Random sets to model uncertainty in remotely sensed objects

Xi Zhao

PhD dissertation committee

Chair

Prof. dr. ir. A. Veldkamp	University of Twente

Promoter

Prof. dr. ir. Alfred Stein	University of Twente

Assistant promoter

Prof. dr. Xiaoling Chen	Wuhan University

Members

Prof. dr. ir. M.G. Vosselman	University of Twente
Prof. dr. A. Bagchi	University of Twente
Prof. dr. R.J. Boucherie	University of Twente
Prof. dr. E.J. Pebesma	University of Muñster

ITC dissertation number 203
ITC, P.O. Box 217, 7500 AA Enschede, The Netherlands

ISBN: 978-90-6164-328-9
Printed by: ITC Printing Department

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

RANDOM SETS TO MODEL UNCERTAINTY IN REMOTELY SENSED OBJECTS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, April 4, 2012 at 14.45 hrs

by

Xi Zhao
born on April 29, 1984
in Hubei, China

This dissertation is approved by:

Prof. dr. ir. Alfred Stein (Promoter)
Prof. dr. Xiaoling Chen (Assistant promoter)

Dedicated to my parents, husband and son

Acknowledgements

When I started to write this acknowledgement part, many people and things came to my mind. They helped me out of all kinds of troubles, supported me through the hard times, and most importantly shaped my beautiful life. It is not possible to express my endless gratitude to all of them. For the people who are not mentioned here, please also accept my sincere appreciations.

The first and most one I would like to thank is my promoter and supervisor Prof. Alfred Stein. You found the doctoral scholarship for my research, flew many times to Wuhan to discuss my progress, and wrote hundreds of emails to exchange ideas during my study period in China. I always enjoyed your guidance in the field of science, and your respected working style positively and unequivocally influenced my growth and world-view. Thanks for introducing me to the field of statistics, inspiring me to challenge the new research topic, always quickly responding to my requests and contributing constructive comments to every publication. I also remembered your encouragements after my first paper was rejected, your warm and comfort words when I came across hard times in my personal life, and your cheerful congratulations when I got married and had a new baby. It was indeed a great pleasure to meet you and good fortune to work with you for four years, during which time I have learnt a lot.

I would also like to thank my Chinese supervisor Prof. Xiaoling Chen. Thanks for your concern and understanding during my study in The Netherlands. Thanks for providing research materials, arranging the Poyang Lake field work and funding my participation in many conferences, as well as your wonderful research team that gave me help. I would not have had the chance to study abroad and complete this thesis without your support.

I wish to express my gratitude for my co-authors and working part-

ners. To PhD candidate Xiang Zhang, my husband and best working partner, I enjoyed every stimulating discussions with you about the research. You have always been very patient to me in person, but critical to my research and giving me constructive suggestions from the GIS point of view. To Dr. Tiejun Wang, thanks for your unconditional supports all the way back from my MSc study. Your way of being a right-hearted person, organizing team work and doing exact research influenced me deeply. To PhD candidate Lian Feng, thanks for your help in the Poyang Lake field work, and for preparing data that have been used in two chapters in this thesis. I wish you much success in your PhD research and a bright career after graduation.

Many thanks go to the people who provided valuable contributions to my research. To Dr. Michael Weir, thank you for reading and editing my proposal, and for your everlasting concern about my research and life. To Dr. David Rossiter, thanks for your attention and questions during my proposal qualifier and for spending time in editing my first paper. I appreciate critical comments on random sets from the specialist Dr. Marie-Colette van Lieshout, which strengthened my confidence in exploring this topic. I enjoyed the discussions with Dr. Ulanbek Turdukulov on selecting software for the experiments, with Prof. Gerard Heuvelink on geostatistics during a summer school, and with Prof. George Vosselman on journal publication. Thanks also to Dr. Guofeng Wu for his generous offer to supply images for my research.

I would like to thank all the colleagues in the EOS department for their continued supports, raising interesting questions during the EOS meetings and in sharing a wonderful time in department activities. Thanks also to all the ITC staff from the logistics departments and the library for your efforts in releasing my research time, in particular Loes, Teresa, Marga, Roelof, Rebanna, Marion, Bettine, Marieke, Lin Pei and Tina.

Special thanks go to my office-mates: Coco Rulinda, Juan Pablo Ardila Lopez and Yaseen Mustafa. I love the multi-cultural environment and I really like to work with you in the wonderful and quiet office atmosphere. We discussed our research findings, shared literature and software skills, talked about our worries and concerns, encouraged each other, cheered for every progress and made jokes during coffee time. Thank you all for accompanying me and bringing happiness during the long research time. I wish you all a bright and prosperous future.

The Chinese student community at ITC is like a big family. I like to thank all my Chinese friends that I met in the foreign country of the Netherlands and with whom I spent a unforgettable time. It is not possible to

mention you all here, and please accept my sincere appreciations if you are not mentioned: Bian Meng, Changbo, Dongpo, Fangning, Fei Teng, Fengfan, Guangyi, Li Xia, Longhui, Minghui, Ningrui, Pu Shi, Qifei, Qiuju, Siqi, Sudan, Tianxin, Xiao Jing, Xiao Wen, Xiaogang, Xiaojing, Xiaotang, Xinping, Xiong Biao, Xuanmei, Xuelong, Yang Chao, Yang Dong, Yang Ying, Yijian, Yixiang, Zeng Zheng, Zhenshan, Zhong Lei, Zhou Liang, Zijin. I would further like to express special thanks to my dear friends: Si Yali, Fangfang and Hao Pu, as well as to Xue Wen and Wang Li. Many wonderful moments bear in my mind: the happy parties we had, laughs and tears we shared, and warm Chinese festivals we spent together. They have contributed to the most exciting and enjoyable time of my life.

Thanks go to all the colleagues in Prof. Chen's research team for your support in China, in particular Liqiao, Xiaobing, Liqiong, Jianzhong, Shoujing and Zhang Li. Many thanks go to Prof. Xiaoping Pang, Prof. Tinghua Ai and Dr. Feng Gao for their persistent concern and supports. I would also like to express my indebted appreciation to my best friends in China: Lili and Jianwei for their help and everlasting friendship.

Finally, I wish to express my deepest gratitude to my parents and parents-in-law for their physical and mental supports. Thank you for creating such a cozy and harmonious family atmosphere, and taking care of my daily life when I was pregnant and delivered the baby. My heartfelt thanks are to my husband Xiang Zhang for his love and everything he did for me and our son. To my son Zhecheng, you are the other wonderful output beside this thesis at the end of my PhD. Thanks for your cooperation during thesis writing, and I wish you a healthy and happy life into a bright future.

Abstract

Remotely sensed images as a major data source to observe the earth have been extensively used in spatial-temporal analysis in environmental research. Information on the spatial distribution and spatial-temporal dynamics of natural landscapes recorded by series of images, however, usually bears various kinds of uncertainties. This thesis proposes a random set method to deepen our insight into the uncertainties that are inherent in these observations of natural phenomena from images. The general objective of this research is to develop different techniques based on random sets to represent image objects with indeterminate boundaries, quantify their extensional uncertainties, and address uncertainty modeling in a spatial temporal change analysis. The methods are applied to classifying wetland vegetation and monitoring wetland inundation in the Poyang Lake area in China.

Main concepts and statistical characteristics of random sets are introduced in the context of geoinformation science in Chapter 2. They characterize the spatial distribution of a random set, including covering function, p-level sets, support set, core set, median set, mean set and variance.

Chapter 3 represents extensional uncertainties of extracted image objects by random sets and quantifies their degree of extensional uncertainty by derived indices. The number of iterations to achieve a stable covering function and the sum of the variances are good indicators of boundary sharpness. The coefficient of variation has a positive relation with the degree of uncertainty. An asymmetry ratio reflects the uneven gradual changes along different directions where broad boundaries exist. Results show that several characteristics of extensional uncertainty of segmented objects can be quantified numerically and spatially by random sets.

Chapter 4 addresses the accuracy assessment of segmented uncertain objects modeled by random sets. Results show that significant correlations exist between a covering function of random set model and actual vegetation coverage, and that their probability distributions

are similar. The accuracy of random set varies from good to moderate according to kappa coefficients derived from the confusion matrixes of support, mean, median, and core sets.

Chapter 5 develops a mixed Gaussian model with three components to serve as a general method to parameterize random sets in a multi-temporal analysis. Transition zones between wetland vegetation and open waters are identified and random set related indices are adopted to show spatial uncertainties of inundated areas. Results reveal that random sets provide detailed spatial configurations of the wetland water regimes as presented by water covering days (WCD). Random sets contribute to WCD mapping by means of smoothing the WCD variation captured by limited snapshots and by providing more details than when using a crisp method.

In Chapter 6, a random spread process model (RSP) is constructed for monitoring the spatial-temporal pattern of wetland inundation, in which random sets are used to model the spatial extent of each stage. Periodicity, trend and random components are captured by monthly and yearly random sets. The Cov-Dist matrix and related operators are used to summarize the spatial pattern and quantify the similarity of different stages in the process. Results improve our understandings of the substantial seasonal dynamic pattern of the lake and reveal a subtle interannual change trend in the monitoring period. Therefore, the random spread process serves as a valuable addition to include uncertainty in wetland inundation modeling.

In conclusion, this research shows that random sets provide a general framework for describing uncertainties of natural landscape extracted from remote sensing images. It enriches spatial and spatio-temporal modeling of phenomena which are uncertain in space and dynamic in time.

Samenvatting

Satellietbeelden die een belangrijke bron van gegevens vormen om de aarde waar te nemen zijn regelmatig geïntegreerd in milieuonderzoek. Informatie met betrekking tot de ruimtelijke verdeling en de ruimtelijk-temporele dynamiek van natuurlijke landschappen zoals die is geregistreerd op basis van een reeks van beelden bevat verschillende vormen van onzekerheid. Dit proefschrift stelt een methode voor op basis van toevalsverzamelingen om tot een dieper inzicht te komen in deze onzekerheden die inherent zijn in deze waarnemingen. Het algemene doel van dit onderzoek is geweest om verschillende technieken te ontwikkelen op basis van toevalsverzamelingen. Deze zijn ontwikkeld om objecten in beelden te representeren met brede grenzen, om existentiële onzekerheden van objecten te kwantificeren en om onzekerheden te modelleren in een ruimtelijk-temporele analyse van veranderingen. De methoden worden toegepast om vegetatie in ecosystemen van waterrijke gebieden (wetlands) te classificeren en de overstromingen daarin, gericht op het Poyang merengebied in China.

De belangrijkste concepten en statistische karakteristieken van toevalsverzamelingen in de context van geinformatiekunde worden uiteen gezet in Hoofdstuk 2. Deze karakteriseren de ruimtelijke verdeling van toevalsverzamelingen, waaronder de bedekkingsfunctie, p-verzamelingen, de basisverzameling, de kernverzameling, de mediaan verzameling, de gemiddelde verzameling en de variantie verzameling.

Hoofdstuk 3 introduceert de onzekerheid in de omvang van toevalsverzamelingen en kwantificeert deze via afgeleide indices. Goede indicatoren voor de scherpte van de grens zijn het aantal iteraties om een stabiele bedekkingsfunctie te krijgen en de som van de varianties. De coëfficiënt van variatie heeft een positieve relatie met de mate van onzekerheid. Een asymmetrie fractie weerspiegelt de ongelijke geleidelijke veranderingen in de verschillende richtingen waar brede grenzen bestaan. De resultaten laten zien dat toevalsverzamelingen in staat zijn om verschillende eigenschappen van existentiële onzekerheid in gesegmenteerde objecten ruimtelijk en numeriek te kwantificeren.

Hoofdstuk 4 richt zich op een schatting van de nauwkeurigheid van objecten die via een segmentatie zijn verkregen. Resultaten laten zien dat significante correlaties bestaan tussen een bedekkingsfunctie van een toevalsverzameling en de waargenomen vegetatie bedekking en dat deze een vergelijkbare kansverdeling laten zien. De nauwkeurigheid van een toevalsverzameling varieert van goed tot matig voor de verwarringsmatrices van de basis, gemiddelde, mediaan en kernverzamelingen op basis van kappa statistiek.

Hoofdstuk 5 ontwikkelt een gemengd Gaussisch model met drie componenten als een algemene methode om toevalsverzamelingen te karakteriseren in een multi-temporele analyse. Overgangszones tussen vegetatie van natte ecosystemen en open water worden geïdentificeerd en indices die horen bij toevalsverzamelingen geven ruimtelijke onzekerheden van ondergelopen gebieden weer. De resultaten tonen aan dat toevalsverzamelingen gedetailleerde configuraties van de ecosystemen laten zien, in het bijzonder in de vorm van het aantal overstromingsdagen (WCD). Zij leveren tevens een bijdrage aan het karteren van WCD via het temperen van de WCD variatie in de objecten die worden waargenomen via een beperkt aantal opnames en ze laten meer details zien dan wanneer een deterministische methode wordt gehanteerd.

In Hoofdstuk 6 wordt een model voor een ruimtelijk verspreidingsproces opgesteld om het ruimtelijk-temporele patroon in overstromingen van natte ecosystemen te monitoren. Toevalsverzamelingen modelleren hierbij de ruimtelijke uitgebreidheid tijdens de verschillende fases. Periodiciteit, trend en toevalscomponenten worden gerepresenteerd via toevalsverzamelingen met een maandelijkse en een jaarlijkse basis. De Cov-Dist matrix en daaraan gerelateerde operatoren worden gebruikt om het ruimtelijke patroon te beschrijven en om overeenkomsten gedurende de verschillende fasen in dit proces samen te vatten. De resultaten helpen om ons begrip van een sterk seizoensgebonden dynamisch patroon van het Poyang meer te begrijpen en laten een subtiele verandering zien in de trend gedurende verschillende jaren gedurende de periode van monitoren. Om die reden levert het toevalsproces nuttige informatie op. Het is daarmee een welkome aanvulling om onzekerheid in het modelleren van overstromingen in natte ecosystemen op te nemen.

Het proefschrift sluit af met de conclusie dat toevalsverzamelingen een algemeen kader vormen voor het beschrijven van onzekerheden in natuurlijke landschappen die worden waargenomen met behulp van satellietbeelden. Ze verrijken het ruimtelijk en ruimtelijk-temporeel modelleren van fenomenen die onzeker zijn in de ruimte en dynamisch in de tijd.

Contents

List of Figures

List of Tables

Introduction

1

1.1 Background

The topics of spatial data quality and uncertainty have received increasing attention in geographical information science and remote sensing for more than a decade [16, 44, 54, 60, 66, 155, 158, 164, 165, 195]. Uncertainty refers to what is 'known imperfectly' [5]. It arises from ignorance in understanding, through measurement towards prediction, and finally impacts on the quality of produced spatial data [5]. Since uncertainty is an issue along the complete image processing chain and propagates from one step to the next, uncertainty analysis and modeling also refers to error propagation modeling [27, 45, 80].

In this section, I overview the framework of modeling uncertainty in image mining chain, review various techniques for handling particular types of uncertainty, and put emphasis on introducing random set theory at the end.

1.1.1 Uncertainty in image mining

Recent research has adopted the image chain as a means to describe the typical remote sensing process and to highlight the sources of uncertainty in each steps [45]. Dungan [35] outlined the key sources of uncertainty in remote sensing, including uncertainty about input variables, their locations and spatial supports, the structure and parameters of the model. Karssenberg and de Jong [80] broke down the complex image analysis into a sequence of applications of simple uncertainty models. They focused on uncertainty in model input variables, model equations and parameters, but neglected spatial location and supports. Following Sinton's [161] categorization, Gahegan and Ehlers [49] restricted the uncertainty in remote sensing to the following properties: value (including measurement and label errors), spatial uncertainty, temporal uncertainty, consistency and completeness. Four stages in image processing that represent the four models of geographic space were considered: the field model, the image model, the thematic model and the object model. Uncertainty characteristics in each data transformation step from field model to object model were formulated.

Image mining is defined as 'the analysis of (often large sets of) observational images to find (un)suspected relationships and to summarize the data in novel ways that are both understandable and useful to stakeholders' [163]. Stein et al. [167] distinguished five steps in image mining: identification, modelling, tracking, prediction and communication with stakeholders. Aspects of spatial data quality in each of these steps were discussed. Compared with the four models in [49], the image mining process jumped over the field model and started from the image model. Moreover, one general step, i.e. identification, replaced two data

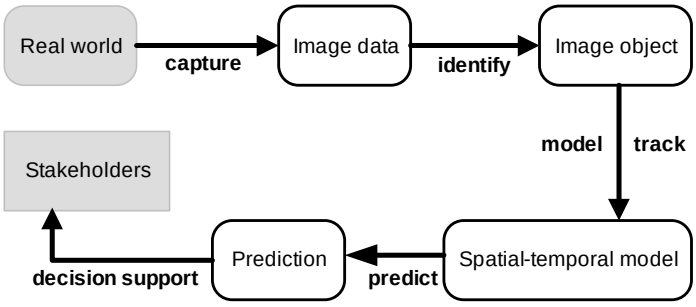

Figure 1.1 Image mining for uncertain objects.

transformations from the image model to the thematic model and to the object model in [49]. We considered the five steps in [167] that linked the original data sources through the chain to end-users as a more comprehensive framework. It was adopted and slightly modified (Figure 1.1) to organize the uncertainty modeling studies in this thesis.

1.1.2 Techniques for handling uncertain objects

Previous research [22, 125, 136, 143] partitioned objects with broad boundaries into two categories based on different causes of the boundary uncertainties: indeterminate/uncertain objects and vague objects. Indeterminate/uncertain objects have boundaries whose position and shape are unknown or cannot be measured exactly, even though the boundaries are sharp. This kind of indeterminate boundaries is due to errors in measurements and the finite representation of numbers in the computer. Vague objects have poorly defined boundaries, which are not due to the limitations of current technology, but intrinsically belong to the nature of objects. The terms vague, indeterminate and uncertain are often used interchangeably, but in this thesis we reserve the term vague for poorly defined objects, indeterminate for our ignorance, and uncertain for the general name of both. If admitting that no measurement is error-free [60], that our knowledge is always limited [84] and that any geographic phenomenon is intrinsically vague [40], then all objects with broad boundaries have both indeterminate and vague characteristics. In addition, many problems are the same when representing and reasoning for both ignorance-based uncertainty and intrinsic vagueness [53]. Therefore, different kinds of uncertainties are indistinguishable for modeling objects with broad boundaries, and all objects with either indeterminate or vague boundaries are called uncertain objects in this thesis.

Various approaches based on probabilistic theory, fuzzy set theory and rough sets have been developed as a means to accommodate the modeling of uncertain objects.

Probability

Probability theory is a major theory for handling uncertainty of geographical entities [195]. Conventionally, it has been applied to modeling positional uncertainties of geographic objects in GIS. Models for depicting the positional uncertainties of these objects have been developed, distinguishing between point features, line features [e.g. 30, 55, 154, 156] and area features [e.g. 68, 157]. Error models for points have been studied for a long time in the fields of geodesy, surveying and mapping. The normal distribution is commonly assumed as the probability distribution for modeling errors in a point location.

Error models for line and area features are based on uncertainty in the coordinates of the constituting elementary vertices. Those methods include the epsilon-band or buffer models [55], the distortion model [82], the confidence region error models [154] and the error-band models [156]. In the epsilon-band model, a buffer of radius ϵ is imposed around each line segment to indicate the error region. A confidence region error model delineates a confidence region within which the true location of the line lies. Shi [154] developed it for lines in N-dimensions by using statistical derivations. An error-band model is based on the assumption that the errors of the two endpoints are independent of each other. Shi and Liu [156] introduced a more generic error model based on a stochastic process to handle the cases where the endpoints of a line segment are dependent. In [30], de Bruin considered uncertainty caused by sampling and approximating a curvilinear feature by a sequence of straight lines, when modeling boundaries in natural landscapes.

Methods based on probability theory model uncertainty with probability distribution functions (PDFs). They have been widely used for studying error propagation. Since analytical solutions for error propagation are only available for a limited number of relatively simple PDFs, Monte Carlo simulation is mostly implemented for uncertainty analysis and sensitivity analysis [26]. Monte Carlo simulation generates multiple realizations of the underlying spatial phenomenon, representing the uncertainty of the simulated variables [67]. For example, the uncertain extent of a lake is simulated by connecting multiple points on the boundary which vary under uncertainty [68]. As such, modeling the positional uncertainty of the lake object requires a separate PDF for each primitive point, together with the statistical dependencies between these points. The large number of primitive points, complexity of the PDF

and the assumptions on statistical dependence, however, complicate the practical implementation.

Fuzzy sets and rough sets

An alternative for handling the uncertain spatial objects is by means of fuzzy set theory [e.g. 20, 32, 40, 41, 134]. Fuzzy set theory treats imprecise definitions of a concept by means of membership functions. Robinson [134] reviewed various kinds of membership functions and commonly used methods for specifying them, including the fuzzy c-means classifier and the semantic import model derived from expert knowledge. Since the assignment of the membership function is subjective in nature, a major obstacle of fuzzy set-based approach is determination of the membership function for a given concept or a spatial object [134].

Schiewe and Ehlers [141] determined the fuzzy widths of interested vague objects based on the mean deviations derived from repeatedly digitizing boundaries from aerial photos. Wang and Hall [180] proposed that the membership function should correspond with sharpness in changes and calculate it by the first order derivative of the property of interest. Van de Vlag and Stein [178] modeled beach area as fuzzy objects where membership degrees are in terms of the thematic definition of class 'beach'. They used a Bayesian procedure to combine prior knowledge with collected data into posterior distributions and used them to improve the membership functions in classification. Fisher [41] addressed the issue of higher order vagueness, and introduced the type 2 fuzzy sets to reveal the depth of the uncertainty. Type 2 fuzzy sets allows more descriptive qualifications of the modeling results, such as reporting fuzzy area as a fuzzy number, and specifying minimum, maximum and average extents of objects.

Several spatial data models have been proposed for vague objects and their topological relations based on fuzzy set theory [32, 89, 144, 147, 193]. Definitions of vague spatial data types and operators were derived from concepts in general topologies [181] and fuzzy topologies [85]. Schneider provided fundamental work for formal definitions of fuzzy types [144, 147], and definitions of topological and metric operators on fuzzy objects [145, 146]. More recently, Dilo et al. [32] gave an overview on existing models proposed to handle spatial vagueness and provided mathematical definitions for vague object types and operators.

Rough sets, an extension of standard mathematical sets like fuzzy sets, are represented by the lower and upper approximation. The lower approximation contains all the elements that certainly belong to the set, whereas the upper approximation contain all elements that possibly belong to the set [124]. Rough sets were used to model vague regions in

[11, 136]. Ahlqvist et al. [1] introduced rough classifications and have shown the possibility of reasoning about areas that have been classified using rough sets. Further, Ahlqvist et al. [2] introduced the idea of rough fuzzy classification and defined various accuracy measures on it to allow the integration of rough and fuzzy geographic data.

1.1.3 Random sets

A random set is a generalization of a random variable, taking sets as its elements. Random set theory has become an inherent part of probability theory since Matheron [104]. This theory studies probability measures on the space of closed subsets with the hit-or-miss topology. The mathematical foundation of random sets is based on Choquet's capacity theorem, which characterize distribution of the set-valued random elements as nonadditive measures. More recent mathematical definitions of concepts in random set theory refer to Molchanov [109, 110, 111], Nguyen [113].

Random set theory has been viewed as a generalization of other uncertainty theories [58, 101]. Close connections between random sets and other theories have been explored and uncovered in several publications:

- Random sets provide essentially mathematical foundation for point process [8, 133, 170].
- Fuzzy sets are equivalent to a weak specification of random sets. The membership function of a fuzzy set analogous to the probability of a random set covering a generic point [56, 100, 114, 129, 182, 192].
- Dempster-Shafer theory of evidence can be expressed formally within the context of random set theory [98, 99, 115, 129].
- Random set theory serves as a unifying foundation for expert system theory [34, 57, 59, 65, 87].

Since traditional probability theory can not capture the full scope of uncertainty, random set theory usually is to be preferred as it is able to deal with interval bounds, fuzzy sets and discrete probability distributions [3, 62, 118, 148]. It allows to model the uncertain variables using probability boxes, possibility and probability distribution functions and families of intervals provided by experts. In engineering models, random sets were typically employed to obtain upper and lower bounds for the reliability of structures and to describe the variability of material parameters and geometrical data [173, 174].

Random set theory provides a means of extending classical single-sensor, single-target point-value statistics to a multi-sensor, multitarget

statistics [58]. It is applicable and useful for general data fusion, to address problems of determining the locations and identities of targets from sensor data with clutter, jamming, and measurement noise effects [100].

Images and noise process can be modeled as random sets in image processing and analysis to solve various problems. Random set theory, in conjunction with mathematical morphology, provides a rigorous statistical foundation for nonlinear image processing to recover images from certain types of noise processes [150]. Molchanov [108] gave an overview of existing definitions of expectations for random sets with emphasis on their use in image analysis. Baddeley and Molchanov [6] introduced an new notion of expectation for random sets (or average of binary image) based on distance function. Friel and Molchanov [48] considered expectations of random sets generated from images as the threshold sets and presented a new approach towards grey-scale image thresholding. Further, Jankowski and Stanberry [76] defined the expected set and the expected boundary based on oriented distance function and applied the method to constructing the boundary of an object in a noisy image. Besides, frameworks for texture classification [36] and context-based classification [12] have been proposed to improve classification accuracy.

Random set theory was developed as part of stochastic geometry and was used to simulate and model patterns in natural or technical process [7, 81, 104]. For example, it was used to model ore-sintering structure by Serra [149] and incidence of heather by Diggle [31] and simulate natural textures in binary images [50]. Due to its capability of studying randomly varying geometrical shapes, Cressie [24] recommended it as a sound theory for modeling uncertain objects. Stoyan [168] used random sets in particle statistics to obtain the mean extent as well as shape fluctuations of sand grains. Stoyan and Stoyan [169] showed that random set as a set-theoretic method can act as a supplement of other powerful means for particle classification. Random set theory has also been used to study the spatial component of object growth process, such as tumor growth [25] and fire spread [151, 179].

1.2 Research scope

This research focuses on modeling uncertain objects in the identification step and the modeling and tracking step of image mining.

The delineation and representation of uncertain objects with gradual or inherently vague extensions such as forest stands, soil types and wetlands, is a recurrent topic in ecological [74, 92] and geo-science studies [16, 46]. The ignorance in understanding and use of the crisp data

model will simplify and cause uncertainties in the identification stage. To avoid this simplification, several conceptual models and spatial data models have been proposed to represent geographical phenomena with uncertain boundaries [e.g. 16, 32, 68]. In addition, assessing the accuracy of modeled uncertain objects is challenging [167]. The information about the uncertainty represented in the modeling results does not necessarily correspond with objects in the field. This is because even on the ground, the delineation of uncertain objects may be impossible, due to vague and ambiguous boundaries. Moreover, detailed reference data that are critical for validating uncertainty modeling results, are often unavailable, especially when field data have mainly been collected for validating hard classification. Two chapters in this thesis address uncertainty issues in the identification step of image mining.

A series of images has been collected and integrated together for monitoring dynamic geographical phenomena. These images are only snapshots of processes, and thus observations from them are limited by spatial resolutions and revisiting intervals of sensors. Analyzing uncertainties in the modeling and tracking stage for spatially continuous and temporally dynamic phenomena may enrich our understanding in spatial temporal modeling and change analysis [167]. Examples are found in modeling saline wetland [142], savanna-forest intergrade [41], coastal landscape [177], glaciers [190] and tundra-taiga transition areas [130]. Two chapters in this thesis focus on uncertainty issues in the modeling and tracking step of image mining.

1.3 Research objectives

The general objective of this thesis is to develop different techniques based on random sets to represent image objects with indeterminate boundaries, quantify their extensional uncertainties, and address uncertainty modeling in a spatial temporal change analysis. The methods have been applied to dynamic wetland monitoring in the Poyang Lake area in China.

