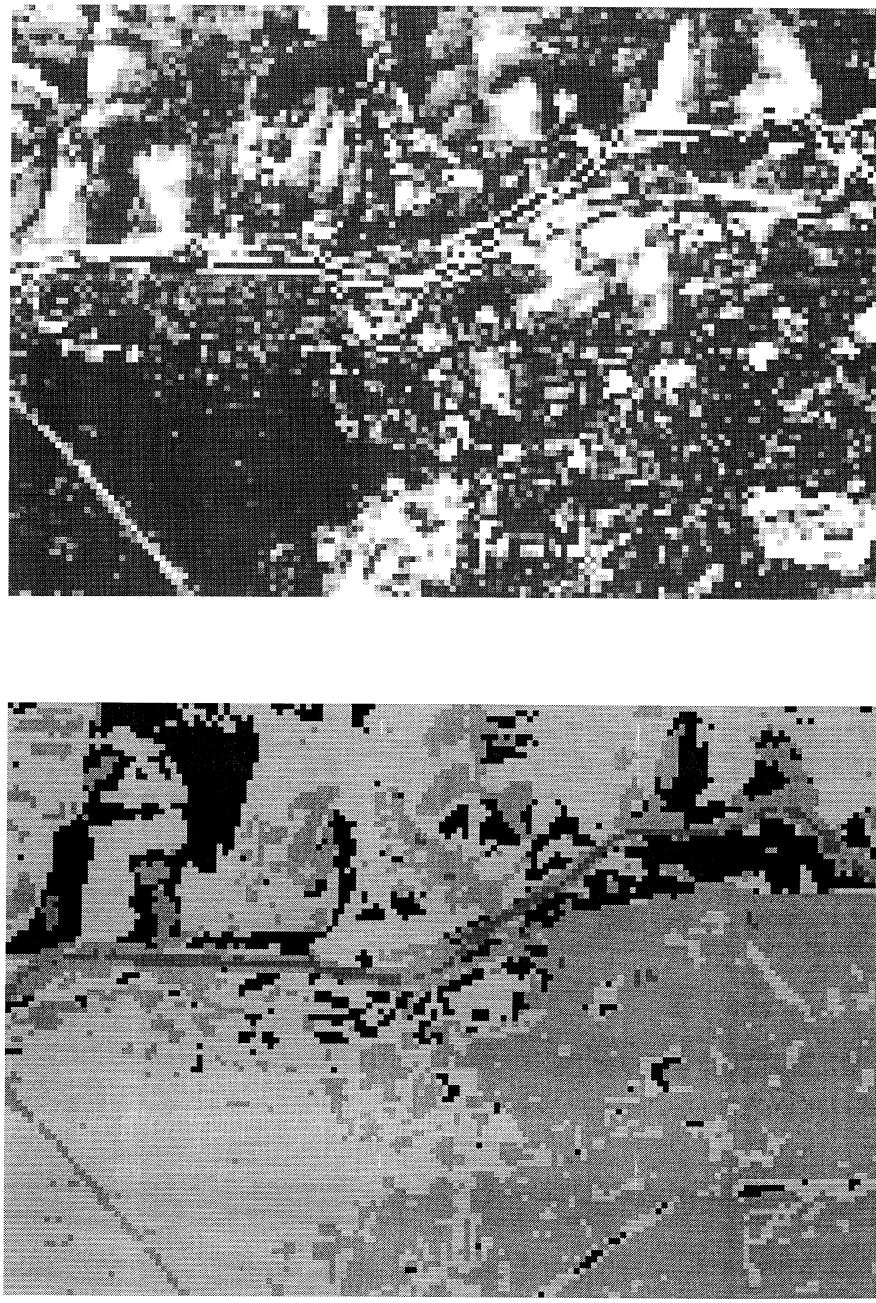

**Figure 11** (a) Maximum probability values. The darker the grey values, the higher are the probability values. (b) Classification result based on the maximum probability values (maximum likelihood classification). The four different grey values (from dark to light) represent four classified land cover classes: 'water', 'urban', 'forest' and 'grass'.

