A Random Sets Model for Spatial Objects with Uncertain Boundaries

Xi Zhao[ab], Xiaoling Chen[ac], Alfred Stein[b]
a State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, Luo Yu road 129, Wuhan, HuBei, China
b International Institute for Geo-Information Science and Earth Observation,
Hengelosestraat 99, Enschede, Netherlands
c The Key Laboratory of Poyang lake Wetland and Watershed Research, Jiangxi Normal University,
Ziyang road 99, Nanchang，JiangXi, China

1. INTRODUCTION

Several conceptual models and data models have been proposed for geographical phenomena with uncertainties that cannot easily be forced into current standard data models. These models of uncertain objects can be categorized into two groups. One group (Clementini and Di Felice 1996; Cohn and Gotts 1996; Roy and Stell 2001) considers the broad boundary of an object as a homogeneous two-dimensional region instead of lines. The well-known "Egg-Yolk" model which describes uncertain region as a pair of crisp regions, one enclosing the other, is a typical example. The other group employs probability theory or fuzzy set theory to further represent gradual changes within the broad boundaries. Probabilistic models for depicting the positional uncertainties of these objects have been developed, distinguishing between point features (Thapa and Bossler 1992), line features (Dunn et al. 1990; Shi and Liu 2000) and polygon features (Shi and Wu 2003). Formal definitions of fuzzy regions and topological relations between them are provided by (Zhan 1998; Cheng et al. 2001; Schneider 2003; Dilo et al. 2007). Definitions of random/uncertain data types are also developed (Glemser and Fritsch 1998; Pfoser and Tryfona 2001; Tøssebro and Nygård 2002) and applied in uncertainty modeling applications (Glemser and Klein 2000). They applied probability theory directly on those spatial data types. However, data analysis when the data are sets, rather than points in spatial-temporal space, is not a nice situation to work with (Nguyen 2006). Therefore, there is a need to propose a general framework for set-valued observations.

This paper aims to implement random sets based on probability theory in a GIS environment for handling regions with uncertain boundaries. A random sets data model is proposed to represent uncertain geographic regions and applied to spatial-temporal modeling, using a wetland monitoring case as an illustration.

2. UNCERTAINTY MODELING WITH RANDOM SET

The term *random sets* firstly appeared to indicate "region…depending on chance" (Kolmogoroff 1933). After the publication of the fundamental book of random set theory by Matheron (1975), various successful applications in spatial science appeared. Random sets are applicable and useful for general data fusion (Goodman et al. 1997), to address problems of determining the locations and identities of multi-targets from multi-sensor data with noise effects. Moreover, random sets are a basic tool in solving problems of geometrical statistics and image analysis (Molchanov 1998) and time series analysis (Nuñez-Garcia and Wolkenhauer 2002). It is the core of stochastic geometry, in which they are stochastic models of irregular or random geometrical structures. Stoyan (1998) applied random set theory in particle statistics to study shape fluctuations of sand grains, showing how random set as a set-theoretic method has its advantages and may act as a supplement of other powerful means such as multivariate statistics. Nguyen (2006) pointed out that uncertainties caused by subjectivity, impression and vagueness bear some relationship with random sets. More recent developments of random sets are provided by (Molchanov 2005; Nguyen 2006; Illian et al. 2008). In the following sections, basic concepts of random sets with respect to this research are introduced and random spatial data types are defined.

2.1 Random Sets

Sets obtained at random can be considered as kinds of random sets. For example, a probability sampling design is a random experiment whose outcomes cannot be predicted with certainty in advance, so a sample in the surveying sampling is a set obtained at random, i.e. random set. Let (Ω, $\sigma_{\Omega,}$ Pr_{Ω}) be a probability space and (Ξ, $\sigma_{\Xi,}$ Pr_x) a measure space. Every σ_{Ω} - σ_{Ξ} measurable mapping v: Ω - Ξ is called a random variable. The distribution or probability law of v is defined as Pr_v (A) = Pr_{Ω} ($\mathrm{v}^{-1}(A)$) = Pr_{Ω} $\{\omega : \mathrm{v}(\omega) \in A\}$ for $\forall$ $A \in \sigma_{\Xi}$. A random set can be seen as a random variable from Ω to U where U is a set of subsets of Ξ, i.e. $U \subseteq \mathcal{P}(\Xi)$. Let ($U$, $\sigma_{U,}$ Pr_X) be a measurable space and σ_U is a σ- algebra defined on U, then a σ_{Ω} - σ_U mapping X: $\Omega \rightarrow U$ is a random set. It is a multi-valued mapping between the space Ω and Ξ. Its distribution is defined as $Pr_X(\mathcal{A})$ = Pr_{Ω} $(X^{-1}(\mathcal{A}))$ = Pr_{Ω} $\{\omega \mid X(\omega) \in \mathcal{A}\}$ for $\forall$ $\mathcal{A} \in \sigma_U$. When all elements of U are singletons, then random set X becomes a random variable.