The specific research questions are:

1. How should random sets be defined and generated to model uncertain objects? (chapter 2-6)
2. How do random sets quantify the extensional uncertainty of image objects? (chapter 3)
3. How can the accuracy of classified uncertain objects that are represented by random sets be assessed? (chapter 4)

4. Does the random set model give more information than a crisp data model on the spatial dynamics of uncertain phenomena? (chapter 5)
5. How should a random spread process be built based on random sets, that allows one to include uncertainty in spatial temporal modeling? (chapter 6)

The specific objectives concerning the wetland application are:

1. To represent vegetation patches with boundaries of a varying degree of vagueness, in order to quantify their extensional uncertainties. (chapter 2, 3)
2. To assess the accuracy of a classified vegetation patch. (chapter 4)
3. To delineate the transition zone between wetland vegetation and open water, and describe its spatial changes. (chapter 5)
4. To compare the spatial configuration of inundation days by either considering the transition zone or ignoring it. (chapter 5)
5. To model wetland inundation as a random spread process, and to investigate its seasonal pattern and interannual trend in changes. (chapter 6)

1.4 Study area

According to the different methods developed in each chapter and the availability of experimental datasets, the wetland application were conducted at various spatial scales, from the whole Poyang Lake, to the Poyang Lake national nature reserve (PLNNR) and vegetaion patches around Banghu lake within PLNNR (Figure 1.2).

1.4.1 Wetland in Poyang Lake

Poyang Lake is the largest freshwater lake in China, located in the Northern part of Jiangxi Province, at the South bank of the middle and lower reaches of the Yangtze River (115° 47′-116° 45′E, 28° 22′-29° 45′N). Main water supplies to the lake are from local precipitation, five source tributaries (Ganjiang, Fuhe, Xiushui, Xinjiang, and Raohe), and from the Yangtze River during summer months. The seasonality of precipitation and water inflow have led to significant variation in the lake's inundation area.

During late spring and early summer (May and June), the five tributaries are flooded due to concentrated rainfall. As a consequence, the water

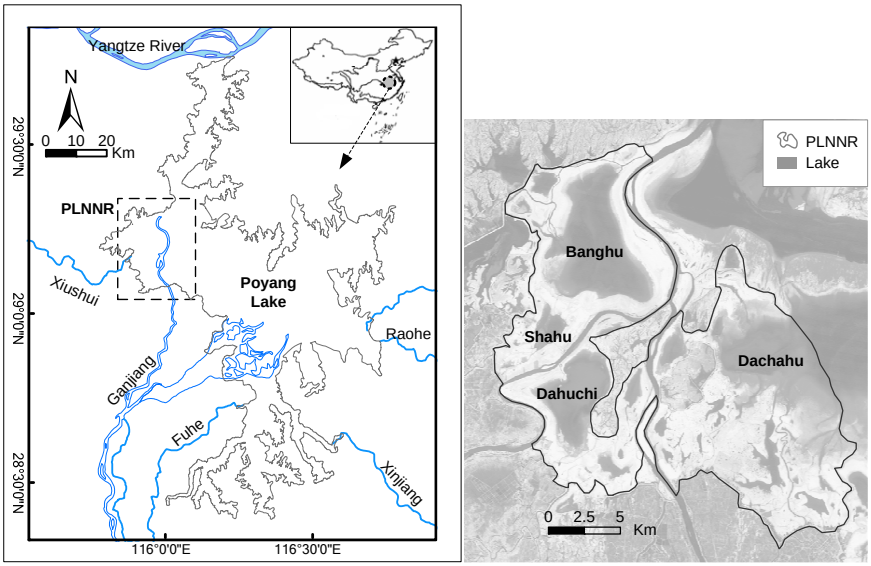

Figure 1.2 Locations and boundaries of Poyang Lake and PLNNR.

level of Poyang Lake swells up without reaching the peak yet. From July until early September, the water supply from the five tributaries decreases, whereas the Yangtze River has its highest level, which causes the water level to rise to its peak. From late September until November, the water level recedes towards its lowest level after the recession of the Yangtze River [72]. Interactions between the lake, rivers and surrounding land constitute an aquatic-terrestrial ecosystem within the Poyang Lake area. Seasonal dynamic pattern of the lake and interannual change trend from 2000 to 2010 were investigated in Chapter 6.

1.4.2 Poyang Lake national nature reserve (PLNNR)

Poyang Lake national nature reserve (PLNNR) is located in the northwest of the Poyang Lake area (115° 55′-116° 03′E, 29° 05′-29° 15′N). Nine lakes in PLNNR are connected to Poyang Lake during high water levels in summer and disconnected when water levels are low in other seasons. Three land cover types dominate the PLNNR: water body, water-vegetation transition zone and wetland vegetation. Water-vegetation transition zones are wide on lake banks due to gradually changing elevations, and they usually consist of shallow standing water, saturated soil, dead algae and sparsely distributed young flush sedge. Such transition zones are suited for different waterfowl species and serve as important habit-

ats of migratory birds. Dynamics of the transition zones and wetland inundation in PLNNR were monitored in Chapter 5.

Two types of emerged vegetation dominate the wetlands grassland: *Miscanthus* dominant vegetation communities occur in high elevation areas and *Carex* dominant communities in low elevation areas. These vegetation types are all blooming in spring and serve as important habitats and forages for spring migration birds. When winter migration birds arrive in autumn, *Miscanthus* dominant community in high elevation areas become senescent. *Carex* dominant communities in low elevation areas are turning green and are growing up gradually. They allow different kinds of winter migration birds forage on leaves and rhizomes in different elevation zones [186]. The experiments in Chapter 2-4 were carried out in the area surrounding Banghu lake within PLNNR where the wetlands grassland has the largest area and shows obvious boundary gradients.

1.5 Thesis outline

This thesis consists of seven chapters, including the introduction, the synthesis and five main chapters which have been submitted or published as peer-reviewed journal paper or book chapter. The contents of the papers have largely been retained, therefore overlaps and repetitions may occur among the main chapters. I unified the formatting styles of figures, tables, citations and units throughout the thesis, and thus they may different from the original papers. The references in all chapters were combined in a common list to avoid redundancy. The structure of this thesis is shown in Fig. 1.3 and the contents of chapters are as follows:

Chapter 1 introduces the research background, defines the research objectives and questions, presents the study area and outlines the structure of the thesis.

Chapter 2 proposes a random set method for uncertainty modeling of spatial objects extracted from images in environmental study.

Chapter 3 models a collection of outcomes from a standard segmentation algorithm as a random set, and quantifies extensional uncertainties of extracted objects using statistical characteristics of random sets.

Chapter 4 addresses the accuracy assessment of segmented uncertain objects and assesses the quality of the random set modeling

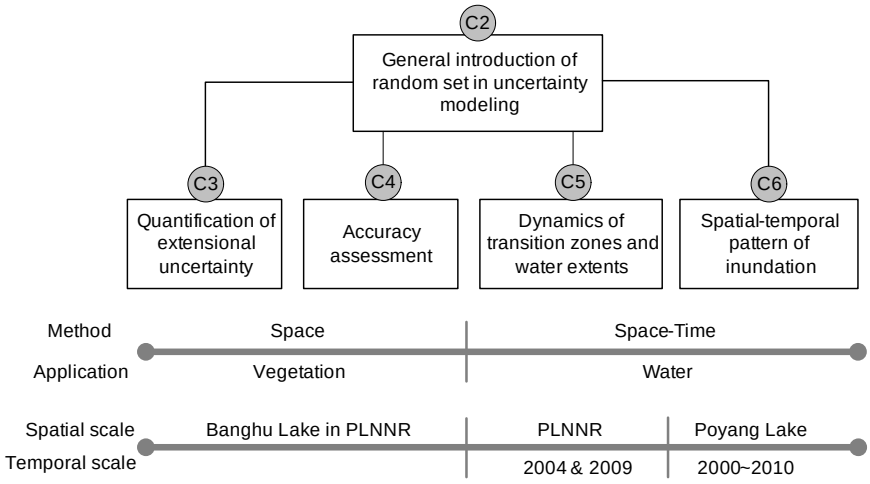

Figure 1.3 Diagram showing the structure of the thesis. C2-C6 refer to Chapter 2 - Chapter 6. The horizontal axis illustrates aspects of method and application domains, spatial and temporal scales.

results, in particular for modeling the extensional uncertainty of a vegetation patch.

Chapter 5 examines the application of random sets to monitor changes of wetland inundation with uncertain spatial extents. It addresses parameterization of a mixed Gaussian random set model for determining transition zones between wetland vegetation and open water from series of images.

Chapter 6 models wetland inundation as a random spread process and uses random sets to represent stochastic inundation areas extracted from multitemporal images, in order to explore the spatial-temporal pattern of the wetland inundation.

Chapter 7 summarizes the main findings and conclusions of Chapter 2 to 6, gives general conclusions and suggests further works.

2 Modeling uncertainties of natural entities extracted from remote sensing images

Published as: Xi Zhao, Alfred Stein and Xiaoling Chen (2010) Application of random sets to model uncertainties of natural entities extracted from remote sensing images. In: Stochastic Environmental Research and Risk Assessment, 24, pp. 713-723.

Abstract

Remotely sensed images as a major data source to observe the earth, have been extensively integrated into spatial-temporal analysis in environmental research. Information on spatial distribution and spatial-temporal dynamic of natural entities recorded by series of images, however, usually bears various kinds of uncertainties. To deepen our insight into the uncertainties that are inherent in these observations of natural phenomena from images, a general data modeling methodology is developed to embrace different kinds of uncertainties. The aim of this paper is to propose a random set method for uncertainty modeling of spatial objects extracted from images in environmental study. Basic concepts of random set theory are introduced and primary random spatial data types are defined based on them. The method has been applied to dynamic wetland monitoring in the Poyang Lake national nature reserve in China. Four Landsat images have been used to monitor grassland and vegetation patches. Their broad gradual boundaries are represented by random sets, and their statistical mean and median are estimated. Random sets are well suited to estimate these boundaries. We conclude that our method based on random set theory has a potential to serve as a general framework in uncertainty modeling and is applicable in a spatial environmental analysis.

2.1 Introduction

Remote sensing plays an important role in data collection for monitoring dynamic geographical phenomena. Series of remotely sensed images as a major data source to observe the earth have been extensively integrated into spatial-temporal analysis in environmental research, for example for monitoring changes in lake sediment [184], forest fire [176], glacier coverage [189] and urban heat island [28]. These images are only snapshots of processes, however, that neither directly represent geographical phenomena as entities, nor the changes in and interactions between the entities [116]. For various reasons, e.g. performing spatial analysis and reasoning on their relationship, geographic entities need to be extracted from images and represented as crisp or uncertain objects [167]. Therefore, spatial data modeling methodology should be general enough to embrace uncertainties and let spatial data models serve as a representation of real-world entities [101].

Traditionally, probability theory has been used for modeling error propagation in spatial objects [195]. It mainly deals with vector data, e.g. point, line and polygon, which have positional uncertainties caused by measurement errors. Models have been well developed for uncertain points [95, 171] and lines [156]. Models for uncertain polygons, however, have been based on models for points and lines, as the uncertain location of the polygon outline is specified by the joint probability distribution function (pdf) of the primitive points of the outline. It is rarely realistic to derive the joint pdf without reducing the complexity of the pdf of each point and relaxing the assumptions on the statistical dependence between the points [68]. Spatial objects on remote sensing images usually have gradual transition boundary due to either the scale of observation or thematic poor definition [167]. In practice, the pdf of each boundary points is difficult to estimate, because the location of points may have different uncertainty due to different slopes of transition zones. In addition, georeferencing of remote sensing data and manual digitization may introduce correlations between boundary points [68].

The alternative for handling the uncertain spatial objects is by means of fuzzy set theory [e.g. 20, 32, 134]. Fuzzy set theory treats imprecise definitions of a concept by means of membership functions. Examples include mountain peaks [42], beach boundaries [20] and the position of the front of a contaminated air detector tube [86]. Robinson [134] reviewed various kinds of membership functions and commonly used methods for specifying them, including the fuzzy c-means classifier and the semantic import model derived from expert knowledge. Since the assignment of the membership function is subjective in nature, a major obstacle of fuzzy set-based approach is determination of the mem-

bership function for a given concept or a spatial object [134]. van de Vlag and Stein [178] used Bayesian procedure to combine prior knowledge with collected data into posterior distributions and used them to improve the membership functions in classification.

Other uncertainty theories such as Dempster-Shafer's theory of evidence and rough set also allow to model uncertain spatial data [e.g. 88, 175]. Each of these theories has its own properties to model particular types of uncertainty and thus does not suffice to construct data models for diverse kinds of uncertain data or information in environmental research.

Random set theory has been viewed as a generalization of other uncertainty theories [58, 101**?**] and the relationships between those theories have been discussed in the past [3, 114, 118, 182]. To the best of our knowledge, however, no research has been done to apply random set theory to uncertainty modeling in geoscience and environmental research yet, though Cressie [24] recommended it is a sound theory for modeling uncertain objects. In this research, therefore, we propose a general spatial data model based on random set theory. In population space, a natural entity with uncertainties can be treated as a randomly varying set, i.e. a random set. One remotely sensed image is then a realization in the sample space of the underlying phenomena. Other samples exist theoretically, and can practically be made by applying other sensors, different image interpreters or observations at different times. The idea is that objects extracted from images should rely on a set of several independent observations rather than on a single realization generated by arbitrary means. A statistical analysis can then be done to better understand the characteristics of the samples and to estimate the parameters of the population. The purposes of such a statistical analysis could be manifold, for example, to average the samples, to explore their variability, or to fit a probability model.

The aim of this paper is to propose a random set method for uncertainty modeling of spatial data collected in environmental study. Random spatial data types are defined based on random set theory and we hypothesize that it can flexibly serves as a general framework to model spatial data with uncertainties. The proposed method is illustrated with two examples on wetland monitoring in the Poyang Lake Natural Reserve in China.

2.2 Concepts and Methods

2.2.1 Random set theory

Random set theory has become an inherent part of probability theory since the publication of the fundamental book by Matheron [104]. Random sets are generalizations of random variables in conventional probability theory. Unlike random variables, though,the probability laws of random sets are defined on elements which are sets rather than points. More recent mathematical definitions refer to Molchanov [110] and Nguyen [113].

Random set theory acts as the core of stochastic geometry, in which they are stochastic models of irregular or random geometrical structures [7] and relate particular closely with point process [119, 120, 131]. As summarized by Mahler [101], random set theory is employed as a foundation for the study of randomly varying population and randomly varying geometrical shapes. Serra [149] used random set to model oresintering structure and Diggle [31] used it to model the incidence of heather, whereas Stoyan and Stoyan [169] applied it 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. Random set theory has also been used to model random spread process [151], such as fire spread [179] and tumor growth [24]. It has further been applied in image analysis [50], time series analysis [117] and geotechnical engineering [118].

2.2.2 Definition of random set and covering function

Random set

Let $(\Omega, \sigma_\Omega, Pr_\Omega)$ be a probability space and (Ξ, σ_Ξ) be a measure space. A random variable $X(\omega)$ is a measurable function from the probability space Ω to the measurable space Ξ. A random set can be seen as a random variable from the sample space Ω to U where U is a set of subsets of Ξ, i.e. $U \subseteq \mathcal{P}(\Xi)$. In other words, the element u of the space U is a set instead of a point in the space Ξ. Let (U, σ_U) be a measurable space and σ_U a σ-algebra defined on U. Then a $(\sigma_\Omega - \sigma_U)$-measurable mapping Γ: $\Omega \to U$ is a random set. Since we match an element $\omega \in \Omega$ with a $u \in U$,i.e. a subset of Ξ, the random set is also a multi-valued mapping from the space Ω to the space Ξ. The multi-valued mapping indicates that random sets can be used to express the imprecision of the measurement due to the inability of making a precise observation.

The distribution of random set Γ is defined as

$$Pr_\Gamma(\mathcal{B}) = Pr_\Omega(\Gamma^{-1}(\mathcal{B})) = Pr_\Omega\{\omega \mid \Gamma(\omega) \in \mathcal{B}\} \quad \forall \mathcal{B} \in \sigma_U \tag{2.1}$$

Matheron [104] defined the probability measure on U by the capability functional of Γ, which is not additive unless Γ is a random singleton. As we do not use the concept of capability functional in this study, we refer for further details to Molchanov [110].

Alternatively, a random set can be presented by a collection of pairs $(\mathcal{A}_i, m_i), i \in \{1, \ldots, n\}$, where $\mathcal{A}_i$ is called a focal elements of the random set, $\mathcal{A}_i \in \mathcal{P}(\Xi)$. For example, n imprecise observations can be regarded as realizations of $\mathcal{A}_1, \ldots, \mathcal{A}_n$, thus constructing a random set, where then m_i is the uncertainty assignment attached to each focal element. The uncertainty assignment m is a mapping: $\mathcal{A} \rightarrow [0, 1]$, with $m(\emptyset) = 0$ and $\sum_{i=1}^{n} m_(\mathcal{A}_i) = 1$. Without evidence supporting that the occurrence of one focal element is more probable than another, equal uncertainty assignments can be chosen.

Covering function

In the special case of $\mathcal{B}$ being a singleton set, e.g. $\mathcal{B} = \{\xi\}$ and $\xi \in \Xi$, the random set Γ becomes a random variable. The distribution function of the random set Γ in this special case is called the one point coverage function or covering function:

$$Pr_\Gamma(\{\xi\}) = Pr_\Omega\{\omega \mid \xi \in \Gamma(\omega)\} \quad \forall \xi \in \Xi. \tag{2.2}$$

The covering function can be interpreted as the probability of the element ξ on space Ξ being covered by the random set. The set $\Gamma_p = \{x \in R^2, 0 \leq p \leq 1 : Pr_\Gamma(x) \geq p\}$ is called a p-level set. Special cases are: $\Gamma_s = \{x \in R^2 : Pr_\Gamma(x) > 0\}$, i.e. the support set, describing the possible part of Γ and $\Gamma_c = \{x \in R^2 : Pr_\Gamma(x) = 1\}$, i.e. the core set, describing the sure part of Γ.

The covering function of a random set can be estimated from its focal elements A_i, by:

$$\hat{Pr}_\Gamma(\xi) = \frac{1}{n} \sum_{i=1}^{n} I_{\mathcal{A}_i}(\xi), \quad \forall \xi \in \Xi \tag{2.3}$$

where $I_{\mathcal{A}_i}$ is the indicator function of $\mathcal{A}_i, i \in \{1, \ldots, n\}$:

$$I_{\mathcal{A}_i} = \begin{cases} 1, & \xi \in \mathcal{A}_i \\ 0, & \xi \notin \mathcal{A}_i \end{cases}$$

Figure 2.1 illustrates a simple example for covering function estimation. $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$ are four focal elements of a random set with

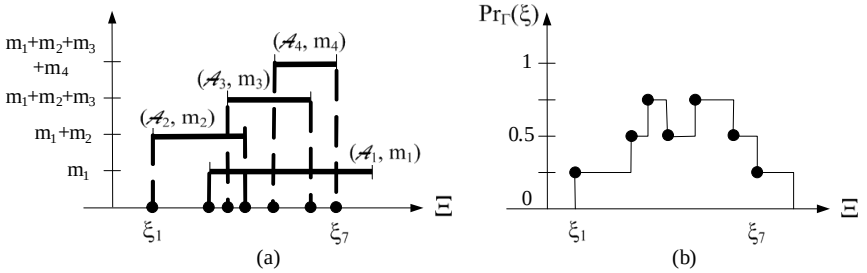

Figure 2.1 Four focal elements with equal uncertainty assignments construct a realization of a random set (left); Covering function of the random set estimated by focal elements (right).

equal uncertainty assignments i.e. $m_1 = m_2 = m_3 = m_4$, reflected by the equal intervals in y-axis of Figure 2.1(a). Also, $\xi_1, \ldots, \xi_7$ are seven elements in the space Ξ, which are covered by some $\mathcal{A}_i$. The values of the covering function at these points derived by Equation 2.3 are shown in the Figure 2.1(b).

Random spatial data model

In environmental research, we may be interested in analyzing phenomena with spatial distribution and temporal dynamics. For example, to monitor a flooded lake, we may need information quantifying how the shape and size of the lake vary as the water level changes within a year. Series of satellite images provide us with spatial and temporal information about the lake and before carrying out further measurements, we need spatial data model to describe, interpret and formalize the spatial information.

Determinate data models, such as points, lines and polygons with crisp boundaries, in current spatial databases, however, do not take uncertainty into account. For example, a vegetation patch in a wetland area extracted from an image can be modeled as a polygon with a crisp boundary, being a subset of R^2. In this determinate model, the measure space equals R^2 and its elements are points in R^2. On the other hand, by using a general data model, taking uncertainty into account, the spatial extent of the patch can be modeled as a random set by a collection of polygons (i.e. the focal set) which is a set of subset of R^2. Therefore, its measure space is a hyperspace whose elements (i.e. polygons or focal elements) are subsets of R^2. N human interpreters or image segmentations with different parameters can be employed to sample the random set and we obtain N focal elements which are polygons in

R^2. According to Equation 2.3, we can estimate the covering function and the probability of points in R^2 belonging to this random set. For example, the point in the center of the vegetation patch is interpreted as belonging to the lake area by N interpreters, then the covering function of this point will be equal to 1. If another point located in the boundary zone and only half of interpreters consider it as part of the patch, then the covering function of this point will be equal to 0.5.

The concepts of random set described in Section 2.2.2 are rather general. We now make more specific definitions for random spatial data types to facilitate the uncertainty modeling of spatial data extracted from remote sensing images. Methods for estimation of the covering function and statistical parameters of some random spatial data types are introduced.

Definitions of random spatial data types

We denote the space of an image as $I \subset R^2$ where the pixel x is the basic element. The measure space that carries random spatial data models is denoted by U, being a set of subsets of R^2, i.e. $U \subseteq \mathcal{P}(R^2)$. Below, two types of random objects are defined: random point and random region. The random regions will also include random lines.

A random point (Rp) is defined as a random set Γ on U that contains a finite collection of points x with a positive covering function, as:

$$Rp_\Gamma = \{x \mid Pr_\Gamma(x) > 0, x \in R^2\} \tag{2.4}$$

where $Pr_\Gamma(x)$ is the covering function of Rp_Γ at x.

A Rp can be presented as a collection of pairs (O_i, m_i), where O_i is the focal element, $O_i \in \mathcal{P}(R^2)$ and m_i is the uncertainty assignment, $m_i \in [0, 1]$. Since focal elements of a random point are points (singletons), i.e. $O_1 = \{x_1\}, \ldots, O_n = \{x_n\}$, the random point becomes a random variable which is a special case of the random set.

A random region (Rr) is defined as a random set Γ on space U that contains a finite collection of random points Rp and the area enclosed by them. Suppose we are interested in modeling the positional uncertainty of a vegetation patch digitized from an image. An example in Figure 2.2 shows an object observed by ten operators. Vertex A has a smaller variance than vertex B due to the smaller uncertainty at vertex A. In this example, each vertex can be modeled by a random point and each random point has ten realizations. By selecting one of the realizations of each random point and connecting them, we can obtain one polygon as a realization of the random region. The conventional probability method focuses on error propagation from the vertices to the polygon. Random sets, in contrast, emphasize the multiple realizations of the polygon as a random set and their variation and mean.

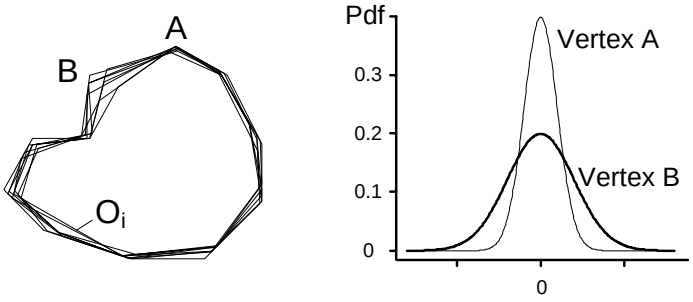

Figure 2.2 Boundary of a vegetation patch represented by a random set (left); different distribution of the coordinates of vertex A and B (right).

A Rr can also be presented by the focal elements O_i with corresponding uncertainty assignments, denoted as a collection of pairs $\{O_i, m_i\}$. The focal elements are regions which are subsets of R^2 (Figure 2.2): $O_i \in \mathcal{P}(R^2)$. If a random region consists of n focal elements with equal probability, then $m_i = \frac{1}{n}$. The probability of point x in R^2 contained by the random region Rr, denoted as $Pr_\Gamma(x) = Pr(x \in Rr_\Gamma)$, can be estimated by the covering function of the random region.

For example, n polygons digitized from an image to model a vegetation patch as a random region, denoted as $O_1, \ldots O_n$, being samples as well as focal elements of the random region. The O_i are assumed to be independent and identically distributed. The probability that pixel $x \in R^2$ is occupied by the random region can be estimated by $Pr_{\cup_{i=1}^{n} O_i}(x)$. An estimator of the covering function of O can be obtained according to Equation 2.3, as:

$$\hat{Pr}_\Gamma(x) = \frac{1}{n} \sum_{i=1}^{n} I_{O_i}(x), \quad x \in R^2 \quad O_i \in \mathcal{P}(R^2). \tag{2.5}$$

A random line (*Rl*) has a similar definition as a random region. The mainly difference is that focal elements of a random line are lines.

Mean and median of random regions

Unlike the case of a random point (random variable), a random region is defined on sets (regions) rather than points of R^2. Dealing with sets is not easy by conventional probability theory. Therefore, we expect that random set models can facilitate calculations on statistical mean and median of sets in R^2.

Since the space U of random regions is non-linear, concepts of mean and median in a linear space cannot be directly applied to random sets. Features such as area of a region, however, can be described in a linear space. Several approaches defining the mean of random compact sets are based on the inverse image of the expectation of these features [108]. Widely accepted definitions of the mean of a random set include the Aumann expectation, the Vorob'ev expectation, the Radius-vector expectation and the Fréchet expectation [169]. The choice of a particular definition depends upon specific application aim. For example, the Aumann expectation is not very attractive for non-convex sets, whereas the radius-vector expectation is a natural choice for star-shaped sets. Here we choose the Vorob'ev expectation, because it considers sets with a finite number of points, such as sets of pixels, which are of interest in image analysis. For the mathematical definitions of the other expectations, we refer to [108, 169].

The mean area EA of the random region Γ is determined by $\mathrm{E}A(\Gamma) = \int_{R^2} Pr_\Gamma(x)\mathrm{d}x$. According to the definition of Vorob'ev expectation, the median set of Γ is the 0.5-level set. The mean set, denoted by Γ_m is defined as:

$$\Gamma_m = \{x \in R^2, 0 \leq p_m \leq 1 : Pr_\Gamma(x) \geq p_m\} \tag{2.6}$$

where p_m is determined for which the set Γ_m has the area $\mathrm{E}A(\Gamma)$. If p_m is not unique then one could take the infimum of all such p_ms. When $p_m = \frac{1}{2}$, the mean and median are identical. In summary, the Vorob'ev mean of random set (Equation 3.2) is estimated by first determining the mean area $\hat{A}(\Gamma) = \int_{R^2} \hat{Pr}_\Gamma(x)\mathrm{d}x$ and then finding a p_m-level set which has area equal to $\hat{A}(\Gamma)$.

2.3 Application

2.3.1 Study area and image pre-processing

The case study considers 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 the Yangtze River, in central China (Figure 2.3). Each year, the water level starts to rise in spring (March to May) due to flooding of its upstream rivers, and keeps its highest level in summer (June to August) as a consequence of flooding of the Yangtze River. The water level falls from September onwards when the flood in the Yangtze River is recessive, and it remains 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 habitat and high quality forage for water fowls [186]. Map-

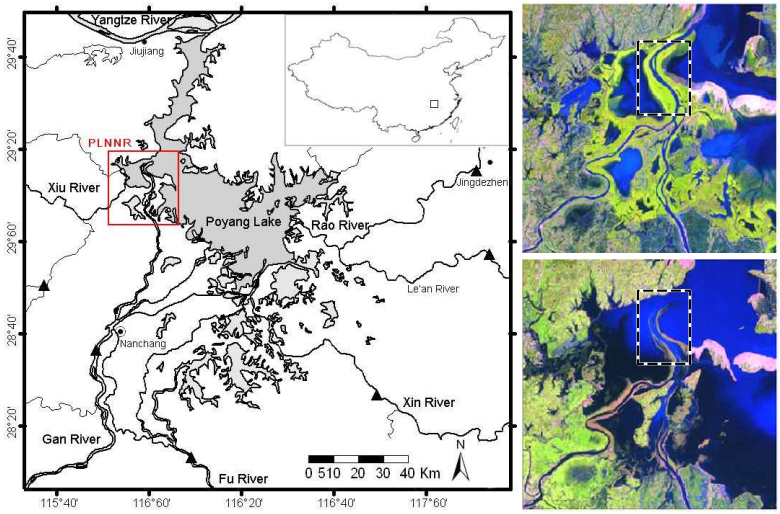

Figure 2.3 PLNNR located in the west of Poyang Lake (left); Landsat TM images covering PLNNR acquired on May 5, 2004 (up right), July 24, 2004 (low right) and the dot-dished boxes indicate the study area for example 1.

ping of wetlands, especially modeling the variance of shape and size of lakes and wet grassland is of great importance for PLNNR managers and decision makers for both ecosystem dynamic monitoring and habitat assessment.