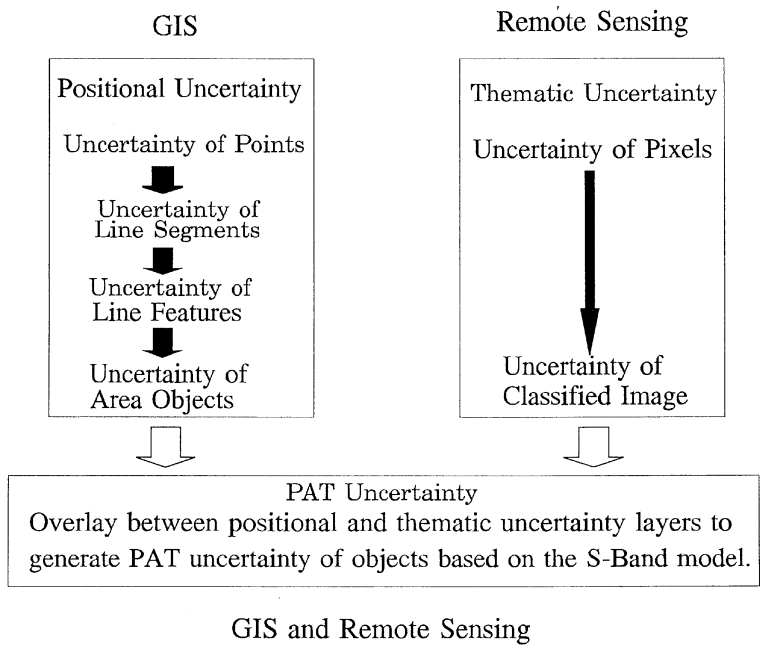


**Figure 12** Diagram of modelling PAT uncertainty of objects using the "S-band" model.

where $P(Z_T(\mathbf{X}) \in C_i)$ is the probability that $Z_T(\mathbf{X})$ belongs to class $C_i$ and $P(Z_p(\mathbf{X}) \in O_j)$ is the probability that point $Z_P(\mathbf{X})$ belongs to area object $O_j$.

Given a boundary layer with positional uncertainty indicators and a classified remote sensing image with thematic uncertainty indicators, an overlay operation in GIS can be used to solve the uncertainty combination problem based on Equation 11. The combined uncertainty for the area indicated in Figure 10a is summarized in Table 3. Table 3 combines information about the size of each land cover class within the area and the uncertainty that is associated with this class. For example, the size of the land cover type 'forest' within the test area is 1280 pixels of which 80 pixels have a certainty between 10 and 20%, 344 pixels between 20 and 30%, . . ., and 240 pixels between 90 and 100%.

One result that is evident from a comparison of Tables 2 and 3 is that after the PAT combination more pixels have lower probability values. For example, within the

**Table 3** The statistics of a classified image with PAT uncertainty indicators. In this table, the column "Sum" is the total area in pixels that were classified as a certain land cover class. The table also describes the PAT uncertainty of each land cover class by providing the number of pixels within each probability interval.

| Prob. | 0–10 | > 10–20 | > 20–30 | > 30–40 | > 40–50 | > 50–60 | > 60–70 | > 70–80 | > 80–90 | > 90–100 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Urban | 0 | 24 | 42 | 7 | 7 | 15 | 12 | 4 | 2 | 4 | 117 |
| Grass | 0 | 9 | 19 | 13 | 18 | 26 | 25 | 29 | 29 | 57 | 222 |
| Water | 0 | 4 | 8 | 0 | 2 | 4 | 1 | 0 | 0 | 0 | 19 |
| Forest | 0 | 80 | 344 | 62 | 95 | 93 | 109 | 133 | 124 | 240 | 1280 |
| | | | | | | | | | | Total: | 1638 |

interval of 90–100%, the number of pixels for the class 'urban' is reduced from 13 to 4, for 'grass' 138 to 57, for 'water' 7 to 0 and for 'forest' from 452 to 240. On the other hand, the number of pixels with high uncertainty values is increased. For example, within the interval 10–20%, the number of pixels for 'urban' is increased from 0 to 24; for 'grass' 0 to 9; for 'water' 0 to 4 and for 'forest' from 0 to 80. With the PAT uncertainty, we can get a quantitative description of the extent to which the overall uncertainty is increased by the combination of positional and thematic uncertainties.

## 6 Conclusions

In this paper, an analytical approach was presented for handling uncertainties in the integration of GIS and remote sensing. Within the five aspects of uncertainty – positional, thematic, temporal, topological and completeness – this paper addressed the first two types of uncertainty in the context of GIS and remote sensing integration.