On the Euclidean space R^n, a random set X associates a probability value to each element $x \in R^n$, which quantifies how likely it is that x belongs to X. The random set X on R^n is a function P_x: $R^n \rightarrow$ [0,1]. This function is called the *covering function* of the random set, which takes values between 0 and 1. The set $X_a = \{x \in R^n \mid P_x(x) \geq \alpha\}$ is called α-level set of X. The set $X_0 = \{ x \in R^n \mid P_x(x) > 0 \}$ is called the *support* set of X. The set $X_1 = \{ x \in R^n \mid P_x(x) = 1 \}$ is called the *core* set of X. We denote the set of all random sets in R^n by $\Re(R^n)$.

A typical situation where observations are sets rather than points in a sample space can be dealt with random sets. When we obtain data with low quality due to imperfection of acquiring procedure or measuring instruments, it is more informative to represent the observations as subsets containing the true values than to ascribe unique values (Nguyen 2006).

2.2 Random Spatial Data Types

In our study we consider geo-space, i.e. the two-dimensional Euclidean space R^2, with geographic entities as subsets of R^2. Based on probability theory and random set concepts, we propose random objects, denoted *Robject*, to model geographic entities with uncertain characteristics. We model *Robject* as sets of points, homeomorphic to R^2. A *Robject* can be either determinate or indeterminate. A *Robject* is determinate if any of its component points has an exact position that can be mapped onto a single point. A *Robject* is indeterminate, if any of its component points can only be mapped to a set of points, i.e., the exact position is unknown. *Random unit* (R_u) is defined as basic elements of R^2 constructing primary data types. A random unit (R_u), denoted as (x, y, p), consists of a pair of coordinates (x, y) and a probability value indicating how likely this unit located at (x, y). The probability value is derived from probability density function which describes the likelihood for each unit to be at the position. For example, uniform distribution tells us that there is an equal chance for each unit.

The random objects (*Robject*) is defined as:

$$Robject \equiv \{ Ru \in \Re(R^2) \mid \exists \, (x, y) \in R^2 , p(x, y) > 0 \} \qquad (4)$$

Three primary random data types: *random point* (R_p), *random line* (R_l) and *random region* (R_r) are defined below.

The random point (R_p) is defined as a random set in R^2 that contains a finite collection of uncertain units with positive probability values, denoted as R_p (R_{up1}, R_{up2}, …, R_{upn}). The probability value attached with each uncertain unit is the likelihood that the point is in this location. The distribution of all elements in a random point is described by its covering function. The uncertain points with $p_i > 0$ form the support set and with p_i =1 form the core set. It is used for spatially modeling a point-like geographic entity. This point entity can be an accident place that stochastically occurred, a pair of imprecise coordinates received by a GPS receiver, or a building that is recorded by mixed image pixels.

The random line (R_l) is to model a linear object with a random component. We distinguish two ways in representing a random line. The first way is by a collection of uncertain lines associated with their probabilities, denoted as R_l {(l_1, p_1), (l_2, p_2),…, (l_n, p_n)}. For each component -uncertain line l_i, it is constructed by a set of uncertain units with a same probability value p_i, denote as (l_i, p_i). The second way is by a collection of uncertain units that constitute the line, denoted R_l (R_{up1}, R_{up2}, …, R_{upn}). The probability value attached with each uncertain unit is that the likelihood of the line goes through the location of this unit. The distribution of all the elements of a random line is described by its covering function. The elements with $p_i > 0$ form the support set and with p_i =1 form the core set. Random lines are used for spatially modeling a linear geographic entity, like a route from city A to city B, a contour line digitized from a scanned map, or a coastline interpreted from a satellite image.