Four Landsat TM images covering PLNNR were acquired at May 5th, July 24th, October 28th and December 15th of 2004 for a wetland monitoring purpose. A topographic map of scale 1:10000 was used as the geographic reference data for image geometric correction. The root mean squared error of all the geometric corrections were less than 10m. Four Normalized Difference Vegetation Index (NDVI) maps were derived by calculating NDVI as:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2.7}$$

where RED and NIR stand for the spectral reflectance measurements acquired in the red and near-infrared band of Landsat TM images, respectively.

Commonly used image segmentation methods to extract spatial objects from images include thresholding, region-based approaches and edge-based approaches. Thresholding only takes the value of image pixels into account, whereas region-based approaches e.g. region grow-

ing approach consider both attribute homogeneity and spatial contiguity of pixels [137]. Since it is arbitrary to choose a single-valued threshold or a specific parameter in segmentation, uncertainties exist in any segmentation results and can have a large effect on the subsequent spatial analysis [94]. Particularly, it is difficult to precisely choose segmentation parameters to identify and extract wetland grasslands with transition zones or vegetation patches with vague boundaries from NDVI images. In the following two examples, we illustrated the application of spatial random data models for modeling uncertainty in the segmentation results of grassland and vegetation patch identification.

2.3.2 Example 1: uncertainty modeling for grassland identification

The thresholding approach was selected to identify grassland (Figure 2.3) from four NDVI images. Suppose that the threshold $NDVI = 0.3$ was adopted for a hard thresholding. To demonstrate how the extent of grassland could vary under different choose of this threshold, we took an NDVI range $[0.1, 0.5]$. This range was divided into 100 equal intervals and 101 thresholds were obtained to slice NDVI images and make samples as binary maps (Figure 2.4). Each sample is a realization of a focal element O_i of a random set. Since we have no emphasis on thresholds, focal elements have equal probability assignments $m_i = \frac{1}{n}$, where $n = 101$. Estimation of the covering function for four Rrs was implemented and has been compared in Figure 2.4. An interpretation of the covering function is that it is the probability of a pixel x covered by the random region O, showing a higher probability of a pixel belonging to grassland where pixel brightness is higher.

We observe that grassland has the largest area in May and most of the vegetation extents are with high NDVI values larger than 0.4. The reason is that all kinds of wetland vegetation species are growing vigorously in spring, including those that grow at high and low elevations. When the wetland is flooded in summer, only vegetation grown on high elevations emerges. During winter, vegetation at high elevation is dying and only vegetation at low elevation (green in Autumn) remains green. Therefore, the grassland shrinks in summer (July) and winter (December). Due to the mixture of fading vegetation species at high elevation and greening vegetation species upon the recession of the Yangtze River at low elevation, the grassland in October has the largest NDVI range. Therefore, variance of vegetation status is expected to be reflected by characteristics of the random set.

We plotted the area of focal elements of the Rrs to explore information on the spread of grassland (Figure 2.5). In addition, the spatial extent of the mean and median of the Rr for October 28 (Figure 2.6) were obtained according to Equation 3.2. We notice that the plot of

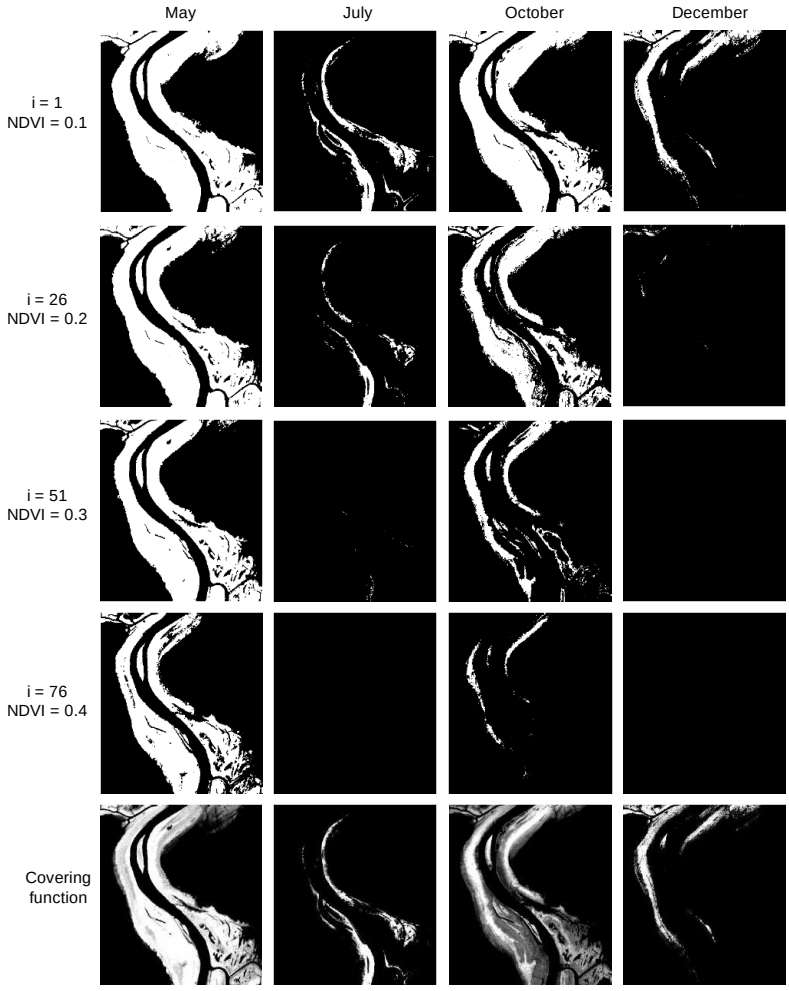

Figure 2.4 Samples of random regions and their covering function. Column 1: May, Column 2: July; Column 3: October, Column 4: December. Row 1: samples at i=1 and NDVI=0.1; Row 2: samples at i=26 and NDVI=0.2; Row 3: samples at i=51 and NDVI=0.3; Row 4: samples at i=76 and NDVI=0.4; pixels in white indicate vegetated area and pixels in black indicate non-vegetated area; Row 5: covering functions.

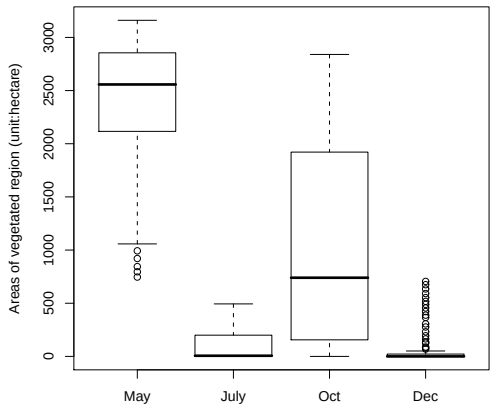

Figure 2.5 Statistical distributions of area of focal elements sampled from four random regions which presenting vegetated area in May, July, October and December.

Figure 2.6 Mean (left) and median (right) of the random region which representing vegetated are in October.

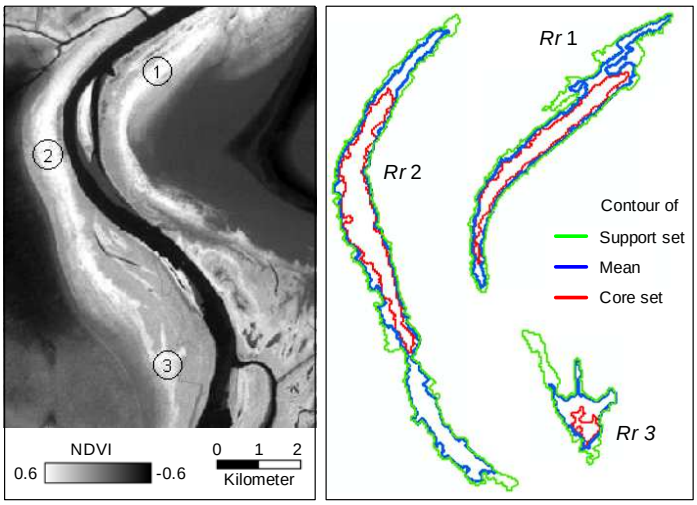

Figure 2.7 Three selected objects on the NDVI image derived from Landsat TM image on October 28, 2004 (left); Contours of Γ_c, Γ_m and Γ_s derived from three random regions.

October in Figure 2.5 has the largest variance and its median skewed to the right, which is consistent with the mean being of a larger area than the median in Figure 2.6.

2.3.3 Example 2: uncertainty modeling for vegetation patch identification

The region growing approach was selected to extract vegetation patches for three example objects in Figure 3.6 from the image on October 28, 2004. Based on a threshold range (minimum and maximum pixel values) or a multiplier of the standard deviation of the original region [137], the region growing algorithm expands from a small seed combining connected pixels within pre-specified limits. Vegetation patches with vague boundaries are sensitive to the setting of parameters in the region growing algorithm. Therefore, by slightly changing the parameters, e.g. under normal distribution with a small value of sigma 0.5, and segmenting iteratively, a set of resulting objects establishes a random set. The covering function can then be determined. We avoid the technical details here, further information can be found in Zhao et al. [198].

Figure 3.6 presents the contours of Γ_c, Γ_m and Γ_s. Apparently, $Rr\ 1$ has similar Γ_c, Γ_m and Γ_s contours, which means that the segmentation

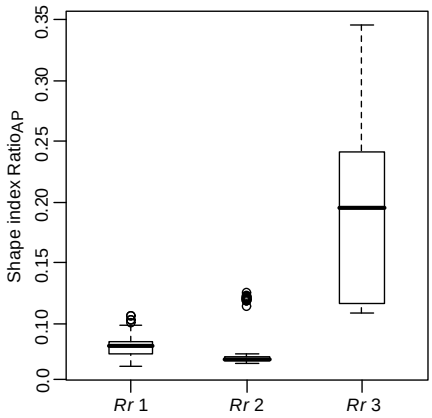

Figure 2.8 Statistical distribution of shape index $Ratio_{AP}$ for all the samples of random regions.

boundaries do not vary much for different parameters. Those for Rr 2 and 3, however, have different extents, implying that they are sensitive to small variation of parameters.

The area-perimeter ratio as a commonly used shape ratio [137], was selected as the shape index to further explore the shape variance of the Rrs. The shape ratio of a sample $Ratio_{AP}(O_i)$ is equal to 4π times the area of O_i and is divided by the square of the perimeter of O_i. The $Ratio_{AP}$ characterizes the deviation from a circular form and changes in scale have no effect. It equals 1 for a disc and it is less than 1 for other shapes. A small value corresponds with a large deviation from a circular shape. We calculated the $Ratio_{AP}$ for all samples and used box-plots to show their statistical distribution in Fig. 3.9. The median values of $Ratio_{AP}$ show that Rr 3 has the smallest deviation from a circular shape and the largest variance, whereas Rr 2 has the smallest variance. For real world objects, there is no apparent reason for the whole boundary of an object to have a constant degree of vagueness. This is the case for Rr 3, where the large variance of the shape of samples indicates that uncertainty in the boundary is not distributed equally in different directions.

2.4 Conclusion and Discussion

In this paper we proposed a stochastic method for modeling uncertainties in earth observation based on random set theory. Usability of random set theory in assessing uncertainties is realized by formalizing

the definitions of random spatial data types. To show that the models are applicable in spatial data analysis in environmental study, we illustrated it with an application in wetland monitoring. Ideas from the theory of random sets are used to obtain means and variances of wetland vegetation area for further exploratory spatial analysis. Random sets are well suited to estimate the broad boundaries of grassland and vegetation patches.

Random sets provide a natural general framework for describing uncertainty, a framework of which many existing formalisms such as probability theory and fuzzy set theory appear as particular cases. The applications in this paper support this selling point of random sets. Note that in the definition of random point, a focal element is a singleton and random set has become a random variable, so that the definition of random point becomes similar with that defined directly on probability theory [e.g. 53]. Moreover, in example 1, we obtain the same results using fuzzy set model by adjusting 0.1 and 0.5 as two parameters in a linear membership function [134], because of the assumption that thresholds are uniformly distributed over the interval $[0.1, 0.5]$. In other words, for some simple cases when a membership function is available, uncertain spatial objects modeled by either random sets or fuzzy theory are similar [101]. Nevertheless, the fuzzy membership function can be difficult to obtain if more complicated assumptions and approaches have to be adopted. In example 2, the empirical covering function is derived from realizations of focal elements which are sampled by the region growing segmentation by using the normal distribution to generate the parameters. Here it is not straightforward to obtain membership anymore, whereas the random set model has a more general applicability. Random sets are also flexible to model dynamic objects in space and time, as an extension to modeling the objects in space. In a previous study [196], we modeled lakes extracted from a time series of images by random sets, where randomness is mainly due to the dynamic process of objects. Therefore, random sets can involve both spatial uncertain objects and spatial-temporal uncertain objects within the same framework.

Random set theory is a theory of a set-valued stochastic variable and serves for rigorous modeling of observed phenomena that can be represented by two dimensional regions. Therefore, it is a straightforward mathematical method for object-based analysis. Individual image pixels contain attribute information of geographical phenomena such as the NDVI, but clusters of pixels forming image objects do not directly express geometric characteristics of the phenomena. In this situation, random sets could be appropriate to model an object as a whole, by set-theoretic operators. In Section 2.3, mean and median of vegetation area are estimated, in which case the vegetated area are considered as a

set of image objects instead of individual pixels.

The set-theoretic approach based on random set theory also provides a method of visualization for positional uncertainty in this wetland application. The variance of vegetation distribution with different NDVI values is reflected in the covering function of random sets. For example, in a wetland vegetation mapping process, we may arbitrarily select a threshold at NDVI=t_1, thus obtaining the support sets and possibly similar spatial distributions for vegetation in May and October. Putting the threshold at t_2, however, we may obtain the core sets, and the mapping results may be largely different. A solution would be to take the mean of the random set rather than thresholding at the arithmetical mean $(t_1 + t_2)/2$. The benefit is that the mean of a random region also considers the uncertainty in a spatial extent.

From the wetlands application point of view, PLNNR managers and decision makers who are concerned about ecosystem dynamics, may become interested in the shape and size variance of lakes and grasslands. Using regular image processing steps to map the grasslands location and calculate the area of lakes are basic operations. Beyond those, newly developed random set approaches may be particularly useful for providing the geometric estimation of objects, especially when they are vague and dynamic on the image observations. Further, based on a time series of random objects extracted from images, interaction between the wetland vegetation and lake can be monitored and explored. Managers who are concerned about the forage and habitat of immigrating birds, i.e. wetland vegetation, may be interested in this information when they make a decision on the water level control of the lake.

In this paper, some concepts of random set theory are introduced and summarized in the context of geoiformation science and environmental science. Despite the great potential of random set theory, these concepts still represent a relatively unfamiliar approach in geoinformation science and hence pose a series of challenges for future research. Definition of general random spatial data model is the first step. The wetland examples in this paper are simple illustration on their feasibility. Many other real world problems may involve further assumptions and complexities. Therefore, the implementation and interpretation of random sets and their applications to other real cases are still necessary.

3 Quantification of extensional uncertainty of segmented image objects by random sets

Published as: Xi Zhao, Alfred Stein, Xiaoling Chen and Xiang Zhang (2011) Quantification of Extensional Uncertainty of Segmented Image Objects by Random Sets. IEEE Transactions on Geoscience and Remote Sensing, 49, pp. 2548-2557.

Abstract

Investigations in data quality and uncertainty modeling are becoming key topics in Geoinformation Science. This study models a collection of outcomes from a standard segmentation algorithm as a random set. It quantifies extensional uncertainties of extracted objects using statistical characteristics of random sets. The approach is applied to a synthetic dataset and vegetation patches in the Poyang Lake area in China. These patches are of interest as they have both sharp and vague boundaries. Results show that random sets provide useful spatial information on uncertainties using their basic parameters like the mean, level sets and variance. The number of iterations to achieve a stable covering function and the sum of the variances are good indicators of boundary sharpness. The coefficient of variation has a positive relation with the degree of uncertainty. An asymmetry ratio reflects the uneven gradual changes along different directions where broad boundaries exist. Our research shows that several characteristics of extensional uncertainty of segmented objects can be quantified numerically and spatially by random sets.

3.1 Introduction

Remotely sensed images provide important information sources for studies in landscape patches. In particular an accurate determination of their extent is a prerequisite for further exploring the relationship between the spatial pattern and the underlying ecological processes of these patches [185]. Gradual boundaries in natural landscapes are regarded as transitional zones, intergrade or ecotones and play functional roles in ecosystems e.g. by controlling the flux of materials [47, 63]. The delineation and representation of natural patches with inherently vague extensions such as forest stands, grasslands and lakes, is a recurrent topic in ecological [74, 92] and Geo-Science studies[16, 46]. Their ignorance or their reduction to lines as often occurs on conventional maps may be a gross simplification. To avoid this simplification, several conceptual models and spatial data models have been proposed to represent geographical phenomena with uncertain boundaries [9, 23, 32]. But few of them provide a method to realize boundary extraction in remote sensing applications as in [159].

Image segmentation is essential in image processing when extracting image objects, as it takes attribute homogeneity and spatial contiguity of image pixels into account. Natural patches, however, usually appear in remotely sensed images as approximately homogeneous collections of pixels delimited by gradual changing boundaries. Uncertainties easily occur when identifying the spatial extents of the vague objects. Recently, uncertainties have been addressed not only in pixel-based manner [132], but also in image segmentation, by developing soft segmentation methods [91, 107], by assessing the effects of uncertainties in landscape studies [4, 126], or by representing and modeling uncertainties of spatial objects [41, 140, 160]. Saha and Udupa [138] utilized class uncertainty and region homogeneity information to improve the estimation of optimum thresholds in segmentation. Beutner III *et al.* [10] estimated uncertainty in MRI-based brain region delineations provided by segmentation methods to automatically detect poor image quality. Lucieer and Stein [94] developed boundary stability index to quantify and visualize the existential uncertainty of spatial objects derived by a split-and-merge image segmentation.

In this paper, we focus on extensional uncertainty, i.e. the uncertainty in identifying geometric elements that describe the spatial extent of an object [112]. On the one hand, extensional uncertainty is tightly related to the uncertain boundary issue. Due to the extensional uncertainties, spatial objects with uncertain boundaries usually face difficulties in determining their spatial extents when extracted from images. To quantify the extensional uncertainty of objects, uncertain boundaries of image objects need to be extracted and modeled first. Based on the analysis of

gradual changes from interior to boundary and to exterior, information on their extensional uncertainty can be obtained and provides more details on the heterogeneity and gradualness of the object boundary. On the other hand, extensional uncertainty also implies existential uncertainty, because if an object has an uncertain extension its existence is uncertain as well. Inversely, an object of which it is uncertain whether it exists cannot have a well-defined spatial extension. Therefore, quantification of extensional uncertainty in image segmentation does not only provide information about the spatial distribution of the boundary, but also contributes to a better understanding of the properties of spatial objects such as their existence and extent.

Many approaches have been proposed to deal with extensional uncertainty, mainly based on fuzzy set theory (e.g. [141, 178, 180]) and probabilistic theory (e.g. [68]). Fuzzy set based methods are grounded on the idea that fuzzy membership values are related to the degree of uncertainty, however, determination of the membership function for a given concept or a spatial object is still a challenge [134]. Wang and Hall [180] proposed that the membership function should correspond with sharpness in changes and calculate it by the first order derivative of the property of interest. Van de Vlag and Stein [178] modeled beach area as fuzzy objects where membership degrees are in terms of the thematic definition of class "beach". Schiewe and Ehlers [141] determined the fuzzy widths of interested vague objects based on the mean deviations derived from repeatedly digitizing boundaries from aerial photos.

Methods based on probability theory represent uncertainty with probability distribution functions (pdfs). Stochastic simulation is now widely used for pdfs that cannot be estimated by analytical approaches (e.g.[154]). It generates multiple realizations of the underlying spatial phenomenon, representing the uncertainty of the simulated variables [67]. For example, the uncertain extent of a lake is simulated by connecting multiple points on the boundary which vary under uncertainty [68]. As such, modeling the positional uncertainty of the lake object requires a separate pdf for each primitive point, together with the statistical dependencies between these points. Due to the large number of primitive points, complexity of the pdf and the assumptions on statistical dependence, however, it is usually difficult to estimate distribution of the simulated object as a whole. As an alternative, we consider the whole object as a set, so its extensional uncertainty can be modeled by a randomly varying set instead of modeling the individual boundary points. Therefore, probabilistic set theory is required beyond conventional probability theory which deals with a small number of random variables.

A random set is a generalization of a random variable, taking sets as its elements. Random set theory acts as the core of stochastic geometry.

It deals with stochastic models of irregular or random geometrical structures [104, 110]. Since it provides a foundation for set-theoretic statistical approaches, random sets have been successfully employed for developing image segmentation methods [12, 37] as well as for the study of randomly varying geometrical shapes [169]. Zhao *et al.* [197] modeled the uncertain boundaries of natural landscape by random sets. This was based on the assumption that interpreting larger uncertain boundaries e.g. by digitizing, thresholding or segmentation will result in bigger differences and more randomness in different outputs. The overall set of outputs can be modeled as a random set, and we can then use its variability to reflect the degree of uncertainty. This paper takes the exploration further to realize the boundary extraction method and quantify extensional uncertainties of segmented image objects modeled by random sets. To do so, we first examine the effect of a small degree of randomness of segmentation parameters on the spatial variance of extracted image objects. Secondly, we model extracted objects by random sets and quantify their extensional uncertainties by newly proposed indices. This method is tested by a synthetic dataset and then applied to a wetland application in the Poyang Lake National Natural Reserve in China.

3.2 Methods

3.2.1 Random set and spatial random region

On the Euclidean space R^n, a random set Γ associates a function $Pr_\Gamma(x) : R^n \to [0,1]$, $x \in R^n$. This function is called the covering function of the random set, taking values between 0 and 1. The covering function can be interpreted as the probability of the element x on space R^n being covered by the random set Γ [197].

We denote the space of an image as $I \subset R^2$ where pixel x is the basic element, $x \in I$. The measure space that carries random spatial data models is denoted by U, where $U \subseteq \mathcal{P}(R^2)$. A random region ($Rr$) is defined as a random set on U and can be presented by the focal elements O_i with corresponding uncertainty assignments m_i, denoted as a collection of pairs $\{O_i, m_i\}$ [197]. The focal elements are regions which are subsets of R^2: $O_i \in \mathcal{P}(R^2)$. If a random region consists of n focal elements with equal probability, then $m_i = \frac{1}{n}$. For example, n interpreters digitizing a vegetation patch from an image may result in n different polygons, denoted as $O_1, \ldots, O_n$, being samples as well as focal elements of the random region. The probability that pixel $x \in R^2$ is occupied by the random region can be estimated by $Pr_{\cup_{i=1}^{n} O_i}(x)$. Not that hereafter, we refer to the random set Γ, random object or random

region as random region.

An estimator of the covering function of random set Γ can be obtained as:

$$\hat{Pr}_\Gamma(x) = \frac{1}{n}\sum_{i=1}^{n} I_{O_i}(x), \quad x \in R^2 \quad O_i \in U \tag{3.1}$$

where I_{O_i} is the indicator function:

$$I_{O_i} = \begin{cases} 1, & x \in O_i \\ 0, & x \notin O_i \end{cases}$$

The set $\Gamma_p = \{x \in R^2, 0 \le p \le 1 : Pr_\Gamma(x) \ge p\}$ is called a p-level set. For example, the median is the 0.5-level set. Other special cases are the support set $\Gamma_s = \{x \in R^2 : Pr_\Gamma(x) > 0\}$, describing the possible part of Γ and the core set $\Gamma_c = \{x \in R^2 : Pr_\Gamma(x) = 1\}$, describing the certain part of Γ.

The mean Γ_m of the random set, has been defined in several ways [169]. Here we choose the Vorob'ev expectation, because it considers sets with a finite number of points, such as sets of pixels, which are of interest in image analysis [169, p.113]. The mean area EA of the random set Γ is then defined as $\mathrm{E}A(\Gamma) = \int_{R^2} Pr_\Gamma(x)\mathrm{d}x$. According to the definition of Vorob'ev expectation, the set Γ_m equals

$$\Gamma_m = \{x \in R^2, 0 \le p_m \le 1 : Pr_\Gamma(x) \ge p_m\} \tag{3.2}$$

where p_m is such that Γ_m has the area E$A(\Gamma)$. If p_m is not unique then it is set equal to the infimum of all such p_m. When $p_m = \frac{1}{2}$, the mean and median are identical.

The corresponding set-theoretic variance of a random set is defined as:

$$\Gamma_{var}(x) = \mathrm{E}(I_{O_i}(x) - Pr_\Gamma(x))^2. \tag{3.3}$$

3.2.2 Random set generation by region growing

Region growing belongs to the region-based segmentation procedures. It assumes that neighboring pixels with similar values should be segmented into the same region [137]. The generation of random sets in our study utilizes the region growing algorithm in the ENVI/IDL software . This algorithm is designed for one input band and needs either one or two user-supplied parameters: a multiplier of the standard deviation of the seed region or the minimum and maximum values of a threshold

www.ittvis.com

range. Based on these parameters, it starts with assigned seeds and analyzes the spectral and spatial properties of pixels. Next, it grows regions by merging neighboring pixels satisfying the homogeneity requirements into homogeneous regions. Several image segmentation procedures exist able to generate random set models as described below. We chose this relatively simple algorithm to demonstrate the application of random sets for quantifying the extensional uncertainty of image objects in segmentation results. This, to our knowledge, has not been done before.

The general idea of random set generation is that the extents of segmented patches should be more sensitive to the variance of parameters in the region growing algorithm when extracting image objects with larger extensional uncertainty. Therefore, by slightly changing parameters and segmenting iteratively, we obtain a set of objects from iterations and construct a random set of which the covering function can then be determined. A growing seed is identified interactively inside the central interior part of the target object and the main parameter of segmentation (denoted as PAR) is initialized by a user-supplied value PAR_0. PAR_0 can be the parameter found experimentally to give the best results in a previous conventional segmentation. Intuitively, values closer to the optimal parameter PAR_0 should be more reliable and thus have more chances to be the preference of users. Therefore we adopt a normal distribution as our proposal distribution, centered about the PAR_0, with a preset variance σ^2, to generate PAR_i. Then we decide on a small number ϵ, e.g. $\epsilon = 10^{-6}$, which is used to stop the iterations. We argue that the variance of a random set becomes stable as iteration times increases, so if the difference of two covering functions between successive iterations (denotes as d_i) is smaller than ϵ, the procedure stops. For the value of σ^2 in the normal distribution, we can easily modify it and repeat the iteration procedure so that the effect of σ^2 on the qualification can be tested by the sensitivity analysis. The main steps are illustrated in Fig. 3.1 and the procedure of random set generation for one object is described as follows:

1. Generate a random number PAR_i $(i > 0)$, from a normal distribution with mean $\mu = PAR_0$ and a preset variance σ^2.
2. Start the segmentation and obtain a sample region O_i and its indicator function I_{PAR_i}.
3. Calculate the covering function $f^{(i)}$ of the random set $\{O_1, \ldots, O_i\}$ by $f^{(i)}(x) = \frac{1}{i}\sum_{j=1}^{i} I_{PAR_j}(x)$ according to equation 4.2.
4. Stop the algorithm at step n_ϵ, if $f^{(i)}$ is stable. The stop criterion is $d_i = \int_{R^2}(f^{(i)}(x) - f^{(i-1)}(x))^2 \mathrm{d}x < \epsilon$ $(i \geq 2)$. If this condition is fulfilled then the algorithm stops at $n_\epsilon = i$; otherwise, repeat for $i + 1$.