The "S-band" model to combine positional and thematic uncertainties within the framework of probability theory and certainty factor modelling was described. A procedure for modelling positional uncertainties of area objects in GIS was outlined based on the cumulative uncertainties of defining line segments, line features, boundary line features, and finally area objects. Within this procedure, modelling the uncertainties of line segments was the most fundamental aspect. Two models were developed for line segments based on the confidence regions and probability distributions of line segments. The ML probability vector was then used to describe thematic uncertainties of classified remote sensing data.

In modelling uncertainties concerned with the integration of GIS and remote sensing, there are a number issues that can be studied further. Modelling temporal and topological uncertainties is the most essential. The "S-band" model can be further extended to include these uncertainty components.

## References

Blakemore M 1984 Generalization and error in spatial data bases. *Cartographica* 21: 131–9

Caspary W and Scheuring R 1992 Error-band as measurers of geographic accuracy. In *Proceedings of European GIS'92*. Utrecht, EGIS Foundation: 226–33

Chrisman N R 1982 A theory of cartographic error and its measurement in digital data base. In *Proceedings of Auto-Carto 5*. Bethseda, MD, American Congress of Surveying and Mapping: 159–68

Dobson J 1993 Commentary: A conceptual framework for integrating remote sensing, GIS, and geography. *Photogrammetric Engineering and Remote Sensing* 59: 1491–6

Dutton G 1992 Handling positional uncertainty in spatial databases. In Bresnahan P, Corwin E, and Cowen D (eds) *Fifth International Symposium on Spatial Data Handling*. Charleston, SC, International Geographical Union: 460–9

Ehlers M, Edwards G, and Bedard Y 1989 Integration of remote sensing with Geographic Information Systems: A necessary evolution. *Photogrammetric Engineering and Remote Sensing* 55: 1619–27

Ehlers M 1997 Rectification and registration. In Star J and Estes J E (eds) *Integration of GIS and Remote Sensing*. Cambridge: Cambridge University Press

Haemers P B M 1990 Integration of GIS and remote sensing data: Accuracy assessment. In *Proceedings of European GIS'90*. Utrecht, EGIS Foundation: 428–36

Lueng Y 1997 *Intelligent Spatial Decision Support Systems*. New York, NY, Springer-Verlag

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 data integration: Error sources and research issues. *Photogrammetric Engineering and Remote Sensing* 57: 677–87

Mace T H (ed) 1991 Special issue: Integration of remote sensing and GIS. *Photogrammetric Engineering and Remote Sensing* 57: 641–97

Mikhail E M and Ackermann F 1976 *Observations And Least Squares*. New York, IEP

Perkal J 1956 On epsilon length. *Bullétin de l'Academie Polonaise des Sciences* 4: 399–403

Perkal J 1966 On the length of empirical curves. Ann Arbor, MI, Michigan Inter-University Community of Mathematical Cartographers Discussion Paper No 10

Richards J 1986 *Remote Sensing Digital Image Analysis: An Introduction*. New York, NY, Springer-Verlag

Shi W and Ehlers M 1993 "S-band", a model to describe uncertainty of an object in an integrated GIS/remote sensing environment. In *Proceedings of IGARSS'93, Tokyo, Japan*: 1721–3

Shi W 1994 Modelling Positional and Thematic Uncertainty in Integration of GIS and Remote Sensing. Enschede, ITC Publication No 22

Shi W and Tempfli K 1994 Positional uncertainty of line features in GIS. In *Proceedings of ASPRS/ACSM'94, Reno, NV*. Falls Church, VA, American Congress of Surveying and Mapping: 696–705

Shi W, Ehlers M and Tempfli K 1994 Modelling and visualizing uncertainties in multi-data based spatial analysis. In *Proceedings of European GIS'94*. Utrecht, EGIS Foundation: 454–64

Star J L (ed) 1991 *The Integration of Remote Sensing and Geographic Information Systems*. Bethseda, MD, American Society for Photogrammetry and Remote Sensing

Zhang G Y and Tulip J 1990 An algorithm for the avoidance of sliver polygons and clusters of points in spatial overlay. In *Proceedings of the Fourth International Symposium on Spatial Data Handling*: 141–50