The random region (R_r) is a region with a random component. It can be modeled in two ways. The first way is by a collection of uncertain regions associated with their probabilities, denoted by R_r {(r_1, p_1), (r_2, p_2),…, (r_n, p_n)}. For each component -uncertain region r$_i$, it is constructed by a set of uncertain units with a same probability value p_i, denote as (r_i, p_i). The second way is by uncertain units which form its random coverage, denoted by R_r (R_{up1}, R_{up2},…, R_{upn}). The probability value attached to each uncertain unit is that the likelihood of the region covers the location of this unit. The distribution of all the elements of a random region is described by its covering function. The areas with $p_i > 0$ form the support set and with p_i =1 form the core set. Random regions are used for spatially modeling areal geographic entities, such as clouds, dunes, field patches and lakes.

2.3 Estimation of the Covering Functions

Since different specified probability measures lead to different random sets, there is a need to look at the distributions of random sets. In general, the question of interest is how to suggest random set models from empirical observations. In previous researches, there are different means of estimating probability distribution functions of random objects. The simplest way is to provide a set of standard functions such as the normal distribution and the uniform distribution. Another alternative is to adopt users' definitions, which are often derived from sampled data. For example, Tøssebro and Nygård (2002) calculated probability of points in uncertain regions as the ratio between distance to support and the sum of distance to support and distance to core. However, the uncertainties often do not conform to normal distributions or are too complex to model by analytical approaches. In this case, we can adopt numerical approaches, such as Monte Carlo method, to obtain the empirical covering function of random sets, like in (Glemser and Klein 2000; Guo et al. 2008). Some researches also estimated uncertainty distribution from fuzzy membership functions (Pfoser and Tryfona 2001; Van de Vlag and Stein 2007).

3. ILLUSTRATIVE EXAMPLES

The case has its study area in the Poyang Lake national nature reserve (PLNNR) (115°55’ -116°03’ E, 29°05’ -29°15’ N), at the southern bank of the middle reach of Yangtze River, central China (Figure 1). Every year, the water level starts to rise in spring (March to May) due to flooding of its source rivers, and keeps its highest level in summer (June to August) as a consequence of flooding of the Yangtze River. The water level reduces from September onwards when the flood in the Yangtze River ebbs to remain stable at its lowest level from December to February. Wetland vegetation submerges in summer and greens up in autumn upon the recession of the Yangtze River, providing desirable habits and high quality forage for waterfowls (Wu and Ji 2002). Mapping the wetlands, especially modeling the form variance of lakes is of great importance for PLNNR managers and decision makers for both ecosystem dynamic monitoring and habitat assessment.

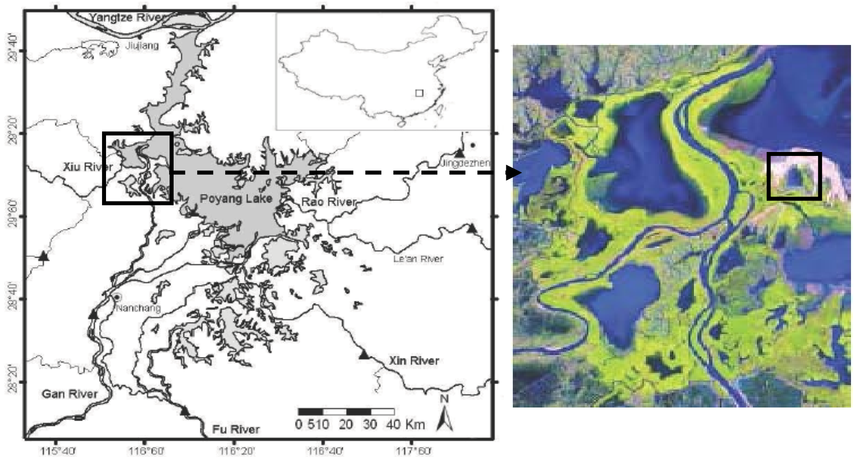


Figure 1: Location of the Poyang Lake and the PLNNR (left) and the study area in the black box (right).