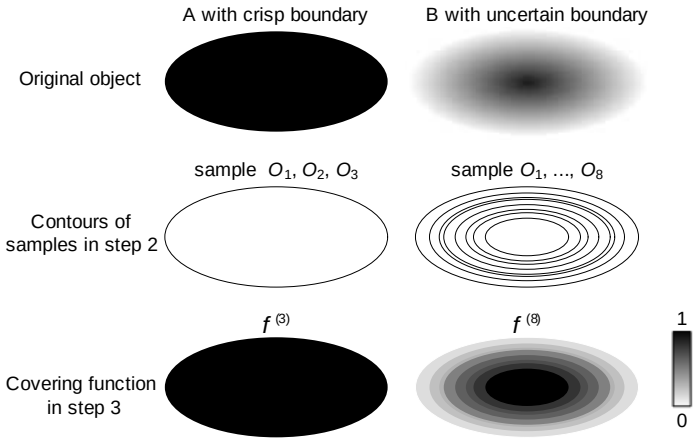

Figure 3.1 Main steps of random set generations for crisp object A (left column) and uncertain object B (right column). The gray scale in presenting the original objects indicates different intensities of the pixels; the gray scale is also used to indicate values of covering functions from 0 to 1.

Note that the first covering function $f^{(1)}$ depends on PAR_1. If PAR_2 is extremely close to PAR_1, i.e., the segmented samples O_1 and O_2 are similar, then $f^{(1)} = f^{(2)}$ and $d_2 < \epsilon$ which means the algorithm should stop. Such a situation can occur by chance. To avoid this, we set the iteration step $n_\epsilon > 2$. The resulting collection of samples $\{O_1, \ldots, O_{n_\epsilon}\}$ has its empirical covering function $f^{(i)}$ in each iteration, which also approximates the mean of the family of indicator functions. The log of the differences between two successive covering functions, i.e. $\log_{10}(d_i)$ provides a convergence curve for random sets as a function of the iteration step i.

3.2.3 Quantifying statistical parameters of random regions

The level sets are used to reflect the spatial distribution of the varying sizes of the random sets to quantify extensional uncertainty of segmented objects. The Vorob'ev mean of random set (Equation 3.2) is estimated by first determining the mean area $\hat{A}(\Gamma) = \int_{R^2} \hat{Pr}_\Gamma(x)\mathrm{d}x$ and then finding a p_m-level set with an area equal to $\hat{A}(\Gamma)$.

The corresponding set-theoretic variance (Equation 3.3) is estimated as $\hat{\Gamma}_{var}(x) = \frac{1}{n_\epsilon}\sum_{i=1}^{n_\epsilon}(I_{PAR_i}(x) - \hat{Pr}_\Gamma(x))^2$. Pixels with $\Gamma_{var}(x) = 0$ are expected in the certain part of an object, whereas high values are expected in the boundary area. The sum of Γ_{var}, denoted as SD, is

defined as:

$$SD = \int_{\Gamma_s} \Gamma_{var}(x)dx. \tag{3.4}$$

The coefficient of variation (CV) is used as a normalized and dimensionless measure:

$$CV = \frac{\int_{\Gamma_s} \sqrt{\Gamma_{var}(x)}dx}{\int_{\Gamma_s} \hat{Pr}_{\Gamma}(x)dx}. \tag{3.5}$$

The CV summarizes the dispersion of the distribution of a random set and allowing us to make comparisons with other objects. A high CV indicates a larger proportion of objects with a high Γ_{var} and thus points to a large extensional uncertainty.

3.2.4 Describing location and shape characteristics

The centroid (or center of gravity) is a commonly used landmark point for locating and tracking objects. Its coordinates are determined by averaging the coordinates of each pixel covered by the object [137]:

$$X_c = \frac{\sum_{i=1}^{N} X_{x_i}}{N} \qquad Y_c = \frac{\sum_{i=1}^{N} Y_{x_i}}{N} \tag{3.6}$$

where the N is the number of pixels that the object occupies and X_{x_i} and Y_{x_i} denote the $X-$ and $Y-$ coordinate of pixel x_i. The centroid of a non-convex object may occur outside the bound of the object, and hence it may not serve to represent the absolute location of the object. In this study we adopt the centroid for tracking the relative shift of the location of a random region. After determining the centroids of the samples $O_1, \ldots, O_{n_\epsilon}$ belonging to each random region, we superimpose the centroids of random regions for comparison purpose. The distances between centroids of samples for each random region are not changed, so that relative variance of random regions are visualized.

According to the different intensities of the covering function, we can assign different weights to each pixel and calculate the density-weighted centroid [137] as:

$$X_{cw} = \frac{\sum \hat{Pr}_{\Gamma}(x_i) \cdot X_{x_i}}{\sum \hat{Pr}_{\Gamma}(x_i)} \quad Y_{cw} = \frac{\sum \hat{Pr}_{\Gamma}(x_i) \cdot Y_{x_i}}{\sum \hat{Pr}_{\Gamma}(x_i)} \tag{3.7}$$

where $\hat{Pr}_{\Gamma}(x_i)$ is the empirical covering function at pixel x_i. For an object with a perfectly symmetrical covering function, the centroid of

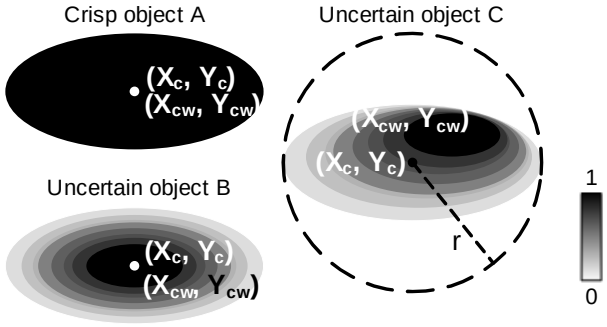

Figure 3.2 Symmetric property of covering functions shown by distance between centroid $(X_c,\ Y_c)$ and density-weighted centroid $(X_{cw},\ Y_{cw})$. Crisp object A and uncertain object B have symmetric covering functions; uncertain object C has asymmetric covering function and its bounding circle and radius r are shown in dash lines. The gray scale indicates values of covering functions from 0 to 1.

its support set and density-weighted centroid coincide (e.g. object A and B in Fig. 3.2). If the covering function has an asymmetric distribution, however, because some parts of the boundary are crisp and other parts are changing gradually, the distance between these two points increases. e.g. object C in Fig. 3.2. We normalize this distance by dividing by the radius of the bounding circle (r), thus producing a ratio between 0 and 1 that measures the asymmetry [137]:

$$Ratio_{ASM} = \frac{\sqrt{(X_c - X_{cw})^2 + (Y_c - Y_{cw})^2}}{r}. \tag{3.8}$$

The radius of the bounding circle can be determined by half the largest distance between any two pixels in Γ_s.

There are various ways to generate a dimensionless shape descriptor. The area-perimeter ratio is a commonly used shape ratio, which can be used to explore the relationship between extensional uncertainty and shape variance of random sets. The shape ratio of a sample O_i is defined as:

$$Ratio_{AP}(O_i) = \frac{4\pi * Area(O_i)}{Peri(O_i)^2} \tag{3.9}$$

where $Area(O_i)$ and $Peri(O_i)$ are the area and perimeter of the sample O_i. $Ratio_{AP}$ characterizes deviation from circular form and is scale invariant. It equals to 1 for a disc and reduces to lower values for an increasing deviation from circularity.

3.3 Experiments

To demonstrate the generalization and assess the validity of the proposed method, we conduct two experiments on different parameters of the algorithm and on different data. Values for the parameters given in the experiments are found experimentally to give the best results for the type of images used. Since quantitative ground truth for assessing extensional uncertainty is difficult to obtain, the first experiment uses simulated blurry images. The real image in the second experiment uses local expert knowledge and a time series analysis to assess the results. In this section, we abbreviate image object a as (a) and the random region used to model (a) as $Rr(a)$.

3.3.1 Synthetic data

We first consider a synthesized test set, consisting of eight test images of 100 ×100 pixels (Fig. 3.3). $(a)-(c)$ result from blurring the disc (d). The left and right arcs of $(ba)-(bd)$ are derived from combining blur boundaries of $(a)-(d)$. For example, (ba) has its left arc from (b) and its right arc from (a). By successively blurring these objects, we test the association between proposed measures and increasing extensional uncertainty. We use the threshold range $[min, max]$ as the parameter in the region growing algorithm and need to generate random numbers for both minimum threshold min and maximum threshold max. But in these simple synthetic data, the center of target objects are in black (pixel values equal to 0) and the boundary pixels are in grey (pixel values between 0 and 1). All the pixels with value 0 definitely belong to target objects, therefore, we set the threshold range at $[0, max]$ to extract the objects. The boundary is uncertain, which means only one parameter max need to generate random numbers. So we fix the minimum threshold equal to 0 and allocate the maximum threshold max to PAR. We use the values $PAR_0 = 0.6$, $\sigma = 0.5$ and $\epsilon = 0.1, 0.01$ for random set generation.

3.3.2 Experimental results on synthetic data

Fig. 3.3 shows the eight image objects and their corresponding random set models presented by the contours of Γ_c, Γ_m, Γ_s. Crisp object (d) has the same extents for Γ_c, Γ_m and Γ_s, indicating no extensional uncertainty. In contrast, the blurry object (a) with a vague boundary shows clear differences among the areas of Γ_c, Γ_m and Γ_s. Larger difference among the contours correspond to large uncertainties of the extensional area of an object.

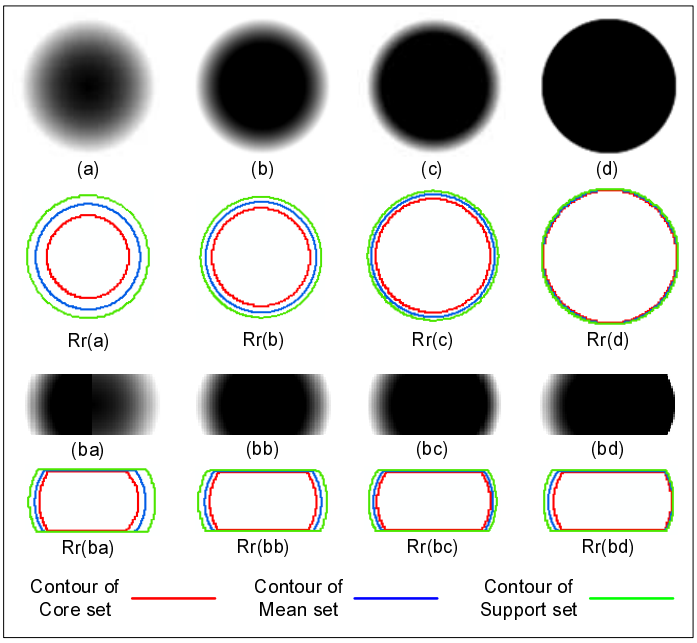

Figure 3.3 Eight synthetic image objects with different uncertain boundaries in gray scale and their corresponding random sets represented by contours of Γ_c, Γ_m and Γ_s.

Several indicators are reported in Table 3.1. $n_{0.1}$, $n_{0.01}$, SD and CV decrease respectively from $Rr(a)$ to $Rr(d)$ and from $Rr(ba)$ to $Rr(bd)$. Since (d) has a crisp boundary without extensional uncertainty, the number of iteration steps for both $n_{0.1}$ and $n_{0.01}$ are equal to the minimum setting value 3, and both SD and CV are equal to 0. Although $Rr(c)$ has a larger SD than $Rr(ba)$, its area is also larger than that of $Rr(ba)$, resulting in a lower CV. The $Ratio_{ASM}$ equals 0 for $Rr(a)-(d)$ and $Rr(bb)$, all having perfect symmetric boundaries. (ba) and (bd) have larger differences in their right arc compared to their left arc than (bc), resulting in larger $Ratio_{ASM}$ than (bc).

3.3.3 Real data

The second experiment analyzes a Landsat TM image with 30m spatial resolution showing vegetation patches with vague and relatively crisp boundaries within a wetland region within the Poyang Lake National Nature Reserve (PLNNR) (115° 55′-116° 03′E, 29° 05′-29° 15′N), in China. This reserve is characterized by wetlands grassland and lakes. Two types

Table 3.1 Quantification of extensional uncertainty of eight synthetic objects by numeric indicators

Object	$n_{0.1}$[a]	$n_{0.01}$[b]	SD[c]	CV[d]	$Ratio_{ASM}$[e]
(a)	42	111	280	0.22	0
(b)	39	90	179	0.11	0
(c)	31	68	111	0.06	0
(d)	3	3	0	0	0
(ba)	25	61	74	0.08	0.04
(bb)	17	60	54	0.05	0
(bc)	14	50	43	0.04	0.01
(bd)	12	35	30	0.03	0.03

[a] the number of iteration steps to get difference $d_i < 0.1$
[b] the number of iteration steps to get difference $d_i < 0.01$
[c] the sum of set-theoretic variance Γ_{var}
[d] coefficient of variation
[e] asymmetry index for covering function

of emerged vegetation dominate the wetlands grassland: *Miscanthus* dominant vegetation communities occur in high elevation areas and *Carex* dominant communities in low elevation areas. These vegetation types are all blooming in spring and serve as important habitats and forages for spring migration birds. When winter migration birds arrive in autumn, *Miscanthus* dominant community in high elevation areas become senescent. *Carex* dominant communities in low elevation areas are turning green and are growing up gradually. They allow different kinds of winter migration birds forage on leaves and rhizomes in different elevation zones [186]. Our experiment is carried out in the area surrounding Banghu lake within PLNNR where the wetlands grassland has the largest area and shows obvious transition zones (Fig. 3.4).

A Landsat TM image of 28 October 2004 is chosen to separate the two general vegetation communities by using their different autumn phenologies. We confine ourselves to the two-class problem of separating foreground (*Carex* vegetation patches) from background (other senescent vegetation and water). A topographic map of scale 1:10000 is used as the geographic reference data for image geometric correction. The root mean squared error of the geometric correction is less than 10m.

The Normalized Difference Vegetation Index (NDVI) as a commonly used vegetation index to characterize vegetation biomass, coverage,

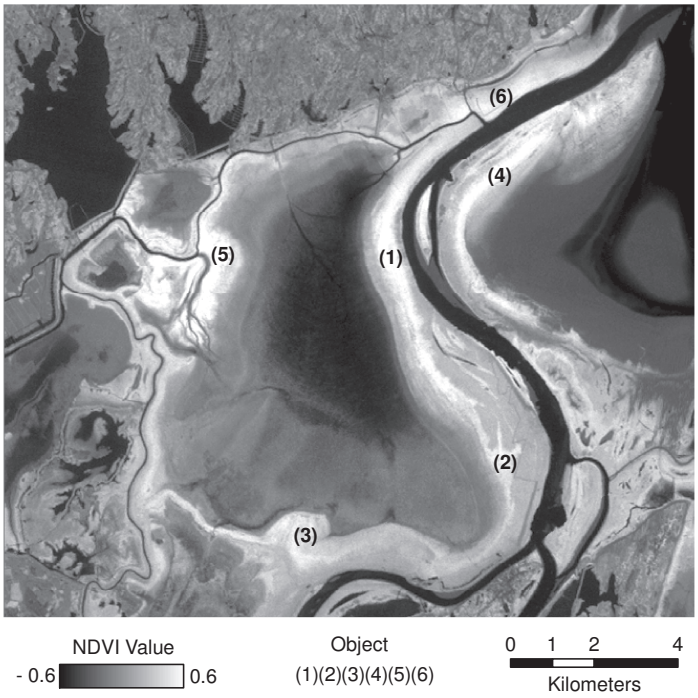

Figure 3.4 Six vegetated areas with high NDVI on the NDVI map of PLNNR derived from Landsate TM image on October 28, 2004 are selected for the experiment on real data.

height and other growing status, provides ideal information to classify general vegetation types. Since it reflects the main spectral characteristics of a vegetated area and can be stored in a single band, we adopt NDVI map derived from red and near-infrared bands as the input image for region growing segmentation. For simplifying the experiment, we select six individual *Carex* vegetation patches (Fig. 3.4) and set the initial seeds manually. Later we will give some recommendations in the discussion section for the case of multiple overlapping objects. In this experiment, we explore the standard deviation multiplier as PAR in the region growing algorithm. The following values are used: $PAR_0 = 3$, $\sigma = 0.5$ and $\epsilon = 0.1, 0.01$. Other choices for σ are used in a sensitivity analysis below.

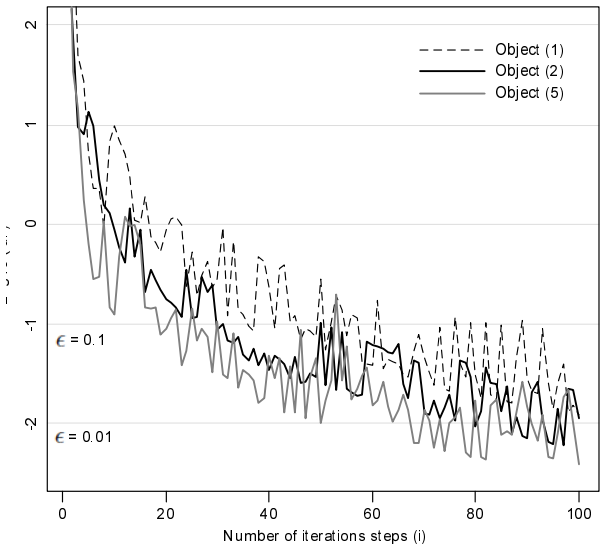

Figure 3.5 Differences between two successive covering functions d_i at logarithmic scale are plotted and form the convergence curves of (1), (2) and (5). Two small numbers $\epsilon = 0.1$ and 0.01 are different criteria for stopping iteration procedures.

3.3.4 Experimental results on real data

Fig. 3.5 shows that (1), (2) and (5) have different convergence curves. Different ϵ values are set for stopping random sets generation and for indicating that the convergence of random sets is achieved by then. The convergence curve of (5) reaches $\epsilon = 0.1$ for the lowest number of iterations, followed by (2) and (1). For smaller criteria $\epsilon = 0.01$, they keep the same order. Iteration steps for all the six objects are reported in Table 3.2. For an object with little extensional uncertainty, NDVI values are homogeneous. The resulting samples $O_1, \ldots, O_i$ have similar extents, irrespective of the segmentation parameters PAR shifts. In other words, its convergence curve rapidly decreases to a small value near or equal to zero and remains stable. In contrast, the convergence curve for an object with a gradual transitional boundary stabilizes slowly.

After segmentation, regions resulting from several iterations form random sets. Fig. 3.6 shows the contours of Γ_c, Γ_m and Γ_s for six Rrs. Apparently, $Rr(5)$ has similar Γ_c, Γ_m and Γ_s contours, i.e. the segmentation boundaries show little variation for different PARs. Those for $Rr(1)$, however, have different extents, implying that they are sensitive to small variation of PARs. This information well explains the observation in

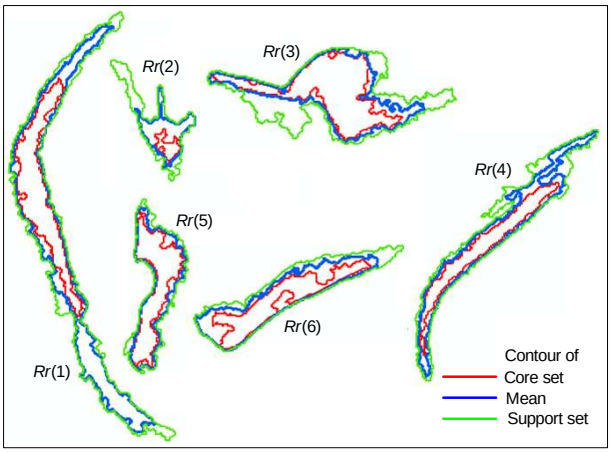

Figure 3.6 Contours of Γ_c, Γ_m and Γ_s derived from six random regions.

Fig. 3.5 that $Rr(1)$ requires more iterations than $Rr(5)$ to achieve stable covering function. We can also obtain this information directly from the set-valued variance in Fig. 3.7. More pixels with a positive Γ_{var} in the object indicate a larger uncertain area. SD values are larger for $Rr(1)$ than for $Rr(2)$ and $Rr(5)$ (Table 3.2), because the number of pixels with non-zero Γ_{var} values in $Rr(1)$ is largest. $Rr(5)$ has a smaller CV than $Rr(1)$ and $Rr(2)$, which again coincides with what we observed from Fig. 3.5 and Fig. 3.6. As the CV takes the object area into account, the value of CV of $Rr(2)$ is larger than that of $Rr(1)$ which indicates that $Rr(2)$ has a larger proportion of uncertain area.

Fig. 3.8 shows that $Rr(5)$ has centroids that cluster most strongly, whereas centroids of $Rr(1)$ and $Rr(2)$ have a relatively large variance. This supports again that (5) has a more crisp boundary than (1) and (2). Centroids of $Rr(3)$, $Rr(4)$ and $Rr(6)$ all have obvious alignments along specific directions. It indicates that objects have the largest extension in that direction and that there is a limitation in a perpendicular direction. For example, in Fig. 3.4, the northwest, southwest and southeast sides of (6) are cut by rivers and the only flexible direction for grass growth is the northeast direction.

The $Ratio_{ASM}$ (Equation 3.8) indicates the asymmetry property of the covering function. An object with a symmetric covering function can have either a crisp or a vague boundary. For a vague boundary, it should have the same gradients in every directions, to make a random set with concentric p-level sets. The covering function of $Rr(2)$ is highly asymmetrical (Table 3.2). $Rr(1)$, $Rr(3)$ and $Rr(5)$ with relatively low values however, have symmetric covering functions. Therefore, their

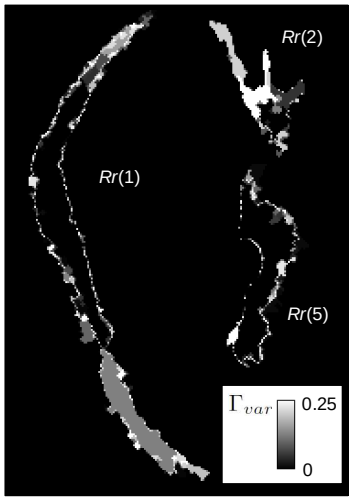

Figure 3.7 Example of set-valued variance of random regions in gray-scale. Pixels have high Γ_{var} are bright, while those with low Γ_{var} are dark.

Table 3.2 Quantification of extensional uncertainty of six objects by numeric indicators in real data experiment

Object	$n_{0.1}$[a]	$n_{0.01}$[b]	SD[c]	CV[d]	$Ratio_{ASM}$[e]
(1)	42	99	274	0.25	0.03
(2)	30	79	117	0.51	0.11
(3)	25	74	145	0.18	0.03
(4)	25	75	148	0.24	0.06
(5)	19	49	69	0.10	0.02
(6)	25	75	114	0.18	0.09

[a] the number of iteration steps to get difference $d_i < 0.1$
[b] the number of iteration steps to get difference $d_i < 0.01$
[c] the sum of set-theoretic variance Γ_{var}
[d] coefficient of variation
[e] asymmetry index for covering function

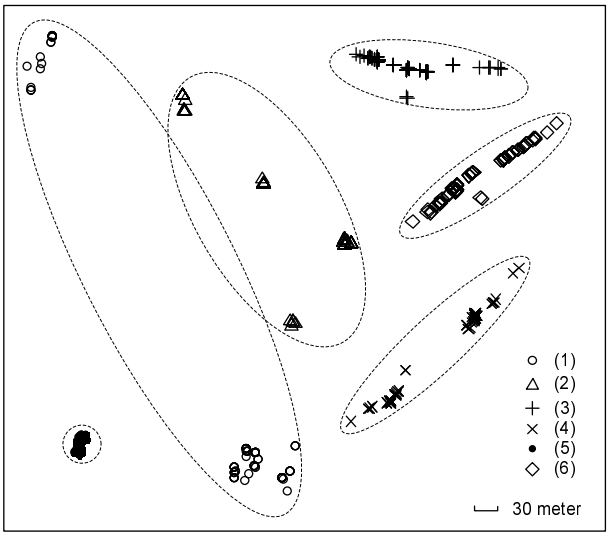

Figure 3.8 Spatial variance of centroids for six random regions. The locations of random regions are superimposed and distances between random regions are meaningless. The size of dash ellipse and distance between centroids of samples for a random region indicate the variance of the location of this random region.

samples of random sets are much more concentric.

In Fig. 3.9, the median values of $Ratio_{AP}$ show that $Rr(6)$ has the smallest deviation from a circular form, followed by $Rr(5)$, $Rr(2)$ and $Rr(3)$. Both (1) and (4) are in obviously elongate shapes having smallest $Ratio_{AP}$ values. $Rr(2)$ has the largest standard deviations of $Ratio_{AP}$ and $Rr(1)$ has the smallest value. This shows that shape variability has no absolute relation with extensional uncertainty. But the shape mean and variance may be useful for users interested in a shape discriminant for classification.

The standard deviation σ of the randomized parameter PAR is chosen to be equal to 0.5 in the above analysis. This value generates 95% of PAR_i within the range of $[PAR_0 - 2\sigma, PAR_0 + 2\sigma]$, resulting in meaningful vegetation segmentation in this experiment. For a sensitivity analysis of this choice, we repeat the random sets generation procedure, by setting different σ values between 0 and 1. The number of steps $n_{0.01}$ needed to achieve stable covering function and CV of random sets are calculated for each σ. In Fig. 3.10, both $n_{0.01}$ and CV increase as σ increases, indicating an increasing extensional uncertainty of segmented random regions. For most of σ in the top figure, $Rr(1)$ and $Rr(5)$ keep

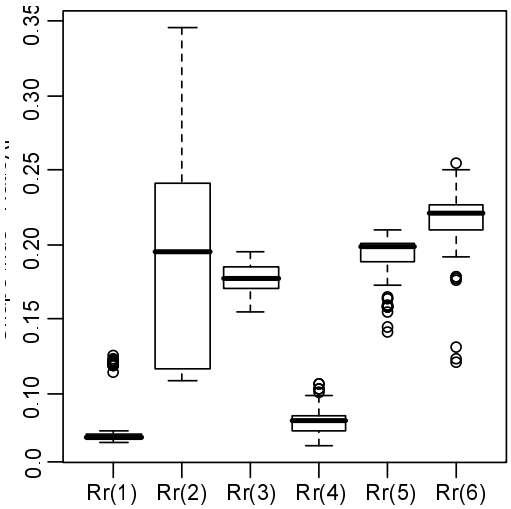

Figure 3.9 Statistical distribution of shape index $Ratio_{AP}$ for all the samples of random regions.

the same ranks among all the six Rrs, while some other Rrs have close values of $n_{0.01}$ at some points and change orders. In the bottom figure, $Rr(2)$ and $Rr(5)$ keep more stable ranks than the other Rrs. But at $\sigma = 0.6$ or 0.7, the CV of $Rr(2)$ shows sudden jumps, and $Rr(3)$ and $Rr(5)$ show fluctuations at $\sigma = 0.9$. The occurrence of these jumps at high σ values implies that moderate shifting of segmentation parameter is preferable for an uncertainty quantification, and that setting a too small or a too large σ is not meaningful.