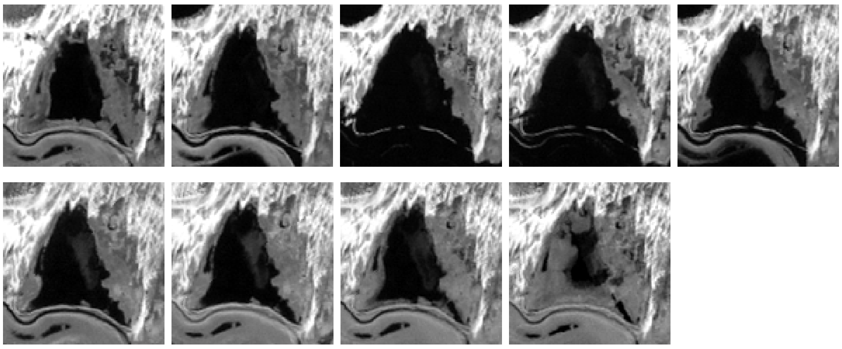


Figure 2: Image object of Meixihu (in black) on the middle Infrared band of nine Landsat TM images of 2004 (The observing time from top to bot-tom and from left to right is 5 May, 22 June, 24 July, 9 August, 26 Sep-tember, 12 and 28 October, 29 November and 15 December)

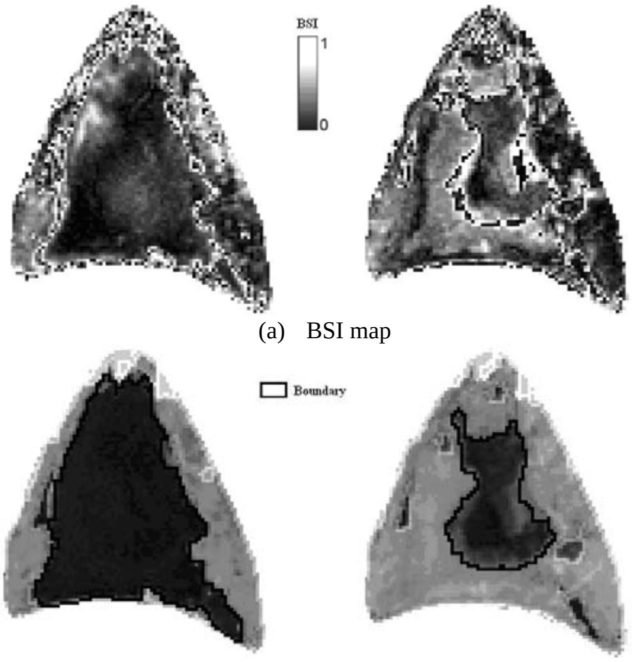
Figure 2 shows a series of Landsat TM images covering the Meixihu Lake within the PLNNR, which were acquired from May to December of 2004 for this wetland monitoring study. A topographic map of scale 1:10000 is used as the geographic reference data. The root mean squared error (RMSE) of all the geometric corrections was less than 10m. To extract image objects – 'Meixihu' from a single image, the split and merge method developed in (Lucieer and Stein 2002) was adopted to segment images and to quantify segmentation uncertainty by using Parbat software (http://parbat.lucieer.net/). In this method, ranges for the splitting thresholds and merging thresholds are chosen and are divided into N steps. At each step, object boundaries, in the form of segment edge pixels, are determined. At step k these boundary pixels are assigned the value 1 and non-boundary pixels the value 0 and are represented on a segment-boundary image. By summing these segmentations, an image with boundary stability index (BSI) for each pixel is generated to indicate the frequency of boundary location (0-1). Higher value of the index means more certainly about the boundary. Objects with a low degree of geometric uncertainty will remain the same at different segmentation levels. Objects with uncertain boundaries will change the shape and size gradually. We set thresholds in split and merge segmentation approach, let Parbat software segment N times, generate a set of N pixel clusters and calculate the BSI for each image pixel.