Since the wetland ecosystem in the study area has seasonal dynamics and because historical field data are not available, validation of the results is difficult to perform. But if we follow an object in a time period, then the temporal changes of its extensional uncertainty modeled by random sets should coincide with the vegetation phenology and field knowledge from PLNNR managers. Such consistency is considered as a partial validation in this experiment. Another two Landsat TM images acquired at November 29th and December 15th in 2004 are used for this purpose. Geometric correction and atmospheric correction are carried out, using the image at October 28th as the reference. The segmentation procedure described in Section 3.2.2 is used to extract (4). Then we modeled extensional uncertainties by random sets and followed their changes in Fig. 3.11. (4) belongs to *Carex* dominant vegetation patch. From Fig. 3.11, we observe that the Γ_c of $Rr(4)$ is shrinking from October to December. This is because the number of pixels with high

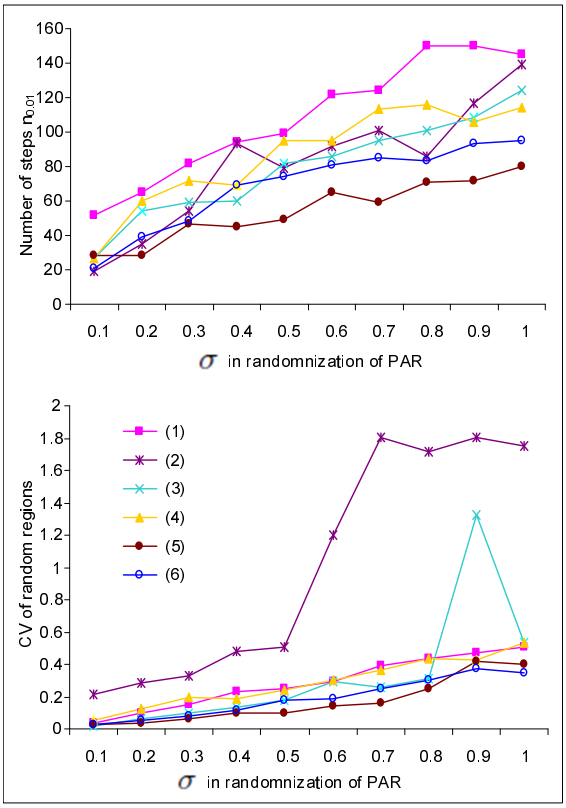

Figure 3.10 Changes of iteration steps $n_{0.01}$ (top) and CV of random regions (bottom) with different σ used in generation of randomized parameter PAR.

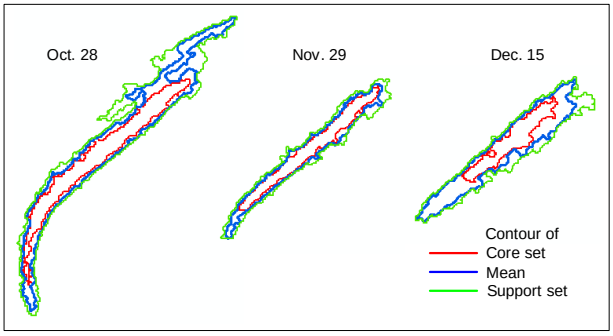

Figure 3.11 Object (4) modeled by random regions at three different image observation times in year 2004.

NDVI intensities reduces on images from autumn to winter. Γ_m and Γ_s in October have convex corner in the north, due to the mixture of *Carex* dominant communities and *Miscanthus* at high elevation near the river. Γ_m and Γ_s in December have convex edges in the south, reflecting the mixture of older *Carex* dominant communities and young germinating *Carex* at low elevation near the lake. All the above observations are consistent with our field experiences in PLNNR.

3.4 Discussion and conclusion

This study interprets extensional uncertainty of extracted image objects in terms of random sets, with the aim of summarizing the essential properties of their uncertain boundaries, using a small number of statistically representative and experimentally accessible parameters. Compared with our previous study [197] which generally introduced the random sets method in spatial uncertainty modeling, this study has made random sets generation and statistics applicable in image segmentation practices. By using the statistical parameters (Γ_c, Γ_m, Γ_s and Γ_{var}) of random sets and the spatial distribution of centroids, we demonstrate that randomness of segmentation parameters has different effects on extracted features when objects have different extensional uncertainties. We also propose indicators, including n_ϵ, SD and CV, to summarize the degree of extensional uncertainty numerically. Objects with relative clear boundaries have covering functions that become stable rapidly as iteration increases and have small SD values. But objects with the same value of SD can have very different CV values if their areas have large differences. Therefore, we can use n_ϵ and SD for indicating the absolute amount of uncertainty, and CV for the relative amount of uncertainty per area unit. Since CV are determined irrespective to the object area, it can be used as general comparison measures.

Mapping centroids of random regions provides the information on the possible variance of the location of objects visually. In change detection or object tracking analysis conducted on series of images, centroid movements of an object within a small range may be due to the difference of a segmentation parameter instead of real movement of the object. Therefore, we need to consider the possible effect of uncertainty occur in segmentation, especially for dynamic objects with uncertain boundaries.

The asymmetry property of covering function reflects the different gradients of transition zones in different positions or directions of the object boundary. In reality, there is no apparent reason for the whole boundary of an object to have a constant degree of vagueness. In most situations, the boundary of an object is indeterminate at one place and

more determinate at other places [160]. For example, some part of a vegetation patch can be cut, due to a river or abrupt terrain elevation at that position, whereas other sides of the patch have gradual transitions to the background. This property can be observed by visualization of the object, whereas could use the $Ratio_{ASM}$ to quantify it successfully.

In this study, we chose a probabilistic approach based on random set theory to model uncertainty. The empirical covering function for each pixel in random sets can also be interpreted as a fuzzy membership value, thus indicating the possibility of the pixel covered by the segmentation object. A similar suggestion is made in [169]. In addition, α-cuts as an important concept of a fuzzy set, used as an efficient interpretation tool to describe the internal structure of a fuzzy boundary, can be viewed as a nested random set in our approach [3, 182]. Therefore, random sets modeling can serve as an alternative procedure to obtain fuzzy membership functions and α-cuts. But, more importantly, our approach allows one to quantify extensional uncertainties directly by random sets, without resorting to other concepts like α-cut. Actually, random set theory is a generalization of other uncertainty theories such as probability, fuzzy set, possibility and Dempster-Shafer theory [58, 101]. In all, random set theory with its roots in probability theory is a solid and reliable way to describe spatial uncertainty.

The region growing algorithm used in this study generally produces good outputs for a random sets analysis. In total, fourteen image objects with different uncertain boundaries have been tested in two experiments, including synthetic blurry circles and natural vegetation patches. The effectiveness and generalization of the proposed method need to be investigated by more experiments on various kinds of spatial objects on different images. In the real data experiment in this study, the region growing algorithm can also be applied on the entire image by setting seeds randomly. In that case, neighboring seeds may grow overlapped random sets if they are close. The random set method, until now, is designed for analyzing individual objects, so whether one object is overlapped with another does not influence the uncertainty quantification procedure for each object. To deal with the overlapping situation, we may turn to studying the topological relationship of uncertain objects, which is out of the scope of this paper. As an alternative, we can also investigate how random set methods can be combined with other segmentation methods which consider space partition such as split-and-merge algorithm. In those cases, other types of uncertainties such as existential uncertainty will also appear [94], and more classes besides vegetation and non-vegetation are then to be determined.

In principle it is possible to apply of the proposed method on multi-bands imagery as well. One approach is to reduce the dimension of the original image bands. This can be done e.g. by principal components or

by a band combination like the NDVI. A second approach is to consider region growing segmentation on individual bands and combine the random sets thus obtained using logical operators. At this stage the best way to proceed is unknown and further exploration is required.

Modeling the extensional uncertainty offers detailed information on the spatial distribution of objects, especially for objects with uncertain boundaries. This information can be important for managers who attempt to locate the resources precisely and calculate the areas or volumes. The gradual changes in boundaries of *Carex* patches reflect the gradient of different wetland vegetation on the bank of lakes. Therefore, moments of random sets and derived indices can be used to characterize the vagueness, diversity and heterogeneity in the grassland boundary zones. The extracted vegetation patches were modeled by random sets in the present study for the purpose of demonstration. The developed method can be applied as well to other objects, such as lakes. For a multi-temporal analysis, changes in the extents and gradualness of wetland vegetation or water areas can be tracked from a series of random sets and indices.

4

Accuracy assessment of extensional uncertainty modeled by random sets

Published as: Xi Zhao, Alfred Stein, Tiejun Wang, Xiaoling Chen and Liqiao Tian (In Press) Accuracy assessment of extensional uncertainty modeled by random sets. Advances in Geo-Spatial Information Science. Shi, Goodchild, Lees & Leung (eds), Taylor & Francis.

Abstract

Random sets have recently been developed for classifying and representing natural landscapes. This study addresses the accuracy assessment of classified or segmented uncertain objects, in particular for modeling the extensional uncertainty of a vegetation patch. Correlation analysis was used to determine the relationship between covering function and measured variables. Statistical distance and Pearson's chi-square test were adopted for testing the similarity of their distributions. The confusion matrix and the k statistic report the thematic accuracy of the classifications. The results show that significant correlations exist between a covering function of random set model and actual vegetation coverage, and thus the probability distributions are similar. The accuracy of the random set varies from good to moderate according to the kappa coefficients derived from the confusion matrixes of support, mean, median, and core sets.

4.1 Introduction

Conventional pixel-based and object-based classification approaches generate maps with exclusive categories. These hard classifications are designed for mapping discrete objects and clear bounded land cover, but they are not appropriate for mapping continuous landscapes in nature such as wetlands [183]. The limitations of hard classification in representing transition zones and uncertain boundaries has been a motivating factor for the development of alternative approaches based on uncertainty handling theories such as fuzzy set theory [191] and random set theory [24, 104]. Several soft classification and uncertainty modeling methods have been developed for classifying and representing natural landscapes such as beaches [178] and grassland [197], and for modeling dynamic phenomena such as fire spread [179] and flooding [166].

The accuracy assessment of soft classification, however, remains a theoretical and practical difficulty. Stehman and Czaplewski [162] and Foody [43] give overviews on accuracy assessment. Accuracy assessment of image objects extracted by object-based classification or segmentation is challenging [194], especially when objects are uncertain [167]. The main difficulty is that the information about the uncertainty represented in the classification results does not necessarily correspond with objects in the field. This is because even on the ground, the delineation of uncertain objects may be impossible, due to vague and ambiguous boundaries. In addition, detailed reference data that is critical for validating soft classification and uncertainty modeling results, are often unavailable, especially when field data have mainly been collected for validating hard classification.

Extensional uncertainty refers to the uncertainty in identifying the geometric elements that describe the spatial extent of the object [112]. The random set model has been employed as a foundation for the study of randomly varying geometrical shapes [169] and introduced by Zhao et al. [198] to represent the extensional uncertainty of image objects extracted from Landsat TM images. The validation of the random set models in [198], however, relies only on sensitivity analysis and partial support from temporal changes analysis. Direct accuracy assessment is lacking due to the absence of detailed and synchronous field data.

In this study, a clean and almost synchronous HJ-1A image was acquired with a comparable spatial resolution (i.e., 30m) two days before the survey and the sampling plan was designed specifically for a) investigating the zonal pattern of wetland grassland and b) assessing the accuracy of a random set model. Since the random set model represents the extensional uncertainty by giving probability values instead of assigning thematic classes, results need to be grouped to assess thematic

accuracy. In addition, the statistical distribution of the random set is expected to be close to the distribution of some variables collected in the field, so that statistical methods can be utilized for measuring and testing the statistical aspects of the model. The objective of this research is twofold: (1) to explore the corresponding measurable variables which were collected on the ground for validating the uncertain image objects modeled by random sets, (2) to quantify the quality of the random set modeling results.

4.2 Methods

4.2.1 Study area and ground survey

The study area PLNNR (115° 55′-116° 03′E, 29° 05′-29° 15′N) is located in the northwest of Poyang Lake in Jiangxi province, central China. Nine lakes in PLNNR are connected to Poyang Lake during high water levels in summer and disconnected when water levels are low in spring, autumn and winter. Several wetland plant species grow in spring and serve as important food sources for migratory birds. During the summer flooding, grasses growing at high elevations (e.g., *Miscanthus*) remain above the water, whereas sedges (e.g., *Carex*) and aquatic species (e.g., *Potamogeton*) at lower elevations are submerged. After the flood, winter migratory birds arrive in autumn from late September, whereas only a few plant species such as *Carex* start to turn green again and thrive before the winter. Other vegetation communities become senescent in autumn and dead in winter. When sedges at lower elevations are shooting up gradually, different kinds of birds forage on the leaves and rhizomes of young sedges and rhizomes of submerged aquatic species in different elevation zones. Taller sedges also provide habitat and shelter for birds [186].

Wetland grassland has the largest zonal width and takes up the largest area around Banghu lake within the PLNNR. For this reason the ground survey was carried out around Banghu lake to investigate the zonal pattern of wetland grassland from October 26th until November 6th, 2009. To verify the gradual change of vegetation, four typical transects L1-L4 were designed, starting from the river bank and perpendicularly crossing different zones until they reach the lake bank of Banghu (Fig. 4.1). These four transects contain 27, 24, 14, 8 sample plots, so that 73 plots were sampled in total.

The sample plots, of size 30×30m, were distributed evenly along transects, and fixed by measuring tape. The location of each plot was measured by a GPS at the centre of the square, with accuracy within 10m. Within the plots, the following variables were recorded: land cover types,

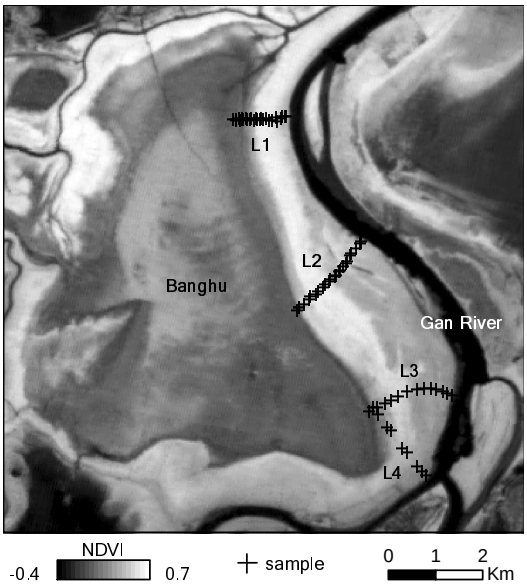

Figure 4.1 Distribution of sample plots along four transects (L1-L4) in Banghu Lake as indicated on the NDVI map derived from HJ-1A image.

vegetation types (communities and species), vegetation height, and percent vegetation cover. Five subplots (1×1m) were also established within each 30×30m plot to measure the vegetation characteristics in detail, such as height and cover. The spectral characteristics of typical vegetation types were measured within the 1×1m subplots using an SVC field-portable spectroradiometer. For each recorded output of the spectroradiometer, approximately 30 separate measurements were taken, which were then averaged for each 1×1m subplot.

4.2.2 Image segmentation and extensional uncertainty modeling

The HJ-1A/1B satellites, launched on September 6, 2008, are mainly used for environment and disaster monitoring in China. The first images were obtained on September 8; the two-satellites-constellation was established on October 13 and passed the in-orbit-test review on February 19, 2009 after more than 5 months of in-orbit testing. The high spatial resolution (30m), short revisiting time (2 days) and its orbit characteristics (sun-synchronous) make HJ-1A/1B CCD a unique sensor for dynamically capturing information on wetlands on the meso- or local scale. To facilitate further spatio-temporal research on the dynamics of Poyang Lake using this free data source, the random set model on a

HJ-1A image was tested.

One HJ-1A image on October 24, 2009 covering Poyang Lake was downloaded from the China Centre for Resource Satellite Data and Applications (CRESDA). As a second level product, radiometric correction and systematic geometric correction were applied before downloading, providing an image with a UTM WGS84 projection. The image was a subset to the full extent of PLNNR and a topographic map of scale 1:10,000 in PLNNR was used as the geographic reference data for more accurate geometric correction. The root mean square error of the geometric correction is less than 10 m. The Normalized Difference Vegetation Index (NDVI) is calculated for the HJ image using

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{4.1}$$

where RED and NIR stand for the spectral reflectance measurements acquired in the red and near-infrared band, respectively.

In this research, the target object was the *Carex* patch located on the east bank of Banghu (Fig. 4.1). Since *Carex* is the dominant green vegetation in October, an NDVI image was used as input for the region growing segmentation. Based on a threshold range (minimum and maximum pixel values), the region growing algorithm expands from a small seed combining connected pixels within pre-specified limits [137]. Vegetation patches with vague boundaries are sensitive to the setting of parameters in the region growing algorithm. Therefore, by slightly changing the parameters, e.g., under normal distribution with a small value of variance, and segmenting iteratively, a set of resulting objects was established as a random set [198]. The image object with extensional uncertainty is thus modeled by a generated random set.

The procedure is as follows:

1. A growing seed is selected inside the central interior part of the target object.
2. The parameters are initialized, i.e., the upper and lower thresholds. For each parameter, random numbers are generated from a normal distribution with the initialized parameter as the mean and a preset variance.
3. Segmentation starts by using one random number, resulting in an object O_i as one sample of the random set.
4. Iterating the above step several times results in a set of objects that jointly establishes a random set.

Several characteristics of random sets can be used to describe the extensional uncertainty of objects. For example, n polygons resulting from n times of segmentation are samples of the random set, denoted

as $O_1, \ldots, O_n$. The probability that pixel $x \in R^2$ is occupied by the random region can be estimated by $Pr_{\cup_{i=1}^{n} O_i}(x)$. An estimator of the covering function of random set Γ is obtained from:

$$\hat{Pr}_{\Gamma}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{O_i}(x), \quad x \in R^2 \tag{4.2}$$

where $I_{O_i}(x)$ is the indicator function of $O_i(x)$. The covering function can be interpreted as the probability of the pixel x on space R^2 being covered by the random set. All the pixels with covering function equal to or larger than p construct a p-level set of the random set. The 0-level, 0.5-level and 1-level sets are called support, median and core sets respectively. The Vorob'ov mean is the p-level set whose area is equal to the accumulated area of the covering function. Further theoretical details about random set models and technical details about segmentation are found in previous work by the authors [197, 198].

4.2.3 Accuracy assessment

To validate the modeling result, all 73 sample plots were used for accuracy assessment. Correlation between several variables collected in the field and the covering function was first explored by regression models. To rate the fit of the regression models, the p-value of the t-test for each explanatory variable, the F value and its significance level, and adjusted squared correlations (R_a^2) were calculated. The probability distribution of the covering function and measured variables were then compared to identify the most similar pair. The statistical distance was calculated by two methods: Kullback-Leibler divergence and total variation distance:

$$\text{Kullback} - \text{Leibler divergence} = \sum_{x} P(x) \log_2 \frac{P(x)}{Q(x)} \tag{4.3}$$

$$\text{Total variation distance} = \sum_{x} |P(x) - Q(x)|. \tag{4.4}$$

Typically, P represents the true distribution of data, observations, or a precise calculated theoretical distribution. The measure Q represents a theory, model, description, or approximation of P. In Equations 4.3 and 4.4, $P(x)$ is the probability of x in the distribution of measured variables, and $Q(x)$ is the probability of x in the distribution of the covering function. In this study, the distributions were estimated from the 73 observations. The variable with the lowest distance has the probability distribution most similar to that of the covering function.

In addition, the samples were divided to four groups according to the 0-25, 25-50, 50-75 and 75-100 percentiles of the covering function and measured variables respectively, and the number of samples falling in each corresponding group was compared. Pearson's chi-square test was adopted to test the similarity of the four groups of the covering function and measured variables. This tests a null hypothesis that the proportions of samples in the four groups are similar for the covering function and the variables.

To measure the thematic accuracy of the random set, the covering function was thresholded to obtain several important p-level sets such as support, mean, median and core, and those sets were validated using confusion matrices and the k statistic. The overall accuracy (OA), producer accuracy (PA), user accuracy (UA) derived from the confusion matrix provide different summaries of the information: OA is a measure of the total match between reference and classification data, taking no account of the sources of error (errors of omission and commission). To assess the accuracy of the individual categories, the PA related to errors of omission was found, together with the UA related to errors of commission. The k statistic indicates whether the confusion matrices are significantly different from a random result. In general, $k < 0.4$ indicates poor performance of a classifier, k between 0.4 and 0.59 indicates moderate performance and $k > 0.6$ suggests a good performance [103].

4.3 Results

4.3.1 Ground survey

From the river bank to the lake bank, transect L1 is approximately 1200m long, whereas the other three transects L2, L3 and L4 are approximately 2000m long. Three different zones occur along all four transects. The first zone is on the river bank. Flowered *Miscanthus* of 1-2m height appears at high elevations near the river bank, often mixed with *Cynodon*, *Carex*, *Polygonum*, and human planted poplar. Some of the flowered *Miscanthus* also have green leaves at lower height, and some shorter *Miscanthus* are not flowered. These are called green *Miscanthus* in this paper. In areas where human activities are intensive, the *Miscanthus* has usually been harvested to approximately 0.5m. The second zone is *Carex* dominant, approximately 500m across in horizontal distance. The height of *Carex* ranges from 0.3 to 0.6m, and it thrives in the autumn, with very high density. Toward the lake bank, the height and density of *Carex* decreases and *Polygonum*, *Artemisia* and *Eleocharis* appear and are mixed together. The third zone is near the lake bank where elevation changes gradually. The indicators of low elevation are high soil

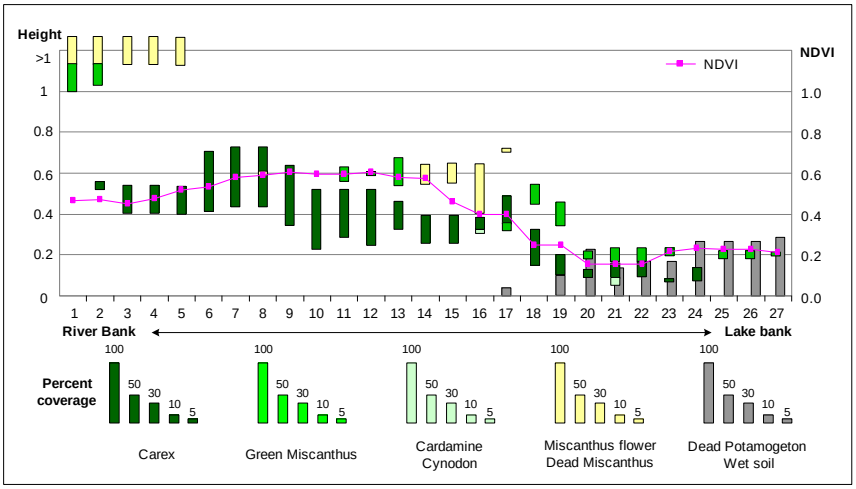

Figure 4.2 Types of dominant vegetation, their percent coverage and heights along transect L1 are compared with NDVI extracted from HJ image at corresponding pixels.

moisture and plant such as *Cardamine* and young *Carex*. The farthest place reached on the lake bank in L1 is covered by 10cm of shallow water with dead *Potamogeton* and *Vallisneria* beneath. Bird droppings and feathers were frequently found and birdcalls were very evident. On the bank of L2-4, we found wet soil, *Cardamine*, dead and dry *Potamogeton* and *Vallisneria* covering to the soil, and shooting up *Carex* with very low density.

Figure 4.2 shows the dominant vegetation types and their coverage at 27 sample plots along transect L1. As illustrated in the legend, the length of the bar indicates the percentage cover of vegetation averaged from 5 subplots. The average height of each vegetation type can be read from the centre point of the bar using the scale on the left. For the three zones categorized above, samples 1 to 5 belong to the first zone, and samples 20 to 27 to the third zone. Samples 6 to 12 are relatively homogenous and belong to the second zone, whereas samples 13 to 17 are in a transition area. From the NDVI values extracted from the HJ image at corresponding sample plots, we found the NDVI reach a peak of around 0.6 for samples 8 to 12, where either homogenous *Carex* plots appear or *Carex* mixed with green *Miscanthus*. As the *Carex* becomes shorter toward the lake, the coverage of wet soil and dead *Potamogeton* increases and NDVI values reduce to 0.2 and remain stable.

Table 4.1 compares field measured NDVI values of vegetation, to which different vegetation types, coverage and heights contribute differ-

Table 4.1 Comparison of NDVI values of vegetation under different status within 1×1 subplots, derived from field spectroradiometer.

Carex	Coverage (%)	100	70	30	5
	Height (cm)	40-60	10	15	5
	NDVI	0.89	0.56	0.34	0.21
Miscanthus flower|green	Coverage (%)	100	100	100	100
	Height (cm)	120|100	70|55	70|30	110|0
	NDVI	0.73	0.69	0.58	0.25
Artemisia	Coverage (%)	100	100		
	Height (cm)	40	25		
	NDVI	0.79	0.59		

ently. Firstly, plots covered by a low percentage of vegetation will have a low NDVI. The cases of *Carex* with a coverage of 100%, 70%, 30% and 5% show decreasing NDVI values. Secondly, vegetation with different heights may have the same NDVI. For example, pure *Carex* plots with heights between 0.4 and 0.6 m have the same NDVI value of 0.89. The possible reason is that *Carex* thrives extremely well at those heights with a high growth density. This causes the NDVI to be almost saturated and not sensitive to differences in plant heights. Thirdly, for cases where NDVI does not achieve saturation, e.g. four *Miscanthus* plots with 100% coverage, the heights of the flowered and green leaves parts of *Miscanthus* do have an impact on the NDVI. The flowered parts of *Miscanthus* are dry in autumn, thus having a low NDVI of approximately 0.25. For subplots where flowered and green leaved *Miscanthus* both exist, the NDVI values reduce as heights of the green leaved sections decrease. Finally, vegetation growing density also influences the NDVI value. For example, Artemisia is a plant species that also thrives in autumn, but at a low proportion and usually mixed with *Carex*. For plots fully covered by *Carex* and Artemisia at the same height of 0.4 m, Artemisia has a lower NDVI of 0.79 as compared to 0.89 for *Carex* alone. The possible reason is that coverage percentage only reflects a proportion of the projected canopy of vegetation on the ground, so that *Carex* with high density has a larger NDVI than Artemisia.

From the above qualitative analysis, *Carex* coverage is found to be the variable most out-standing, in providing high values of NDVI. Other variables such as the height of *Carex* also contribute to the NDVI values. They may, therefore, act as corresponding variables to covering function derived from the random set model.

4.3.2 Extensional uncertainty modelled by random sets

To select the parameters in random region growing algorithm, we calculated the mean and standard deviation of the NDVI values for the 68 samples in which *Carex* coverage was larger than 0. Pixels with pure *Carex* have maximum NDVI values around 0.6. The average NDVI of the 68 samples equals 0.44 and the standard deviation equals 0.13. Since plots with NDVI > 0.6 are dominated by *Carex*, 0.7 was fixed as the maximum threshold to simplify the random set generation procedure. We further selected 0.44 as the initialized minimum threshold, and generated 200 random numbers from a normal distribution with a mean of 0.44 and standard deviation of 0.13. Finally, the growing seed was placed at the location of one sample plot where 100 percent *Carex* was recorded, and 200 randomized threshold intervals were used to obtain 200 objects and model them as a random set Γ.

The main characteristics of the random set Γ were estimated, including the covering function, the support set, the median set, the mean set, the core set and the variance. The support set illustrated in Figure 4.3a includes all pixels with probability values above 0.2, indicating the possible spatial extension of this *Carex* patch. The reason for adjusting the support set to the 0.2-level is given in section 3.5 below. The median and mean sets correspond to the 0.5-level set and the 0.32-level set respectively. The pixels with a covering function value larger than 0.95 are enclosed by the contour of the core set which almost ensures that the pixel belongs to a *Carex* patch. The differences between the spatial extension of the support and core sets indicate the extensional uncertainty of the *Carex* patch. The higher uncertainty corresponds to a higher variance of the random set in Figure 4.3b.

4.3.3 Correlation between covering function and measured variables

The covering function of Γ is derived from the region growing segmentation using NDVI data of the HJ image. Since many variables measured in the field are closely related to the NDVI, such as coverage and height, their relationships with the covering function were further explored. Table 4.3 summarizes five linear regression models in which the covering function is the response variable. In model 1, four variables were tested, and only two of them: *Carex* coverage and *Carex* height have p-values below a 0.05 significance level. The non-significant p-values of *Miscanthus* coverage and *Miscanthus* height suggest that they can be removed from the model. The model was refitted without *Miscanthus* coverage and *Miscanthus* height in model 2 and an R_a^2 value equal to 0.57 and a high F value were obtained. Apparently, 57% of the variance

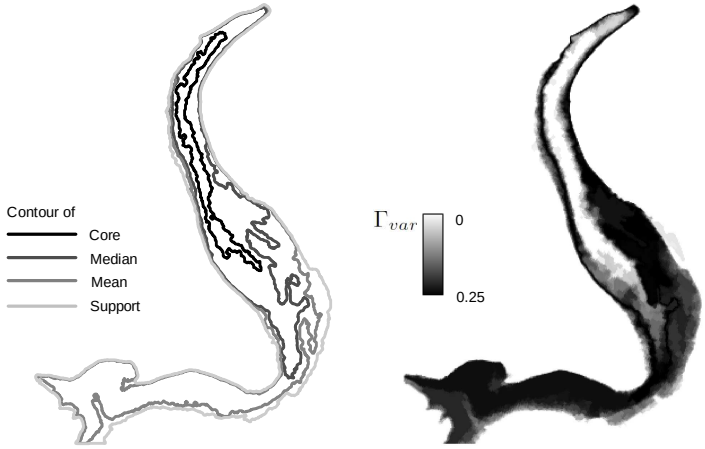

Figure 4.3 Extracted object and its extensional uncertainty described by concepts from random set theory: (a) core set, median, mean and support set; (b) variance.

Table 4.2 R_a^2-values and F-statistics of the correlation relationships between covering function and different explanatory variables.

	Explanatory variables	R_a^2	F value
Model 1	*Carex* coverage***, *Carex* height*, *Miscanthus* coverage, *Miscanthus* height	0.58	84.74***
Model 2	*Carex* coverage***, *Carex* height*	0.57	104***
Model 3	*Carex* volume (i.e. coverage $\times$ height)***	0.59	48.55***
Model 4	*Carex* coverage***	0.54	26.31***
Model 5	*Carex* height***	0.49	70.4***

*** indicates p-value under a 0.001 significance level
* indicates p-value under a 0.05 significance level

can be explained by the model and the F value shows that the model is statistically significant. In model 3, coverage was multiplied by height to obtain the volume of *Carex*, which is used as the single explanatory variable in the simple regression model. The value of R_a^2 equals 0.59, which is the highest among all five models. Models 4 and 5 test how the Carex coverage and height, as single variables respectively relate to the covering function. The R_a^2 and F values show that *Carex* coverage can explain the variance of the covering function well and *Carex* height explains variation of the covering function less well.

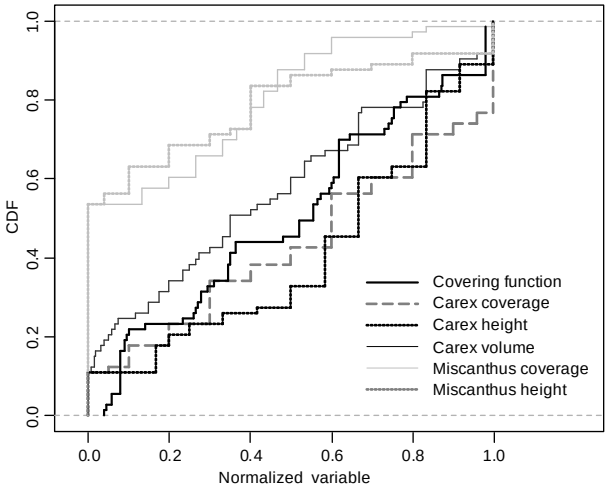

Figure 4.4 Empirical cumulative probability curves.

4.3.4 Assessment of probability distribution similarity

To test whether the probability distribution of the random set can reflect the distributions of the measured variables, the empirical cumulative probability functions of the random set and the measured variables estimated from the 73 sample plots are plotted in Fig 4.4. From visual comparison, curves of *Carex* coverage, height and volume are closer to the covering function curve than those of *Miscanthus* coverage and height. The Kullback-Leibler divergence and total variation distance reported in Table 4.3 confirm that the distribution of random set best matches the distribution of *Carex* volume, followed by those for *Carex* coverage and height. The above results coincide with the conclusions drawn above from the regression models in section 4.3.3.

The degree of similarity between *Carex* volume, *Carex* coverage, *Carex* height and the covering function for the four percentile groups was tested by Pearson's χ^2 test and is reported in Table 4.4. The p-value of tests 1 and 2 are larger than the significance level 0.001, hence failing to reject the null hypothesis that the distributions of the test variables are similar. The p-value of test 3 is smaller than 0.001, and apparently rejects the null hypothesis on similarity, based on the percentile distinction. The test result shows that, the frequency distribution for the 73 samples observed in the four percentile groups for *Carex* coverage and volume are similar to those for the covering function, whereas the *Carex* height is significantly different. *Carex* height, therefore, is excluded as a variable in the following validation analysis.

Table 4.3 Statistical distance between covering function and different explanatory variables.

Reference distribution $P(x)$ of	Distribution of the model $Q(x)$	Kullback-Leibler divergence	Total variation distance
Carex volume	Covering function of random set	0.02	0.15
Carex coverage		0.06	0.25
Carex height		0.11	0.36
Miscanthus coverage		0.71	0.73
Miscanthus height		0.60	0.76

Table 4.4 Pearson's χ^2 tests on four percentiles groups of covering function and relevant variables.

Percentiles	≥ 0%-<25%	≥ 25%-<50%	≥ 50%-<75%	≥ 75%-≤ 100%	p-value
Covering function	25	16	15	17	
Carex volume	19	18	19	17	
Carex coverage	18	21	16	18	
Carex height	33	23	8	9	
Test 1:	Covering function & *Carex* volume				0.39
Test 2:	Covering function & *Carex* coverage				0.26
Test 3:	Covering function & *Carex* height				0.0006***

*** indicates p-value under a 0.001 significance level

4.3.5 Assessment of the support, mean, median and core sets for classifications

To validate the uncertain object from the classification point of view, the coverage percentage of *Carex* recorded at 73 sample plots was selected as the single variable for group testing samples. The *Carex* volume explains a little more about the proportion of the variance of the covering function, but in practice, it is not straightforward to decide the thresholds for making hard classes of the *Carex* volume for accuracy assessment. In contrast, the vegetation coverage thresholds of such as 0, 20, 50, 80, 100 percents are natural choices for categorizing different coverage levels. Secondly, the vegetation coverage information can be an easily measured variable in the field, whereas height measurement may not be easy or possible. The volume estimated, therefore, by multiplying the coverage by height is also difficult to obtain. Thirdly, the coverage information can be easily calculated from high resolution reference

Table 4.5 Accuracy assessment of the support set determined by different p-level sets.

p-level set, p =	0.05	0.10	0.15	0.20	0.25
OA (%)	87.67	82.19	83.56	86.30	84.93
Lowest UA or PA (%)	11.11	50.00	52.94	70.59	66.67
Kappa coefficient	0.14	0.44	0.50	0.62	0.59

images, but the vegetation height, thus the volume, is more difficult to extract from images. Therefore, only the *Carex* coverage variable is used to report on the classification accuracy assessment.

Due to the limitation of NDVI in sparsely vegetated areas, low p-level sets of the random set derived from NDVI maps, for example the 0.05-level set, cannot be regarded as accurately describing the extent of vegetation with coverage larger than 5%. In general, NDVI does not work well for sparsely vegetated areas with less than 15% -20% coverage (Purevdorj et al. 1998), and therefore, the 0.2-level set is expected to be the appropriate support set with high mapping accuracy for *Carex* coverage above 20%. In Table 4.5, the 73 samples are divided into two classes: *Carex* coverage above 20% and below 20%, and the test classification accuracies for different low p-level sets are shown. The result shows that all the OA, lowest UA or PA and k are highest for the 0.2-level set. Therefore, the 0.2-level set was adopted as the adjusted support set in this study.

Accuracy assessment was also applied to the core set by comparing with samples which were more than 95% covered by *Carex*, and the mean and median set by comparing with samples which were more than 50% covered by *Carex*. Table 4.6 details the mapping accuracy as given by OA, PA, UA and k derived from confusion matrixes. The highest overall accuracy is achieved by the support set with OA of 86.3% and k = 0.62. According to Mather [103], the support set and the mean set have good k values, whereas the median set and the core set have moderate k values. The lowest PA occurs in the class "presence" of the core set, showing that there is more area of *Carex* in the field than indicated by the core set. The mean set has a higher accuracy than the median set when mapping *Carex* coverage larger than 50%. To provide an overall assessment, a three class confusion matrix was constructed: *Carex* coverage below 20%, coverage above 95% and coverage between 20% and 95%. The overall accuracy and k coefficient derived from the three class confusion matrix equal 71.2% and 0.51, respectively.

Table 4.6 Comparison of OA, UA, PA and k coefficients for the support set, mean set, median set and core set.

	Support 0.2-level set		Mean 0.32-level set	
Class Coverage (%)	Presence ≥ 20	Absence < 20	Presence ≥ 50	Absence < 50
PA (%)	91.07	70.59	88.89	71.43
UA (%)	91.07	70.59	83.33	80.00
OA (%)	86.30		82.19	
k	0.62		0.62	

	Median 0.5-level set		Core 0.95-level set	
Class Coverage (%)	Presence ≥ 50	Absence < 50	Presence ≥ 95	Absence < 95
PA (%)	77.78	82.14	47.37	98.15
UA (%)	87.50	69.70	90.00	84.13
OA (%)	79.45		84.93	
k	0.58		0.54	

4.4 Discussion and conclusions

In this research, we applied the random set model for representing uncertain boundary of a wetland plant species (i.e., *Carex*) patch, and perform an accuracy assessment on the modeling results. We find that *Carex* volume and coverage can be the corresponding measurements on the ground, by which the covering function of random sets can be quantified and interpreted adequately. The accuracy of the random set varies from good to moderate according to the k coefficients derived from the confusion matrixes of support, mean, median, and core sets. The support set has a higher accuracy than the core set after making some adjustments, whereas the mean set has a better performance than the median set.

To extract a *Carex* patch, the autumn season was selected when *Carex* thrives and contributes most to a high NDVI value whereas other wetland vegetation is dry or dead. In conventional crisp segmentation, the extracted *Carex* patch is represented by a crisp boundary. The area inside the boundary is regarded as dominated by *Carex* and the area outside is not. This hard partition of continuously vegetated area

ignores the extensional uncertainty, if looking at the original HJ image where pixel values change gradually from one class to the other or in the field where changes between different types of vegetation dominance are smooth. The random set model takes the extensional uncertainty into account, by representing the randomness around human interpreted boundaries of vegetation. In this study, the randomness is propagated, in our recognition and modeling process, from the real world to the object world in geoscience, by means of taking an image and making segmentation. In addition, the randomness involved in the former step influences the next step in the process. For example, the vegetation type, coverage, height, and soil background, all contribute to the NDVI values of image pixels. If using NDVI as the input of the segmentation, the extracted boundary of a vegetation patch/object will be influenced by the randomness in NDVI, resulting in extensional uncertainty.

Regression analysis and statistical distance calculation were applied to explore which factors are correlated to the extensional uncertainty modeled by random set and whether their probability distributions are similar. The result shows that *Carex* volume has the most significant relationship with covering function, followed by *Carex* coverage. The distribution of the covering function, however best reflects the distribution of *Carex* volume based on their closest statistical distance. Therefore, the random set best describes the randomness of *Carex* volume among the variables collected in the field. Pearson's χ^2 tests also confirm this conclusion. Since the sample size i.e. 73 pixels is not large, four percentiles groups were used in the Pearson's χ^2 tests, to avoid any grouping bias that may occur when small samples are divided into large groups.

To simplify the accuracy assessment by confusion matrix, *Carex* coverage only was selected as the reference variable. The other variables, such as vegetation type, height and density, also influence the modeling result, and should be considered together. These variables, however, may belong to different scales, such as nominal (e.g. vegetation type), ordinal (e.g. big or small density) and ratio scale (e.g. height and coverage). Hence, they are difficult to integrate into one general variable which might be a better match with the covering function of a random set. The classification accuracy of the random set model is assessed by OA, UA, PA and ęÊ. The accuracy of the support set is better, after a reasonable adjustment, than that of the core set. This result highlights the fact that the accuracy of a random set model for quantifying extensional uncertainty will be affected by the quality of the input data. In the other words, the uncertainty modeled by random set in this study is only the randomness brought about by the segmentation process, and does not include the randomness in the earlier image recording step.

5 Monitoring the spatial dynamics of wetland inundation by random sets

Published as: Xi Zhao, Alfred Stein and Xiaoling Chen (2011) Monitoring the dynamics of wetland inundation by random sets on multi-temporal images. Remote Sensing of Environment, 115, pp. 2390-2401.

Abstract

General models are required for better understanding the dynamics of wetland inundations and their water regimes. Extraction by imagery snapshots and use of a crisp data model do not value the inherent uncertainties in space and time. This study addresses parameterization of a mixed Gaussian random set model in a multi-temporal analysis. The model is applied to monitor annual variation of wetland inundation extents from a series of Landsat TM images in 2004 and HJ images in 2009 on the Poyang Lake national nature reserve (PLNNR) in China. We use related indices to represent spatial uncertainties of inundated areas and to delineate the transition zone between wetland vegetation and open water. The PLNNR water regime is investigated by accumulating a series of random sets during one year and determining the water covering days (WCD) at the pixel level. Random sets provide detailed spatial configurations of the WCD which has a strong negative correlation with the underwater DEM. Comparing 2004 and 2009, the study shows that almost half of the PLNNR area experienced drought. We conclude that the mixed Gaussian random set model with three components presented in this study serves as a general method to parameterize the random set model for large datasets. Moreover, it is well suited to capture detailed information on spatial temporal dynamic of wetland inundation and contributes to our understanding of wetlands water regimes from multi-temporal images.

5.1 Introduction

Wetlands are considered as a diverse ecosystem for providing wildlife habitats. They have important functions in minimizing the effects of flooding and erosion as well as in recharging of ground water. A wetland is usually characterized by a dominance of wetland plants, soils subjects to waterlog and indicators of wetland water regime [97]. Remote sensing images have been used to obtain information about these criteria for identifying wetlands, detecting their changes over time and monitoring the effects of human influences [e.g., 64, 77, 83, 102, 127]. Wetlands experiencing periodic flooding exhibit changes in the spatial distribution and temporal duration of inundation areas. A better understanding of wetland dynamics is important for the wetland ecosystem monitoring and habitat assessment [97].

Observing wetlands from satellite images is limited by spatial resolutions and revisiting intervals of sensors [122]. Moreover, conventional data processing such as hard classification and use of the crisp data model simplifies the extraction and representation of wetlands. Exploring more general models that can better represent spatially continuous and temporally dynamic phenomena and their uncertainties, therefore, becomes an attractive topic in spatial temporal modeling and change analysis [167]. Examples are found in modeling saline wetland in Spain [142], savanna-forest intergrade in Bolivia [41], coastal landscape in the Netherlands [177], glaciers in the Tibetan plateau [190] and tundra-taiga transition areas in Russia [130].

Mapping the extents of wetlands is a basic practice for wetland inventory and change tracking. It could suffer, however, from problems in identifying the uncertain boundary that appears as a set of mixed pixels in images [122]. Due to regular interactions between water and vegetation as well as gradual elevation gradients, the transition chain from dense vegetation through sparsely distributed vegetation, saturated soil, shallow and turbid standing water to open water is common and typical in a wetland that is periodically flooded. To extract inundated areas in such a natural landscape from images, we thus need to consider wide transition zones from open water to vegetation. The spectral signatures of water are condensed in the spectral feature space and therefore open water is found easily separated from vegetation [122]. Inundated wetlands, however, have less condensed spectral signatures, due to the transition zones from open water to vegetation which contain the mixture of shallow water, wet soil, flood debris, dead flattened algae and sparsely distributed young flush sedge. These transition zones cause a wide spread of spectral signatures of inundated wetland that may overlap with signatures of wetland vegetation. A crisp delineation of inundated extents leading to misclassification and artificially imposed

boundaries, therefore, is not appropriate for mapping inundated areas in wetlands.

A better representation of inundation extents has the potential to enrich our understanding of water regime. Water regime variation is the driver of wetland dynamics and a key factor for maintaining wetland ecosystem functioning and wetland vegetation development [17, 19]. Artificially imposed crisp boundaries of inundation areas lead to a stepwise and imprecise map of water covering duration from monthly collected images, like in [73]. Ideally, the dynamics of the physical process will be approached by increasing the number of observations between snapshots. Practically, this is prevented by fixed orbits and return intervals of satellites and lack of cloud free optical images. As an alternative, we propose to fill the gaps by characterizing the inundation status in a detailed way for every snapshot. It starts from our understanding that the status of wetland inundation at one moment is the result of a flooding or recession. Differentiating zones of different water status instead of water and non-water areas, should give more information on these processes.

Extraction and representation of objects with uncertain extents such as transition zones, is considered as an uncertainty modeling process [167]. Several conceptual models and spatial data models have been proposed to address it in a theoretical way [e.g., 9, 23, 32]. Other research realized the extraction of uncertain objects from images based on different uncertainty theories, such as probability theory, fuzzy set theory and random set theory [e.g., 41, 69, 159, 198].

This study focuses on the use of random sets. Random set theory deals with stochastic models of irregular or random geometrical structures [104]. Random sets have been successfully employed for developing image segmentation methods [37] as well as for the study of randomly varying geometrical shapes [169]. Zhao et al. [198] adopted a random set method to quantify extensional uncertainty of spatial objects and modeled the broad boundaries that were extracted from images. Their method was based on the assumption that objects e.g. by digitizing, thresholding or segmentation, with uncertain extents can be represented in an objective way by a probability based method, avoiding subjective choice of interpretations. In this way, the overall set of outputs was modeled as a random set whose variability reflects the degree of uncertainty of the object extent. Previous studies focused on modeling the extensional uncertainty of objects at a single time point and parameterized the random set model by user-defined parameters [198]. Tuning parameters by users or based on collecting training data on each image, however, involves subjective effects from users and are time consuming for large datasets. A semi-automatic routine, therefore, may facilitate applying the random set model in a spatial temporal analysis.

The challenge is thus to develop a general parameterization method to determine scene-specific parameters for multi-temporal images.

The aim of this research is to examine the application of random sets to monitor changes of wetland inundation with uncertain spatial extents. Specific objectives of the study are to: (i) refine the random set model that was developed in [198], and adapt its parameterization for monitoring seasonal dynamics of wetland inundation from series of images; and (ii) to investigate annual variation and interannual changes of wetland inundation extents. This study is applied to the Poyang Lake national natural reserve (PLNNR) in China.

5.2 Study area and Data

5.2.1 Study area

Poyang Lake is the largest freshwater lake in China, located in the middle reach of Yangtze River. It exchanges water with the Yangtze River through a 40 km long water channel in the northern part, and weakens flood peaks by storing discharges from its five tributaries. Interactions between the lake, rivers and surrounding land constitute an aquatic-terrestrial ecosystem within the Poyang Lake area. Poyang Lake national nature reserve (PLNNR) is located in the northwest of the Poyang Lake area (115° 55′-116° 03′E, 29° 05′-29° 15′N). Nine lakes are present in PLNNR, and we selected four of them (Dahuchi, Shahu, Banghu and Dachahu) that have larger areas and are ecologically meaningful to report in the result sections (Fig. 5.1).

During late spring and early summer (May and June), the five tributaries are flooded due to concentrated rainfall. As a consequence, the water level of Poyang Lake swells up without reaching the peak yet. From July until early September, the water supply from the five tributaries decreases, whereas the Yangtze River has its highest level, which causes the water level to rise to its peak, and the nine lakes in PLNNR merge into one single water body. From late September until November, the water level recedes towards its lowest level after the recession of the Yangtze River [72]. Lakes in PLNNR are disconnected by then and have different water levels throughout the winter (from December to February) (Fig. 5.2).

Three land cover types dominate the study area: water body, water-vegetation transition zone and wetland vegetation. Water body indicates the open water areas, including rivers, lakes and water channels. Vegetation indicates wetland plants on the land surface (Fig. 5.3(a)), including grass at high elevation such as *Miscanthus* spp. and sedge at low elevation such as *Carex* spp.. Water-vegetation transition zones are usually

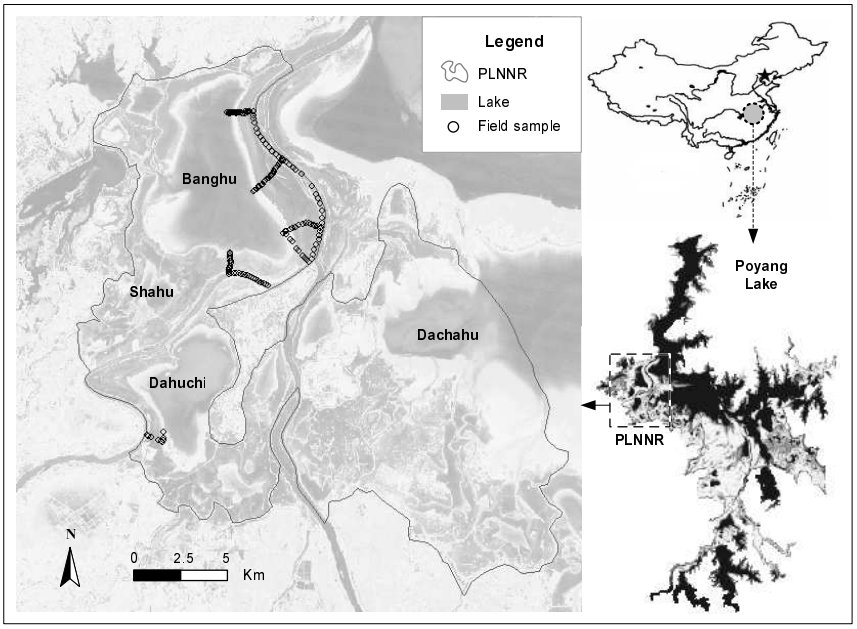

Figure 5.1 Location of Poyang Lake national nature reserve (PLNNR) and field samples collected in October and November, 2009. We selected four lakes in PLNNR: Dahuchi, Shahu, Banghu and Dachahu, for reporting in the result section.

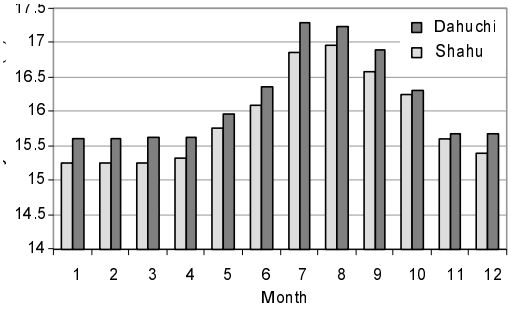

Figure 5.2 Monthly water levels of Dahuchi and Shahu in PLNNR in 2004.

Figure 5.3 Photos of water and vegetation in PLNNR, (a) wetland vegetation (b) river and narrow transition zone (c) lake and broad transition zone.

narrow on the river banks which are formed by natural levees due to sudden changes in elevation (Fig. 5.3(b)). The transition zones are wider on lake banks due to gradually changing elevations, and they usually consist of shallow standing water, saturated soil, dead algae and sparsely distributed young flush sedge (Fig. 5.3(c)). Such transition zones that are suited for different waterfowl species and serve as important habitats of migratory birds. In addition, they are an ideal environment for snails (*Oncomelaniahupensis*) growing and thus have high risk of schistosome (*Schistosoma japonicum*) infection. Identification of these zones, therefore, provides important information for mapping the food habitat of endangered birds like Siberian Crane (*Grus Ieucogeranus*) [186] and for predicting the distribution and abundance of schistosome (*Schistosoma japonicum*) [188]. In the rest of this paper, we name the three target classes as open water, transition zone and vegetation, respectively.

5.2.2 Images and pre-processing

HuanJing (HJ) images, acquired by HJ-1A/1B satellite, were obtained from the China Centre for Resource Satellite Data and Applications. The HJ-1A and HJ-1B satellites were launched on September 6, 2008, and passed the in-orbit-test review on February 19, 2009 after more than five months of in-orbit testing. The HJ-CCD sensor onboard HJ-1A/1B has four bands in the visible and near infrared spectrum: visible blue Ch-1 (0.43-0.52 μm), visible green Ch-2 (0.52-0.60 μm), visible red Ch-3 (0.63-0.69 μm) and near infrared Ch-4 (0.76-0.90 μm), with 30 m spatial

Table 5.1 Multi-temporal images and their references for geometric corrections.

TM		HJ		ETM+
No.	Date	No.	Date	Reference
		$I^{09(1)}$	14 Apr 2009	16 Apr 2000
$I^{04(1)}$	5 May 2004	$I^{09(2)}$	15 May 2009	16 Apr 2000
$I^{04(2)}$	22 Jun 2004	$I^{09(3)}$	19 Jun 2009	6 Jun 2001
$I^{04(3)}$	24 Jul 2004	$I^{09(4)}$	17 Jul 2009	8 Jul 2001
$I^{04(4)}$	9 Aug 2004	$I^{09(5)}$	19 Aug 2009	8 Jul 2001
$I^{04(5)}$	26 Sep 2004	$I^{09(6)}$	3 Sep 2004	10 Sep 2001
$I^{04(6)}$	12 Oct 2004	$I^{09(7)}$	6 Oct 2009	10 Sep 2001
$I^{04(7)}$	28 Oct 2004	$I^{09(8)}$	24 Oct 2009	24 Nov 1999
$I^{04(8)}$	29 Nov 2004	$I^{09(9)}$	24 Nov 2009	24 Nov 1999
$I^{04(9)}$	15 Dec 2004			29 Jan 2001

resolution and two days the revisiting time. Landsat TM images were obtained from the Chinese Remote Sensing Satellite Ground station. The first five bands of Landsat TM5 have a 30 m resolution: visible blue Ch-1 (0.45-0.52 μm), visible green Ch-2 (0.52-0.60 μm), visible red Ch-3 (0.63-0.69 μm), near infrared Ch-4 (0.76-0.90 μm) and mid-infrared Ch-5 (1.55-1.75 μm). One series of HJ images in 2009 and one series of Landsat TM images in 2004 were chosen based on data availability and image quality (Table 6.1). We labeled an image I in year T with a time order t as $I^{T(t)}$.

Acquired HJ and Landsat TM images are products with systematic geometric corrections. We used six Landsat ETM+ images with precise geometric correction (L1T product) as reference images (Table 6.1) to implement image-to-image geometric corrections, and the root mean square errors were below 15 m. Atmospheric correction was carried out according to the cosine approximation model described by [18]. After the atmospheric correction, the digital numbers (DN) were transformed to surface reflectance. Corrected images were then spatially subset to the PLNNR region.

Near infrared band and mid-infrared band are commonly used to discriminate water and land, due to the spectral absorption features of water in infrared portions of the electromagnetic spectrum [e.g. 15, 139]. Several multispectral indices have been developed based on the combination of visible bands and infrared bands for this purpose, such as the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), the modified NDWI (MNDWI) [52, 96, 105, 135, 187]. Logical combinations of these indices have been proposed,

showing improved discrimination capability in specific cases [93, 121]. Due to lack of mid-infrared band in HJ images and the large proportion of vegetation present in the study area, we chose the commonly used NDVI index to map water and its transition zones to vegetation. These NDVI maps are the input to further processing.

A digital elevation model (DEM) from 1998, covering major parts of PLNNR was used for validation. Its vertical accuracy is 0.1 m and spatial resolution is 20 m. We reprojected the DEM to the same projection UTM-WGS84 zone 50N with images and resampled it to 30 m resolution. In addition, the daily water levels of Dahuchi and Shahu are available for 2004. They are combined with the DEM to assist evaluation of the model results.

5.2.3 Field samples

A ground survey has been carried out in PLNNR between October 26th and November 6th, 2009. Its aim was to investigate the spatial configuration of wetland vegetation and to validate classifications. Five sampling lines have been designed, starting from the river bank and perpendicularly crossing different zonations of vegetation until they reach the lake bank. 95 plots were located along the five sampling lines and 57 plots were located in water along the cours of the vessel that took us between start points of the sampling lines (Fig. 5.1).

Ten to twenty seven plots were located on each sample line with size of 30×30 m, and averaged distance between the plots is 90 m. The location of each sample plot was measured by a GPS at the center of the plot, with a accuracy less than 10 m. Within the plots, the following variables were recorded: vegetation types, coverage and height, soil color and coverage, and presence of standing water. According to the density of wetland vegetation (less than 5%), wetness of the soil and distance from the open water, 20 out of 95 samples were categorized as reference for the transition zone and 75 samples as references for vegetation. For the remaining 57 samples collected on the river, only GPS locations were recorded and they were categorized as references for open water. To validate the thematic accuracy of $\Gamma^{09(8)}$, we compared all 152 samples with the pixels classified as open water, transition zone and vegetation, respectively.