$$BSI = \frac{\sum_{k=0}^{n} I_k}{n} \qquad (1)$$

The number of segmentation steps n was set to 50 in our experiment. Boundaries in all the 50 segmentation results construct a random line that models the uncertain boundary of Meixihu. An image with a BSI map was calculated to depict object stability which represents the random line by uncertain points (figure 3a). This BSI map also indicates the empirical covering function of the random line. The final extracted boundaries in fig. 3b are constructed by pixels with the local maximum BSI values. Comparing the uncertain distribution in (figure 3a) with segmented boundaries in (figure 3b), we observe fewer pixels with high BSI values around the edge of the object on the left image than on the right image. In other words, the flooded Meixihu object has a thinner certain boundary than the dry Meixihu object. This result is consistent with our previous observation in figure 2 that the water boundary in the flooding season (June) looks less gradual than during the dry season (December). In this case, random set model which is constructed by multiple segmentations presents the uncertainties that may occur in image processing. The gradual object boundary modeled by random line can be used in further image analysis which may ask the inclusion of uncertainty.

We now consider N image objects extracted from a time-series of N images that record the changing process of a dynamic entity. These N image objects construct a random area X for modeling its dynamic shape and size. Every year, the spatial extent of the lake changes with the same trend, but vary due to different environmental conditions, which is a typical stochastic process. In space and time, this Meixihu object can be considered as an object with uncertain boundaries due to its dynamic extents. The N image objects are considered as samples of a random set X. Pixels on each image that belong to one image object are assigned with value 1 and otherwise 0, then summed up and divided by N. This results in an image with object covering index (OCI) to indicate the frequency of object covering extent, i.e. covering function of the random set. Covering frequency in figure 4 can be interpreted as possibility which indicates the membership of the cell to the dynamic object 'flooded region'. All cells with possibility larger than 0 were grouped to indicate the maximum extent of the flooded area.

$$OCI = \frac{\sum_{k=0}^{n} I_k}{n} \qquad (2)$$



(a) BSI map

(b) Object boundary

Figure 3: Meixihu area observed at June-22-2004(left) and December-15-2004(right).

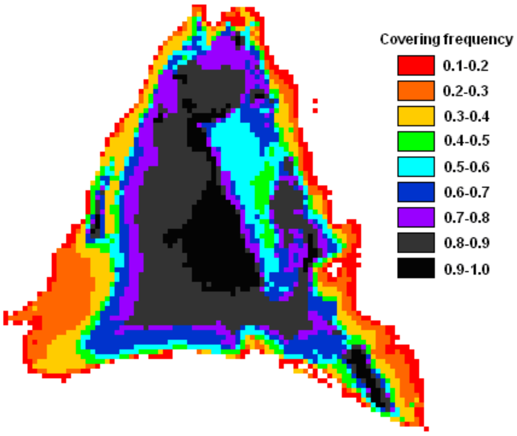
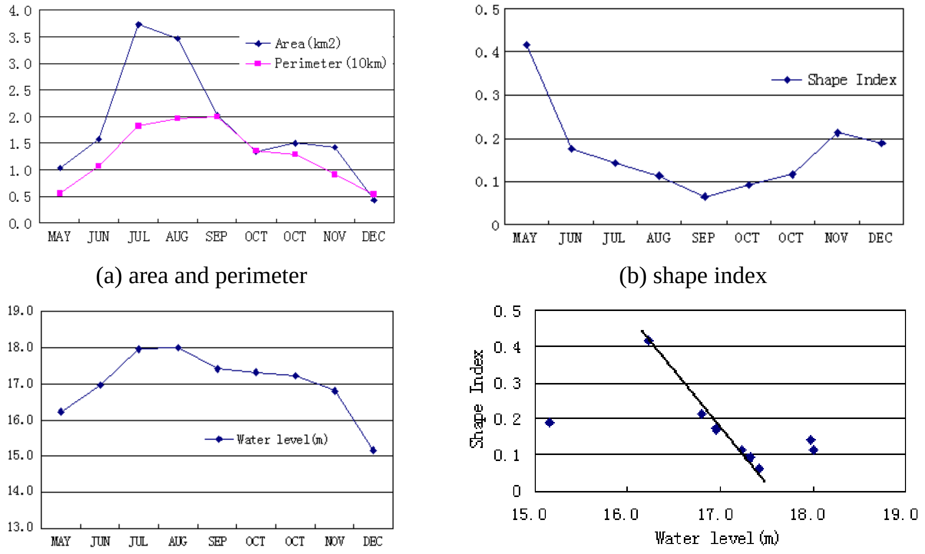


Figure 4: *OCI* map represent empirical covering function of random region modeling dynamic Meixihu Lake.

Random set can be described by other random variables, whose values lie in a space mathematically simpler than the space of random area. Very simple random variables assigned to random set X are its area A(X) and its perimeter U(X). To explore the shape changes, we also need reprehensive shape index, which do not change as scale changes, compactness C(X). Compactness has a maximum value of 1 for a circle. The change of shape and presence of irregular boundaries will decrease the value.

$$C(x) = 4\pi \frac{A(x)}{U(x)^2} \tag{3}$$

To study the geometrical dynamic of Meixihu, several form parameters were derived and analyzed with other driving factor, i.e.water level. From the results shown in figure 5a, we can see the area and perimeter of Meixihu have the same change trend: the values increase rapidly from May to June then to around twice in July, followed by a decrease from September to December. There are two extreme points in the change curve of area index: the peak point in July and the second peak point in August. Compared with water extents in figure 2, we found that it was the topological change of Mixihu causes these abnormal points. The water in Meixihu was overflowed and merged with other water bodies of Poyang Lake in flooding months - July and August. Therefore, when calculating area and perimeter indices, newly flooded areas connected with Meixihu were regarded as parts of Meixihu object and counted in. Figure 5b shows the shape of the Meixihu object is most close to a circle in May.



(a) area and perimeter

(b) shape index

(c) water level

(d) relationship of shape index and water level

Figure 5: Changes of area, perimeter, shape index and water level of Meixihu at nine image observation time.

After the water level beginning to rise up, the object boundary becomes more and more irregular. The lowest compactness appears in September, when Meixihu was connected with a linear-shape object (a river). After flood recession, the lake will be disconnected with other water bodies again and the compactness shape index is expected to return back to the high level as it in May. However, *Carex*-dominated grassland green up on the newly exposure wet soil around the lake, which makes the object shape concave and thus the index still stay low. Figure 5c shows water level of the Meixihu lake, which was derived from daily recorded hydrological data averaged into month unit. We can see obviously, the change curves of both area and perimeter have the same trend with the water level fluctuation, while the curve of compactness shape index has inverse trend. To check the relationship of the water level and shape index, we found six pairs of data are quite well correlated, except the three extremes: the highest water level in July and August and the lowest water level in December. To justify this negative relationship between water level and shape index, more images are necessarily required.

4. DISCUSSION AND CONCLUSIONS

This paper proposed a random sets method to model objects with uncertain boundaries. Random data types: point, line and region were proposed and defined conceptually. We also illustrated examples from a wetland monitoring case, to show these random data types can be applied in image processing. Random set model could be used to present uncertainties of image processing operators and as well as dynamic variation of shape and size. One advantage of the random set model is that it can involve both spatial uncertain objects and spatial-temporal uncertain objects under a uniform framework, i.e. modeling as random point, line, and region instead of adding new time-related data models.

The examples in section 3 are illustrative rather than detailed, and many other real world problems may involve further assumptions and complexities. As a conceptual model proposed in this paper, lower level of data type definition needs to design for databases. Considering the finite representation which can be stored in computer, continuous probability covering function needs to transform and approximate for each object by discrete cells, which forms a probability matrix. In brief, the implementation and interpretation of random sets and its applications to other real cases are necessary.

Random sets are built by overlaying of the interpreted images in this paper, which is only a simple illustration of model construction. But the usability of the random set model is not shown directly through the examples. It is the future work after this paper that to take the benefits from random set models, such as Boolean model, in uncertain modeling.

ACKNOWLEDGEMENTS

This work was funded by the 973 Program (Grant No.2009CB723905), Sino-Germany Joint Project (Grant No. 2006DFB91920), National Key Project of Scientific and Technical Supporting Programs (2007BAC23B05), NSFC project (Grant No. 40721001) and ITC PhD fund. The authors would like to express their many thanks to Dr. David Rossitor for his valuable discussions and the AGILE reviewers for their valuable comments.