5.3 Methods

5.3.1 Parameterization of random set model

Let the intensity of an image I within a window W be denoted as $f : W \to [0,1]$. Each pixel $x \in W$ has an intensity value $f(x) \in [0,1]$. The

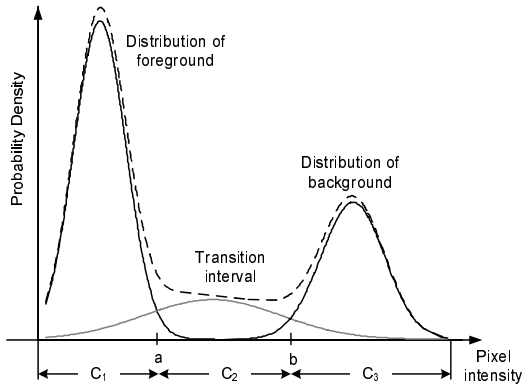

Figure 5.4 An image density function and related mixed Gaussian model.

intensity function f can be interpreted as a collection of sets $F = \{x \in W, R \in [0,1] : f(x) \geq R\}$. F becomes a random set when R is a random variable. The distribution of a random set is uniquely determined by f and the random variable R [48].

An image can be partitioned into a foreground and a background class, given their clear separation in the histogram and absence of uncertain pixels in between [123]. But since this idea case rarely exists, many efforts have been made to find the optimal threshold between the foreground and background [e.g. 15, 172]. In this study, we treat the uncertain areas between the foreground and background as the third class, the transition zone, with intensity values in the interval $[a, b]$, where $0 \leq a \leq b \leq 1$ (Fig. 5.4). We consider wetland inundation as the area of interest (AOI) consisting of open water and the transition zone, and the vegetation as the background. We aim to extract the extent of AOI in an nondeterministic way and model it as a random set to quantify its extensional uncertainty.

The essential part of creating the random set model is to generate realizations that fully characterize its distribution. To obtain these realizations, we first determine the probability distribution of R and generate random numbers in the transition interval $[a, b]$ as multiple thresholds. The Gaussian distribution was chosen as in [198], based on the assumption that pixel values close to object boundary have a higher probability to be selected by the image interpreters when labeling boundary pixels, than pixel values far away. As a more general parameterization method, we use the mixed Gaussian model to fit as a density distribution to the image histogram and to determine the transition interval $[a, b]$ (Fig. 5.4). In a spatial-temporal monitoring,

series of images with different histograms have different proportions of transition zones. For example, an image captured at the summer flooding peak has a narrow transition interval than an image captured at flood recession time. We therefore selected the mixed Gaussian model with three components (i.e. the distributions of open water, vegetation and transition zones). Transition intervals with various widths are determined by tuning the weight of the transition zone component.

Let the three classes be denoted as: open water (C_1), transition zone (C_2) and vegetation (C_3). We assume that the intensity of pixels belonging to class $C_i, i \in \{1,2,3\}$ follows the Gaussian distribution. We denote $C_i \sim \mathrm{N}(\mu_i, \sigma_i)$ to signify that C_i has the Gaussian distribution with mean μ_i, standard deviation σ_i and density function $(\sqrt{2\pi}\sigma_i)^{-1}\exp[-\frac{1}{2}(z-\mu_i)^2/\sigma_i^2]$ in the one-dimensional case. The density function of I is the mixed density distribution of C_i:

$$d(z) = d(z;\theta_i,\mu_i,\sigma_i) = \sum_{i=1}^{3} \theta_i \mathrm{N}(\mu_i,\sigma_i)(z) \tag{5.1}$$

where $z = f(x)$, θ_i is the weight coefficient for C_i and $\sum_{i=1}^{3}\theta_i = 1$. Without loss of generality, we assume that $\mu_1 < \mu_2 < \mu_3$. The lower limit of the transition interval is to be determined at a where $\theta_1\mathrm{N}(\mu_1,\sigma_1)(a) = \theta_2\mathrm{N}(\mu_2,\sigma_2)(a)$, and the upper limit at b where $\theta_1\mathrm{N}(\mu_2,\sigma_2)(b) = \theta_3\mathrm{N}(\mu_3,\sigma_3)(b)$. This way we identify three classes, as shown in Fig. 5.4:

Open water (C_1): $f(x) < a$

Transition zone (C_2): $a \le f(x) \le b$

Vegetation (C_3): $f(x) > b$

5.3.2 Modeling extents of wetland inundation by random sets

To model the extents of wetland inundation, we generated 200 random numbers from the distribution of $R \sim \mathrm{N}(\mu_2,\sigma_2)$ in $[a,b]$. This results into different realizations $O_1,\ldots,O_n$ of a random set by thresholding the NDVI image using $R_1,\ldots,R_n$ as the multiple thresholds: $O_i = \{x \in W, i \in \{1,\ldots,n\}, R_i \in [a,b] : f(x) \le R_i\}$. The covering function, also named the one-point covering function in Stoyan and Stoyan [169], serves to characterize the distribution of the random set Γ. The covering function $Pr_\Gamma(x)$ at pixel x is defined as $P(\Gamma \cap \{x\} \neq 0) = P(x \in \Gamma)$. Pixels with $Pr_\Gamma(x) \ge p$ constitute a p-level set of Γ. The 1-level set and the 0.5-level set are called the core set (Γ_c) and median set, respectively. The support set (Γ_s) is defined as $\{x \in W : Pr_\Gamma(x) > 0\}$. Pixels with $Pr_\Gamma(x) = 1$ are labeled as open water, with $Pr_\Gamma(x) \in (0,1)$ as transition zone and with $Pr_\Gamma(x) = 0$ as vegetation.

The covering function at x is estimated as:

$$\hat{Pr}_{\Gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{O_i}(x), \tag{5.2}$$

where the O_i are realizations of Γ and $I_{O_i}(x)$ is the indicator function of $O_i(x)$. The corresponding set-theoretic variance is estimated as:

$$\hat{\Gamma}_{var}(x) = \frac{1}{n}\sum_{i=1}^{n} (I_{O_i}(x) - \hat{Pr}_{\Gamma}(x))^2 \tag{5.3}$$

and pixels with high variance are expected to occur close to the boundaries between the classes.

Two indices SD and CV, derived from $\Gamma_{var}(x)$, summarize the dispersion of the distribution of a random set [198]. The SD is the sum of Γ_{var} over Γ_s, defined as:

$$SD = \int_{\Gamma_s} \Gamma_{var}(x)dx. \tag{5.4}$$

A high SD corresponds to a broad transition zone and a large extensional uncertainty, i.e. the transition from open water bodies has a gradual transition to the surrounding vegetation. The CV is the coefficient of variation, defined as:

$$CV = \int_{\Gamma_s} \sqrt{\Gamma_{var}(x)}dx \;/\; \int_{\Gamma_s} Pr_{\Gamma}(x)dx. \tag{5.5}$$

CV as a normalized and dimensionless measure allows us to compare extensional uncertainties of AOI with different sizes. A high CV indicates a larger proportion of the AOI with a high Γ_{var} and thus points to a large extensional uncertainty. These two indices are used to quantify the extensional uncertainty of wetland inundation.

5.3.3 Identifying crisp extents of wetland inundation

For a comparison purpose, we classified the image into two crisp classes: water and non-water, and contrasted the results with those from random sets. We adopt the mean set of the random sets that is a binary set to identify the crisp extents of water. Several definitions of the expectation of a random set exist in the literature. The Vorob'ev expectation is determined by the covering function that is the average of the indicator functions [169]. Replacing the indicator functions by distance functions, then the expectation of a random set is characterized by the distance average [6]. Here, we use the mean set based on the oriented distance

function (ODF), because the ODF takes the distance of each pixel to the random set into account [76]. All pixels in the crisp classification are classified both according to their grey value as well as to their positions with respect to the AOI.

For the observed realizations $O_1, \ldots, O_n$, we denote the ODF of Γ at pixel x as $b_{O_i}(x)$:

$$b_{O_i}(x) = d_{O_i}(x) - d_{O_i^c}(x), x \in W \tag{5.6}$$

where $d_{O_i}(x)$ and $d_{O_i^c}(x)$ are the distances from x to O_i and the complement of O_i, respectively. We use the Euclidean distance that $d_{O_i}(x)$ equals the distance from x to the nearest pixel in O_i. Let $\hat{b}_O(x) = \frac{1}{n}\sum_{i=1}^{n} b_{O_i}(x)$ denote the sample mean ODF at pixel x. Pixels with positive $\hat{b}_O(x)$ are classified as non-water areas, whereas those with non-positive $\hat{b}_O(x)$ are classified as water areas. This definition allows us to determine the water areas with a crisp boundary by the mean set $\hat{\Gamma}_{ODF} = \{x : \hat{b}_O(x) \leq 0\}$.

5.3.4 Analyzing changes in wetland inundation

For each $I^{T(t)}$, we determine the transition interval $[a^{T(t)}, b^{T(t)}]$ and generate random thresholds to slice the image into different binary realizations. The covering function $\hat{Pr}_\Gamma(x,t)$, set-variance $\hat{\Gamma}_{var}(x,t)$, mean set $\hat{\Gamma}_{ODF}^{(t)}$ and indicators $SD^{T(t)}$ and $CV^{T(t)}$ are estimated for each $\Gamma^{T(t)}$. Areas of $\Gamma^{T(1)}, \ldots, \Gamma^{T(t)}$, $SD^{T(1)}, \ldots, SD^{T(t)}$ and $CV^{T(1)}, \ldots, CV^{T(t)}$ report the seasonal dynamics of the inundated area and their extensional uncertainties. A comparison may reveal similarities and differences between 2004 and 2009.

To analyze the annual variation of inundation, multi-temporal maps of wetland inundation are used to produce water covering day (WCD) maps. The number of days that a pixel x is classified as water provides the water covering days at x within a year. Since our calculation depends on images at discrete time points, we assume that the extent of inundation on $I^{T(t)}$ remains the same from $I^{T(t-1/2)}$ until $I^{T(t+1/2)}$. We thus define the water covering days at pixel x in the year T as:

$$WCD^T(x) = \sum_{j=1}^{t} p(x,j) \times \left(\frac{|d_{j+1} - d_j|}{2} + \frac{|d_j - d_{j-1}|}{2}\right) \tag{5.7}$$

where $p(x,j)$ is the probability that pixel x is covered with water at time j and d_j is the Julian date of j, $d_j \in \{1, \ldots, 365\}$. To avoid a long time span of no-data before d_1 and after d_t, we use the average of $p(x,1)$ and $p(x,t)$ as the probability for the no-data period. Thus, in Equation 5.7, we set $d_0 = d_t - 365$ and $d_{t+1} = d_1 + 365$. If we model the inundation

areas as crisp sets by $\Gamma_{ODF}^{(t)}$, then $p(x, j) = 1$ if $x \in \Gamma_{ODF}^{(j)}$ and $p(x, j) = 0$ otherwise. By modeling the extensional uncertainty of inundation areas using random sets $\Gamma^{T(t)}$, the $p(x, j) = Pr_{\Gamma}(x, j)$ have values between 0 and 1. Therefore random sets provide different WCDs as compared to crisp sets. Since *in situ* data about water covering days are not available, we consider an underwater DEM as the alternative to validate the WCD. We expect that WCDs with detailed spatial configurations based on random sets (WCD-RS) will better match underwater elevations and thus have a stronger correlation with the DEM than WCDs based on ODF (WCD-ODF). For two lakes i.e. Dahuchi and Shahu, of which the daily water level was recorded in 2004, the WCDs (WCD-WL) are calculated by combining water levels with the DEM. WCD-RS supposedly has a stronger positive relationship with WCD-WL than WCD-ODF. The Pearson correlation coefficient and related statistic test are used for this purpose.

To obtain interannual change information of WCD, we generate WCDs for both 2004 and 2009 and produce the difference map ($\Delta WCD = WCD^{2009} - WCD^{2004}$) to locate the changes of wetland inundation between these two years. Theoretically, pixels with $\Delta WCD > 0$ show areas turning wet, whereas areas with $\Delta WCD < 0$ show places turning dry, and areas with $\Delta WCD = 0$ show no changes. Considering that lag of image acquisition dates can be as large as a month in the two multi-temporal series, we set a tolerance and categorize $\Delta WCD \geq 30$ as wet turning areas, $\Delta WCD < -30$ as try turning areas and $-30 \leq \Delta WCD < 30$ as areas with no change.

5.4 Results

5.4.1 Extents of wetland inundation represented by random sets

The extent of wetland inundation is represented by the random set model for image $I^{09(8)}$ in Fig. 5.5. This is taken as an example to show the support set Γ_s, the core set Γ_c, the covering function $Pr_{\Gamma}(x)$ and the variance $\Gamma_{var}(x)$ of a random set. Pixels in Γ_c (shown in blue) surely belong to open water, whereas those outside of Γ_s (shown in yellow) surely belong to vegetation. The gradual changes in transition zone are represented by $Pr_{\Gamma}(x)$ with different values (shown in green) between Γ_c and Γ_s. Pixels in bluish green have a high probability of inundation and those in yellowish green have a low probability. A clear distinction exists between the narrow transition zones separating rivers and vegetation and broad transition zones between lakes and vegetation. The lake bank of Dachahu has a wide and vague transition zone, followed by Dahuchi, whereas Shahu has a relatively narrow transition zone to surrounding

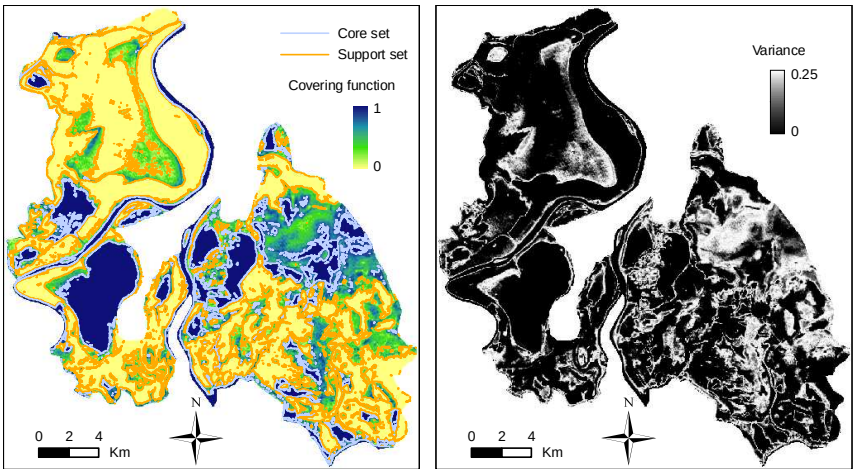

Figure 5.5 The random set model for image $I^{09(8)}$ is shown by its covering function, contours of the support set and the core set (left), and the set-theoretic variance (right).

Table 5.2 Result of accuracy assessment for the map of open water, transition zone and vegetation on $I^{09(8)}$.

	No. of pixels	Producer's Accuracy (%)	User's Accuracy (%)
Open water	57	100	98
Transition zone	20	100	80
Vegetation	75	92	100
Overall Accuracy: 96.1%			
Kappa Statistic: 0.93			

wetland vegetation. Since Banghu experienced severe drought and the lake was dry during our field work, it shows a large area of vegetation and wide transition zones, but no core area for open water in Fig. 5.5.

The set-theoretic variance $\Gamma_{var}(x)$ shows in grey scale values with light tones indicating high variations in uncertain transition zones, and dark tones indicating low variation in vegetation and water classes. For pixels inside Γ_c or outside Γ_s, $\Gamma_{var}(x)$ equals 0, whereas pixels far from the contours of Γ_c or Γ_s have higher $\Gamma_{var}(x)$ values. The extent of Dahuchi has the large uncertainties in its western and southwestern parts, mainly caused by location of its east bank next to a populated town at a higher elevation. In contrast, its unconsolidated west shore

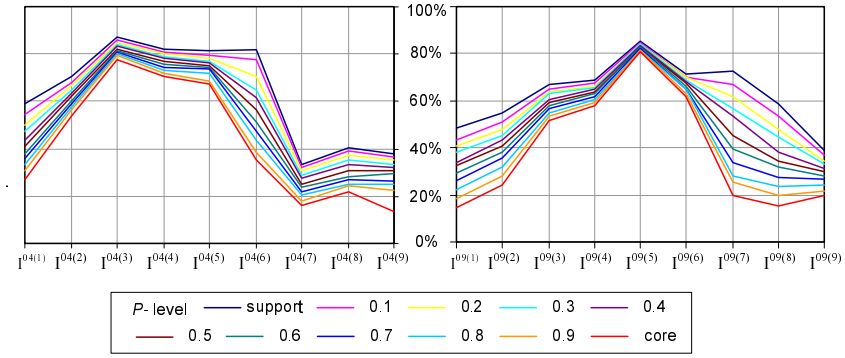

Figure 5.6 The proportion of wetland inundation areas in the PLNNR at different p-level sets.

connected to a natural wetland grassland has a wide, gradual transition zone.

The $\Gamma^{09(8)}$ map has an overall accuracy of 96.1% and a kappa statistic of 0.93 (Table 5.2). Misclassifications occur at six plots that were sampled as vegetation near narrow transition zones close to rivers. These six plots correspond to mixed pixels in the image, caused by contributions from both vegetation and open water to the spectral reflectance. As a consequence, five were misclassified as transition zone and one as open water.

5.4.2 Seasonal dynamics of wetland inundation extents

The seasonal dynamics of inundated extents are demonstrated by the temporal profiles of areas of p-level sets (Fig. 5.6) and four maps selected from different seasons (Fig. 5.7). The area of wetland inundation has the largest proportion between 78% and 87% on $I^{04(3)}$ and between 81% and 85% on $I^{09(5)}$, occurring during flooding peaks. At the mean time, areas of different p-level sets have the smallest range, which implies a small proportion of transition zone between flood and vegetation, i.e. 9% and 4 % respectively. Wetland inundation has the most varying extents between 35% and 82% on $I^{04(6)}$ and between 20% and 73% on $I^{09(7)}$ in early October when floods start to recede. The transition zones occupy approximately half the total study area. This is mainly because large areas just experience flood recession and have shallow standing water and large extents of saturated soils remain on the land surface. In months such as April and November when the water level is stable, the proportion of transition zones is still relatively large, around 20%. These

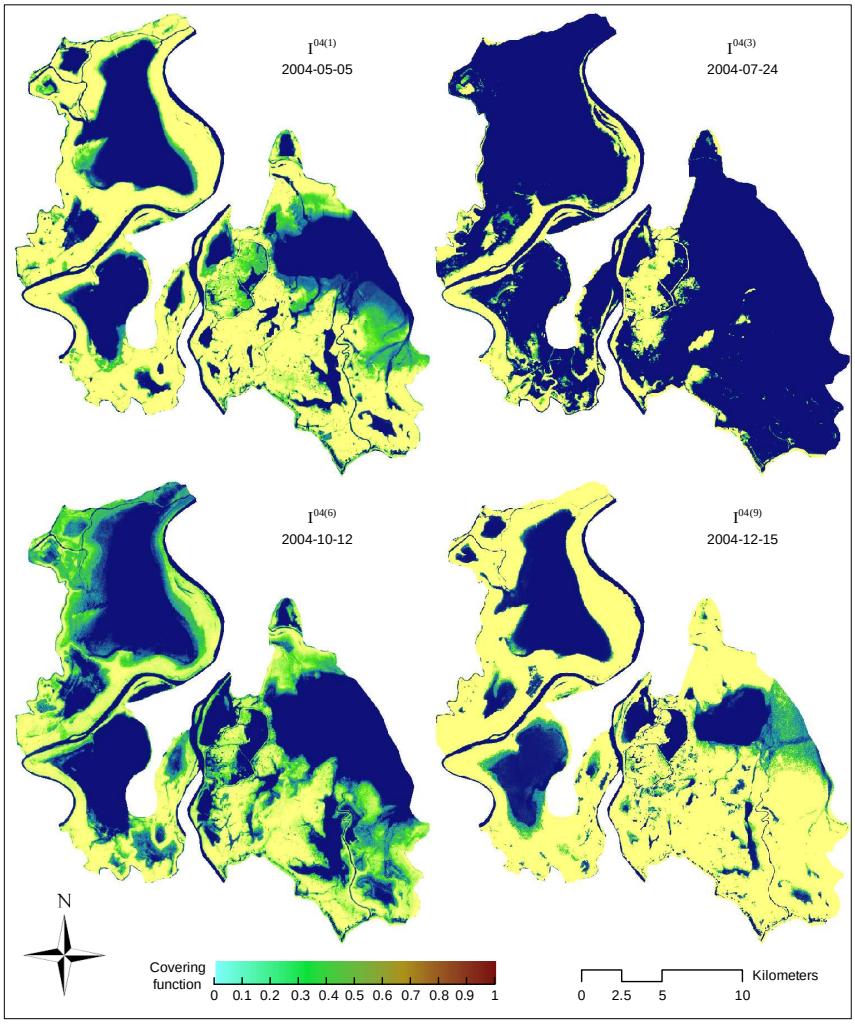

Figure 5.7 Four maps of random sets for illustrating the spatial pattern of wetland extents in different seasons.

broad transition zones could hinder a crisp partition between water and non-water and subsequent area calculations of inundation.

The proportion of water classes derived from the ODF based hard classification is compared with the proportion of open water class and transition zone class derived from random set model in Table 5.3. The proportion of water (ODF) from each image is larger than the proportion of open water, but smaller than the sum of open water and transition

Table 5.3 Proportion of inundated wetland in PLNNR, classified as water in ODF and the combination of open water and transition zone in random sets.

Image	Water (ODF)	Open water	Transition zone
$I^{04(1)}$	41	28	31
$I^{04(2)}$	62	54	17
$I^{04(3)}$	83	78	9
$I^{04(4)}$	78	71	11
$I^{04(5)}$	75	67	15
$I^{04(6)}$	60	35	47
$I^{04(7)}$	25	17	17
$I^{04(8)}$	32	22	19
$I^{04(9)}$	29	14	24
$I^{09(1)}$	29	15	24
$I^{09(2)}$	39	25	30
$I^{09(3)}$	60	52	15
$I^{09(4)}$	64	58	11
$I^{09(5)}$	83	81	4
$I^{09(6)}$	68	62	9
$I^{09(7)}$	42	20	53
$I^{09(8)}$	34	16	43
$I^{09(9)}$	29	20	19

zone constituting the inundated area. Images with the same proportion of water derived using a crisp method, may have different proportions of open water and transition zones than those derived from random sets. For example, when ignoring transition zones, water takes the same proportion (29%) on images $I^{04(9)}$, $I^{09(1)}$ and $I^{09(9)}$. After taking transition zones into account, $I^{04(9)}$ has a 14% open water area, which is slightly smaller than 15% on $I^{09(1)}$ but much smaller than 20% on $I^{09(9)}$. $I^{09(9)}$, however, has a smaller proportion of transition zone than $I^{04(9)}$ (19%) and $I^{09(1)}$ (24%). Consequently, the total inundation areas are similar on these three images, but their components: open water and transition zone take different proportions. The extraction of transition zones, therefore, offers extra information for locating the inundation extents and calculate the areas of inundation.

Temporal profiles of two indices *SD* and *CV* in Fig. 5.8 show dynamics of extensional uncertainty of wetland inundation. The size of the uncertain areas indicated by *SD* decreases from April to July or August,

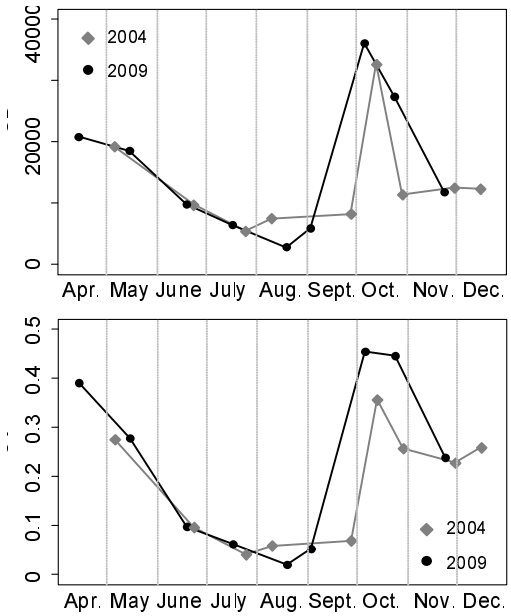

Figure 5.8 Trends of the extensional uncertainty of wetland inundation in 2004 and 2009, represented by temporal profiles of SD and CV.

it jumps to its highest value in early October, and decreases afterwards. The temporal profile of CV in Fig. 5.8 has similar seasonal dynamics as that of SD. One exception is the more gentle decrease from October to November, since the total inundated area also shrinks as the area of transition zone decreases. In addition, the uncertain areas in November and December are slightly larger than those in June and September according to the SD; their differences are more prominent as expressed by the CV because the CV takes small extents of inundation in winter into account.

5.4.3 Annual variation of wetland inundation extents and interannual changes

The annual variation of wetland inundation indicated by the water covering days (WCD) is shown in Fig. 5.9. Centers of lakes are permanent water covering areas in 2004. The lake shores have gradual changes in WCD, showing the dynamics of lake extents within a year. Places that are located near the lake bank and submerged during the whole flooding season (shown in yellow) provide ideal wetland soils for sedges

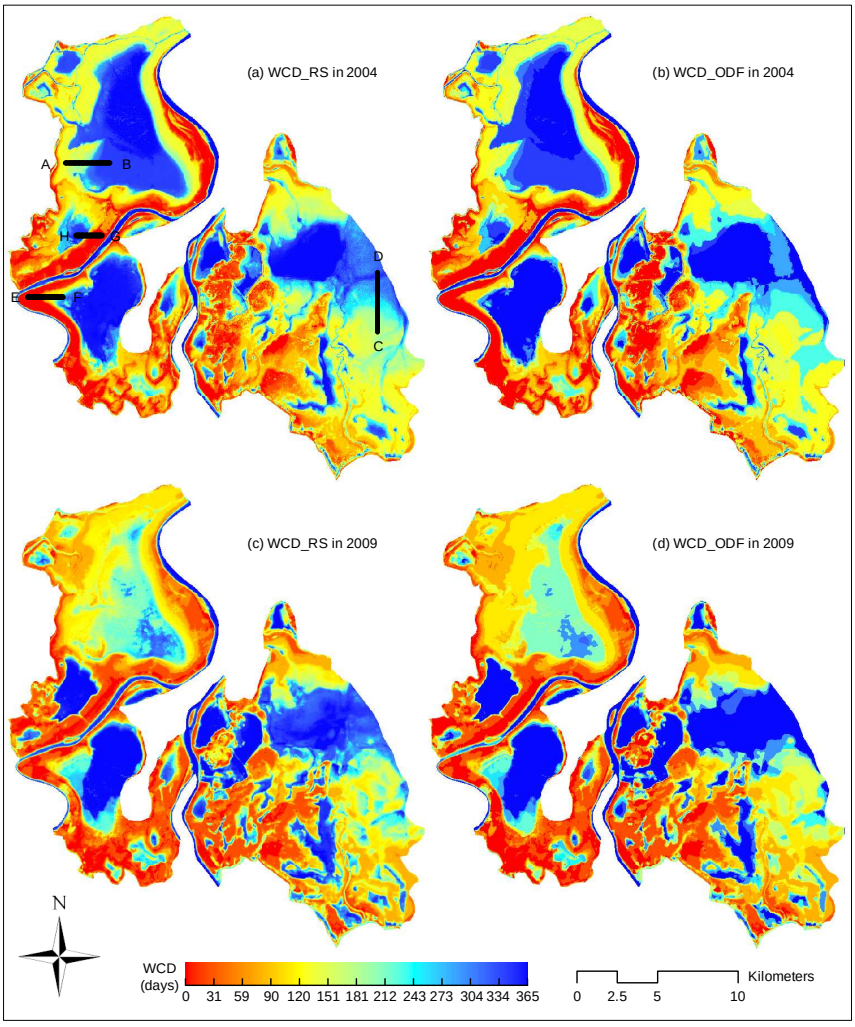

Figure 5.9 Four transects and the water covering days (WCD) maps in year 2004 and 2009, derived from random sets and ODF mean sets.