BIBLIOGRAPHY

Cheng, T., M. Molenaar and H. Lin, 2001. Formalizing fuzzy objects from uncertain classification results International Journal of Geographical Information Science 15(1): 27-42.

Clementini, E. and P. Di Felice, 1996. An algebraic model for spatial objects with indeterminate boundaries. *Geographic objects with indeterminate boundaries*. P. A. Burrough and A. U. Frank. London, Taylor&Francis**:** 155-170.

Cohn, A. G. and N. M. Gotts, 1996. The 'egg-yolk' representation of regions with indeterminate boundaries. *Geographic objects with indeterminate boundaries*. P. A. Burrough and A. U. Frank. London, Taylor&Francis**:** 171-187.

Dilo, A., R. A. De By and A. Stein, 2007. A system of types and operators for handling vague spatial objects. International Journal of Geographical Information Science 21(4): 397-426.

Dunn, R., A. R. Harrison and J. C. White, 1990. Positional accuracy and measurement error in digital databases of land use: an empirical study. International Journal of Geographical Information Systems 4(4): 385-398.

Glemser, M. and D. Fritsch, 1998. Data Uncertainty in a Hybrid GIS. International Archive of Photogrammetry and Remote sensing, Stuttgart, Germany.

Glemser, M. and U. Klein, 2000. Hybrid modelling and analysis of uncertain data. International Archive of Photogrammetry and Remote sensing, Amsterdam.

Goodman, I. R., R. P. S. Mahler and H. T. Nguyen, 1997. Mathematics of data fusion. Dordrecht, Kluwer Academic.

Guo, Q., Y. Liu and J. Wieczorek, 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science 22(10): 1067-1090.

Illian, J., A. Penttinen, H. Stoyan and D. Stoyan, 2008. Statistical Analysis and Modelling of Spatial Point Patterns (Statistics in Practice). Chichester, Wiley.

Kolmogoroff, A. N., 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlin, Springer-Verlag.

Matheron, G., 1975. Random sets and integral geometry. New York, Wiley.

Molchanov, I., 1998. Grey-scale images and random sets. Proceedings of the fourth international symposium on Mathematical morphology and its applications to image and signal processing Amsterdam, The Netherlands

Molchanov, I., 2005. Theory of random sets London, Springer.

Nguyen, H. T., 2006. An Introduction to Random Sets. Boca Raton, Chapman&Hall/CRC.

Nuñez-Garcia, J. and O. Wolkenhauer, 2002. Random Set System Identification. IEEE transactions on fuzzy systems 10(3): 287-296.

Pfoser, D. and N. Tryfona, 2001. Capturing fuzziness and uncertainty of spatiotemporal objects, Springer-Verlag.

Roy, A. J. and J. G. Stell, 2001. Spatial relations between indeterminate regions. International Journal of Approximate Reasoning 27(3): 205-234.

Schneider, M., 2003. Design and implementation of finite resolution crisp and fuzzy spatial objects. Data & Knowledge Engineering 44: 81-108.

Shi, W. and W. Liu, 2000. A stochastic process-based model for the positional error of line segments in GIS. International Journal of Geographical Information Science 14(1): 51-66.

Shi, W. and H. Wu, 2003. A probabilistic paradigm for handling uncertain objects in GIS by randomized graph algebra. Progress in Natural Science 13(9): 648-657.

Stoyan, D., 1998. Random sets: models and statistics. International Statistical Review 66(1): 1-27.

Thapa, K. and J. Bossler, 1992. Accuracy of spatial data used in geographic information systems. . Photogrammetric Engineering and Remote Sensing 58(6): 835-841.

Tøssebro, E. and M. Nygård, 2002. An Advanced Discrete Model for Uncertain Spatial Data. WAIM, Berlin, Springer-Verlag.

Van de Vlag, D. E. and A. Stein, 2007. Incorporating uncertainty via hierarchical classification using fuzzy decision trees. IEEE Transactions on geoscience and remote sensing 45(1): 237-245.

Wu, Y. and W. Ji, 2002. Study on Jiangxi Poyang Lake National Nature Reserve.

Zhan, B. F., 1998. Approximate analysis of topological relations between geographic regions with indeterminate boundaries. Soft Computing 2: 28-34.