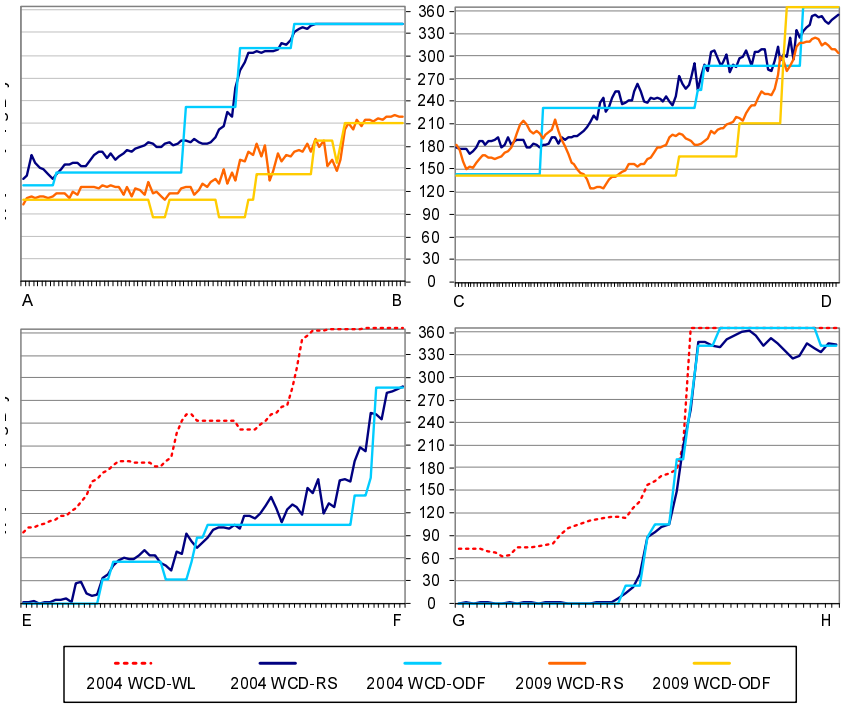

Figure 5.10 Water covering days extracted on four transects.

Table 5.4 Correlations between DEM and WCD-RS, WCD-ODF at four transects represented by correlation coefficients (r, $p < 0.001$).

	2004				2009			
	AB	CD	EF	GH	AB	CD	EF	GH
WCD-RS	-0.88	-0.96	-0.84	-0.95	-0.81	-0.76	-0.78	-0.93
WCD-ODF	-0.89	-0.92	-0.77	-0.95	-0.62	-0.76	-0.71	-0.93

like *Carex*. Uplands with lowest WCDs (shown in red) are dominated by grass that is emerged during the short flooding time. Comparing the two columns of Fig. 5.9, we observe stepwise changes in crisp WCDs, whereas WCDs given by random sets are smoother and reflect more details in water covering change. This difference especially occurs on natural lake banks. We now use transect AB on the bank of Banghu, transect CD on the bank of Dachahu, transect EF on the bank of Dahuchi and transect GH on the bank of Shahu (Fig. 5.9(a)) to make a further quantitative comparison.

Fig. 5.10 shows the water covering days extracted from all 92 pixels on AB, 121 pixels on CD, 73 pixels on EF and 53 pixels on GH. Corresponding pairs of WCD-RS and WCD-ODF curves have similar trends. Compared with WCD-RS, the WCD-ODF curve consists of pieces of line segments with no fluctuations within each segment, but apparent discrepancies between two neighboring segments. These steep and sharp turns are replaced by gentle slopes in WCD-RS, which are more consistent with the gradual gradient of underwater DEM and follow Tobler's First Law [106]. The differences between the corresponding WCD-RS and WCD-ODF curves are relatively large near the pixels where a large change in values occurs either downward or upward. For example, the largest differences occur at the ascending step from $WCD = 144$ to $WCD = 231$ between 2004WCD-RS and 2004WCD-ODF in both AB and CD. For transects EF and GH, the WCD-WL curves are plotted to compare with WCD-RS and WCD-ODF in 2004. Although with a similar trend, they do not match as well as expected, probably due to the fact that the DEM stems from 1998. The WCD-WL values are significantly correlated with both WCD-RS and WCD-ODF, but the correlations are lower with WCD-ODF ($r = 0.80$ for EF and 0.98 for GH) than those with WCD-RS ($r = 0.90$ for EF and 0.99 for GH).

Table 5.4 gives the correlation test results between the DEM and WCD-RS, WCD-ODF. Relationships between WCDs and the DEM extracted from these pixels are negative and all are significant at the $p = 0.001$ level. All but one correlation coefficients in the first row of Table 5.4 are almost higher in an absolute sense than those in the second row, showing that WCD-RS has a stronger correlation with the DEM than WCD-ODF. In addition, negative correlations in 2004 are stronger than those in 2009. This may be caused by the change in underwater terrain within the five years.

Fig. 5.11 shows the ΔWCD map and the corresponding histogram. The outstanding bin in the histogram shows a large number of non-change areas with $\Delta WCD = 0$. There are more pixels on its left side than on its right side, resulting in a left skewed histogram. This indicates that more areas in the PLNNR turn dry rather than turn wet between 2004 and 2009. Pixels with ΔWCD less than one month, regarded as a non-change area, occupy 39.8% of the total study area. Area proportions of moderate drought or inundation within three months occupy 28.2% and 10.9%, respectively. Other more severe changes in drought and inundation of more than three months occupy 13.2% and 7.9% of the study area, respectively. In total, 41.4% of the study areas have different degrees of decreasing water covering frequency, whereas 18.8% of the total area has an increasing inundation duration.

Large areas of Banghu experienced changes in wetland water regimes between 2004 and 2009 according to Fig. 5.9 and Fig. 5.11. The perman-

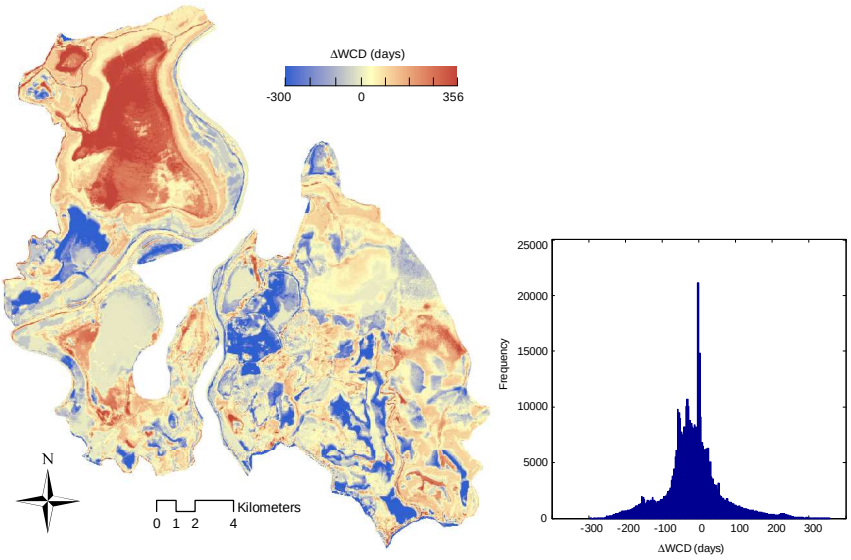

Figure 5.11 The difference water covering days (ΔWCD) map and corresponding histogram.

ent water covering areas shrink to the small corner in the southeast of Banghu, whereas the other places have less water covering days in 2009. These changes are also observable from transect AB in Fig. 5.10. Dahuchi experienced less modification and a decreasing WCD mainly on the west bank and south bank. In contrast, Shahu has larger areas covered by water during the whole year 2009 and its lake bank becomes more crisp. The wetland in Dachahu appears more fragmented in 2009. Some parts have less WCD and some have more, which makes the previous continuous shore break into patches.

5.5 Discussion

This study addresses the parameterization of random sets, by doing so it makes the model applicable and efficient for analyzing large datasets. The mixed Gaussian model has been used in the past to determine object boundaries by Shi et al. [159]. They fitted an image histogram by two Gaussian models for the distribution of object and its background, and identified the object boundary as the cross area outside the two standard deviations of both distributions. We found that this method is not applicable if the aim is to locate a large interval between two modes. For example, fitting the histogram in Fig. 5.12 with two Gaussian

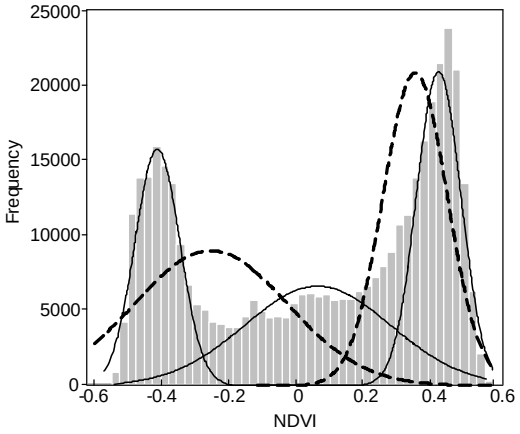

Figure 5.12 Histogram of NDVI image $I^{04(1)}$ and fitted two Gaussian components (dashed curves) and three Gaussian components (solid curves).

distributions leads to the unsatisfactory result that the distribution of the foreground does not match the left mode. Instead, if the distribution of transition zones is considered as the third component in the mixed Gaussian model, this model is more flexible to detect either wide or narrow transition intervals for various kinds of image histograms. This is in particular relevant for spatial temporal studies. In addition, this unsupervised parameterization method eliminates the sampling subjectivity and the time consuming training phases.

This study reveals that the parameterization of random sets successfully detects the distributions of vegetation, open water and transition zone. But it may not be appropriate for detecting specific species of wetland vegetation given their similarity in spectral signatures. Since different wetland vegetation communities have different phenology, we are able to classify and model them by choosing images from the right season [122]. Zhao et al. [198] selected one Landsat TM image in autumn to model the *Carex* patches with gradual transition to *Miscanthus* by random sets. These two vegetation communities are not separable in spring but are separable in autumn due to the second growing phrase of *Carex*.

We partially validated the WCDs with underwater elevation according to a common sense that a place at lower elevation should be covered by water for a longer time than a place at a higher elevation. The correlations were tested for four transects, being located in different lakes with local terrains. The results reveal that random set based WCDs

outperform crisp set based WCDs because of the smoothing effects and that random sets can complement the gaps among limited snapshots for the WCD mapping. The smoothing effects could be explained by utilizing a detailed characterization of the current status to reflect flooding and recession processes in the past. For example, a place classified as open water should have a larger probability to be covered by water during the whole observation interval (days from current image to previous image) than a place classified as a transition zone. Those classified as transition zones should be covered by water for a longer time than those classified as vegetation.

Random sets identify uncertain extents of inundated wetlands in a nondeterministic manner and the results show that its derived dynamic modeling resolves the deficiencies of exploring water regime pattern by deterministic methods. This random set model is worthy of further investigations into other wetlands or natural landscapes with uncertain extents. The spectral information about inundated areas is derived from a multispectral index NDVI, in order to apply the random set model to a single input band. Other indices, such as NDWI, MNDWI and NDVI-NDWI, can be used to detect the inundated areas, like suggested in [29, 73, 75, 93]. To implement the random set model on multiple bands, e.g. namely NDVI, NDWI and mid-infrared, we propose to first run the random set model on each of these separately and then combine the resulting random sets by logical operators. Another option is to first combine all input bands by logical operators and then model the collection of results by random sets. If more than two types of land cover dominate the study area, irrelevant land cover could be masked before applying the method to the AOI and its background.

The classified open water, transition zone and vegetation have been validated by samples collected at one field survey. Since the survey was not designed for validating the water extents, it did not allow us to further investigate the relationship between covering function and water coverage or depth. If such an investigation is considered in the future, collecting samples from several other reference data will be required besides field surveys. Many constraints prevent the accessibility: (1) the centers of lakes in PLNNR are only approachable by small boats during the flooding season when water levels are high. But they become difficult to access during autumn and winter because lakes are at low water levels and their flat muddy bottom terrain easily stick the boats; (2) Shahu and Dahuchi are staging and overwintering sites for migratory birds, so human disturbances are to be avoided in late autumn and winter. In those situations, we may turn to find other references such as satellite image with high spatial resolution, air photography and radar data as alternatives.

The wetland inundation in PLNNR experienced great changes between

2004 and 2009. Drought with different severities occurred in nearly half of the study area, whereas only a small part of the areas was covered by water for a longer time. The most severe and the largest areas of drought happened in Banghu, although its lake bottom is deeper than Shahu and Dahuchi and thus it should have a longer water covering time. One reason is that Shahu and Dahuchi are the only two lakes to which the Chinese government has its ownership, where levees have been constructed to control the water levels for protecting the wetland landscapes and habitats of migratory birds. The other seven lakes, however, are only partially protected and do not have a water level control mechanism. Their changes such as drought in Banghu provide a hint on the changes of Poyang Lake. All products in this study provide detailed information on the water regimes for PLNNR managers, and can be used for assessing the previous preserve effect, assisting identification of future protected areas and triggering further study on the Poyang Lake hydrological dynamics.

5.6 Conclusions

This study presents a mixed Gaussian model with three components that serves as a general method for determining the transition intervals from images with various density distributions. The covering function of the random set and related set-theoretical variance enable a more detailed characterization of wetland inundation extent than a hard classification. Temporal profiles of p-level sets and two indices SD and CV, show the seasonal dynamics of inundated areas and related extensional uncertainty. Open water and transition zones can be identified be means of the core set and support set. Transition zones have the largest area and occupy almost half of the PLNNR immediately after flood recession in early October. In other months when the water level is stable, the proportions of transition zone are still approximately twenty percent and can hardly be ignored in a classification.

Besides uncertainty modeling, random sets are used to create artificially imposed water boundaries through the ODF mean set for comparison purpose. Random set based WCDs and ODF based WCDs are generated to represent annual variation of wetland inundation areas. Comparing them with the DEM shows that detailed changes become detectable through random set modeling and that derived water covering days have significant and higher correlation with underwater elevations. We therefore reach the conclusion that uncertainty modeling of inundation extents by random sets contributes to the WCD mapping, through smoothing the WCD variation captured by limited snapshots, and through providing more details than using a crisp method.

6 Use of a random spread process to model spatial-temporal pattern of wetland inundation

Submitted as: Xi Zhao, Alfred Stein, Xiaoling Chen, Xiang Zhang and Lian Feng. Use of a random spread process to model wetland inundation from multitemporal images. Remote Sensing of Environment.

Abstract

Storage and representation of spatial-temporal information with uncertainties in image mining is an important step for spatial-temporal modeling of land changes. In this study, we modeled wetland inundation as a random spread process and used random sets to characterize stochastic properties of inundation extents extracted from multitemporal images. Periodicity, trend and random components were captured by monthly and yearly random sets. The Cov-Dist matrix and related operators summarized and visualized the spatial pattern and quantified the similarity of different stages in the process. The method was implemented on the Poyang Lake wetland area in China, and MODIS images for a period of eleven years were used. Results improved our understandings of the substantial seasonal dynamic pattern of the lake and revealed a subtle interannual change trend from 2000 to 2010. Besides, probability bounds were derived from the model for a region that was attacked by flooding. This ability of supporting decision making is shown in a simple management scenario. We conclude that in these respects the random spread process is a valuable addition to include uncertainty in wetland inundation modeling.

6.1 Introduction

Remotely sensed images have been integrated into spatial-temporal analysis procedures as a major source of landcover data in image mining. Image mining is a relatively new methodology focusing on extracting implicit knowledge, image data relationships, or other patterns that are not explicitly stored in large sets of remote sensing images. Stein et al. [167] structured a framework for image mining as a chain from identifying objects of interest, their parameterization, towards modeling spatial-temporal patterns on the basis of realistic assumptions about their distribution. Within this framework, work has been done on the storage of and handling of objects [32, 90]. For a complete spatial-temporal modeling, however, additional research on storage and representation of information extracted from land changes is required.

In image mining, spatial information is usually replaced by characteristics in lower dimension to simplify the analysis from large data sets. For example, the spatial information of a vegetation patch is represented by its area, perimeter, averaged vegetation index and its inherent variation. Previous research in spatial-temporal modeling has largely followed this line and has been dedicated to improve model fitting of the characteristics and clear up noise raised in image recording and processing [13, 51, 61, 70, 78]. Under such dimension reduction, however, interesting spatial information may be lost. Zhao et al. [197, 198] proposed a random set based model, in which the spatial spread of the phenomena at a moment in time is considered as a set, in order to maintain the information of the spatial distribution. They found that random sets were able to represent image objects with uncertain boundaries and that they provide useful information on the variation in extent using basic parameters like the covering function, the mean, level sets, and the variance.

Study of wetland inundation processes from a series of images is of great interest for ecology [15, 33, 71]. Due to its important ecological meaning and engineering function, the Poyang Lake area in China has received considerable attention [73, 128, 152]. Most studies so far focused on the spatial dynamics of the lake by its area and the seasonal dynamic pattern of the lake size [e.g. 73], whereas the shift of its extension or the frequency distribution of the inundation area has received attention only recently [199]. A typical Poyang Lake wetland flood spreads on flow-resistant vegetation that grows on a surface of a variable topography under varying weather conditions. It is thus representative for natural flooding systems elsewhere. Such floods depend on random (for observers) irregularities of both surface condition, such as depth, slope and flow resistance, and on inflow and outflow that both vary in time. These features point to a stochastic shape of the inundation area. In

addition, changes in shape are not identical every year, although the area of the lake has seasonal increase and decrease pattern in a year [71]. Inundation extent, therefore, requires to be modeled as a random spread process, whereas its spatial properties are best understood using a random set model for wetland inundation spread.

The random spread process (RSP) model was proposed by Vorob'ov to model fire spread based on random sets [179]. In earlier studies, some models of stochastic spreading processes were used to investigate tumor growth [24]. Each stage in an RSP is modeled as a random set due to the stochastic nature and the uncertainty of its geometric shape. Random sets have been applied as well to study the form variation of a group of objects [24]. Stoyan and Stoyan [169] used random sets in particle statistics to obtain the mean extent as well as shape fluctuations of sand grains. They showed that random set as a set-theoretic method can act as a supplement of other powerful means for particle classification. Recently, Zhao et al. [199] found that random sets reveal subtle changes of wetland hydrology that were not visible from a crisp approach. They concluded that the random set model enriches spatio-temporal modeling of uncertain spatial phenomena.

This paper aims at constructing a random spread process (RSP) model from a time series of images in order to explore the spatial-temporal pattern of the wetland inundation. The model focuses on the dynamic of water covering extent in the Poyang Lake wetland area. This area was observed with a large amount of images during a period of eleven years. A similar mathematical estimation of random sets is applied as in previous research [198]. In this study, however, we adopt it to analyze a time series of images, in particular to model the periodicity, the trend and the random components in the spatial-temporal change of the inundation extent. A simple management scenario is used to show how the model can be utilized by local managers.

6.2 Study area and data

6.2.1 Study area

Poyang Lake is located in the Northern part of Jiangxi Province, at the South bank of the middle and lower reaches of the Yangtze River (115° 47′-116° 45′E, 28° 22′-29° 45′N). Main water supplies to the lake are from local precipitation, five source tributaries (Ganjiang, Fuhe, Xiushui, Xinjiang, and Raohe), and from the Yangtze River during summer months. The seasonality of precipitation and water inflow have led to significant variation in the lake's inundation area. An extensive levee system of approximating 6400 km was built around the Poyang Lake

to control flood and to facilitate management [152], leading to numerous small lakes and tributaries, especially during the dry season when the lake is divided into many connected and disconnected segments separated by the exposed floodplains. For this reason, it is difficult to define the exact boundary of Poyang Lake. Generally, the large water body during the wet season together with three lateral lakes in the south (Kangshan basin, Junshan and Qinglan Lakes, Fig. 6.1) was considered as the lake boundary.

The size of the lake area in Poyang Lake is not well known. In most of studies on Poyang Lake, it is only mentioned briefly as "no more than 1000km^2 in low-water seasons to approximately 4000km^2 in high-water seasons" [79, 128, 153, 184]. Studies which concern the lake size change either adopt a different region boundary for calculating the lake size as there is no unique and official boundary definition of Poyang Lake [38], or their conclusions are derived from images at a different temporal scale such as images within a year e.g. 2000 [73]. In order to address the inherent uncertainty in these estimates, this study uses the random set process.

6.2.2 Images and preprocessing

We used the same series of images from the previous study and the pre-processing steps were detailed in [38, 39].

MODIS Level-0 data collected by the Terra and Aqua satellites between 2000-2009 and 2002-2009, respectively, were obtained from the U.S. NASA Goddard Space Flight Center (GSFC). Of the 9000 data granules covering the study region between February 2000 and December 2010, 620 cloud free data granules were selected for this study (Table 6.1).

The Level-0 data were processed using SeaDAS software (version 6.1) to generate calibrated at-sensor radiance, first corrected for Rayleigh scattering and gaseous absorption effects, and then converted to Rayleigh-corrected reflectance for each pixel (R_{rc}). R_{rc} data were geo-referenced to a rectangular projection with an error of less than 0.5 pixel.

The water areas were extracted from each image by using the Floating Algae Index (FAI). The FAI, based on band subtraction from the red to the shortwave-IR, i.e. R_{rc} data at 645nm, 859nm, and 1240nm, is stable against environmental and observing conditions such as aerosol type and thickness and solar geometry [71]. A gradient image was derived from the FAI image, where a pixel's gradient was determined by the FAI difference from the neighboring 3×3 pixels. Because of the strong water absorption, water reflectance at 859 nm is much lower than of land, resulting in strong FAI gradient at the water/land boundary. The mean FAI value of the pixels with maximum-gradient was chosen as the

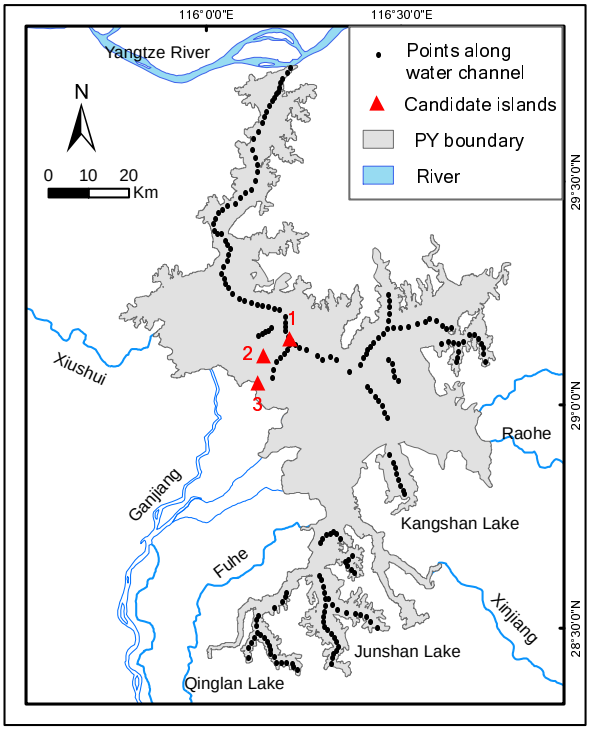

Figure 6.1 The boundary of Poyang Lake defined in this study.

Table 6.1 Number of cloud free MODIS images used in this paper.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
Jan	0	2	5	6	2	2	6	8	6	9	3	49
Feb	1	1	1	4	5	1	1	8	6	4	3	35
Mar	5	2	1	3	5	3	2	3	6	4	5	39
Apr	1	1	2	3	5	7	3	3	4	8	2	39
May	2	3	2	4	3	1	3	6	5	9	8	46
Jun	2	1	1	1	2	6	1	1	1	1	1	18
Jul	4	5	9	6	4	2	2	2	2	4	3	43
Aug	1	1	3	4	3	1	3	3	2	5	9	35
Sep	3	5	5	7	7	8	9	7	5	5	4	65
Oct	2	4	11	14	12	10	4	8	4	11	9	89
Nov	5	7	5	9	10	5	6	13	15	10	11	96
Dec	2	2	1	7	12	12	10	3	9	4	4	66
Total	28	34	46	68	70	58	50	65	65	74	62	620

threshold to separate water from land. Pixels with FAI values smaller than the threshold were considered as water.

6.3 Method

6.3.1 Definitions of random sets in flooding mapping context

The probability $\Pr(y \in \Gamma)$ that the point y belongs to the random set Γ is called the covering function, $\Pr(y \in \Gamma) = EI_{\Gamma}(y)$. The support set is defined as $\Gamma_s = \{y : \Pr(y \in \Gamma) > 0\}$. All points y with $\Pr(y \in \Gamma) \geq p$ constitute a p-level set. The 1-level set is called core set Γ_c. The 0.5-level set is called median set Γ_{med} which minimizes the square deviation:

$$\sum_{y}(\Pr(y \in \Gamma) - I_{\Gamma_{med}})^2 = min \sum_{y}(\Pr(y \in \Gamma) - I_A(y))^2 \tag{6.1}$$

among all sets $A, A \subseteq \Gamma_s$. In this sense Γ_{med} is the set that best fits the random set Γ.

The random spread process (RSP) is defined as a sequence of random sets, $\{\Gamma^t\}_{t\geq 0} = \{\Gamma^0, \ldots, \Gamma^t\}$, which describe the states of the process at discrete times t. Considering a water flooding process as an RSP, at each moment t, the water flooding in the point x attacks the neighboring unflooded point y with a given probability $\Pr_x(y)$. All y which have started being flooded between time t and $t+1$ from the inundated point x form a random set S_x, called the local flood spread random set (l.f.s.random set). The $\Pr_x(y)$ are the covering function of S_x: $\Pr_x(y) = \Pr(y \in S_x)$. The union of all random set S_x for all flooded locations x of Γ^t yields the next random set Γ^{t+1}, which describes the state of the flooding at time $t+1$: $\Gamma^{t+1} = \bigcup_{x\in\Gamma^t} S_x$.

Given the deterministic outcome F of random set Γ^t, the conditional probability $\Pr(y|F)$ that point y ($y \notin \Gamma^t$) attacked from the set $\Gamma^t = F$ will be flooded at time $t+1$ ($y \in \Gamma^{t+1}$) equals

$$\Pr(y|F) = \Pr(y \in \Gamma^{t+1} | y \notin \Gamma^t, \Gamma^t = F). \tag{6.2}$$

If we have a sample of n sets Γ^{t+1}: $\Gamma_1^{t+1}, \ldots, \Gamma_n^{t+1}$, then the covering function $\Pr(y \in \Gamma^{t+1})$ can be estimated by:

$$\hat{\Pr}(y \in \Gamma^{t+1}) = \frac{n_y}{n} \tag{6.3}$$

where $n_y = \#\{i : y \in \Gamma_i^{t+1}\}$ is the number of the set Γ^{t+1} containing the point y.

The risk of attacking some part, B, by water at the moment $t+1$, is defined as the probability:

$$\Pr(\Gamma^{t+1} \cap B \neq 0) \tag{6.4}$$

$$\hat{\Pr}(\Gamma^{t+1} \cap B \neq 0) = \frac{\#\{i : \Gamma_i^{t+1} \cap B \neq 0\}}{n}. \tag{6.5}$$

The hazard of occupying B is defined as the probability:

$$\Pr(\Gamma^{t+1} \supseteq B) \tag{6.6}$$

$$\hat{\Pr}(\Gamma^{t+1} \supseteq B) = \frac{\#\{i : \Gamma_i^{t+1} \supseteq B\}}{n}. \tag{6.7}$$

Equation 5 and Equation 7 are used to calculate the probability bound of being flooded in the management scenario.

6.3.2 Flooding process modeling based on random sets

To model the monthly trend of flooding, we adopt the month as the time unit. The RSP is denoted by $\{\Gamma^t\} = \{\Gamma^1, \ldots, \Gamma^{12}\}$, where Γ^t is the random set for month t, representing the stochastic shape of the flood in that month. Observations within that time unit, e.g. the flood extents extracted from images at 2000-01-01 and at 2001-01-20, are realizations of the random set Γ^1. Since the flooding extension in the month t in every year generally has similar spatial ranges (modeled by Γ^t), the $\{\hat{\Gamma^t}\}$ estimated from several years' observations should reflect the monthly dynamics of the flood process, as well as the randomness in the process.

For a monthly random set, the support set and core set delineate the extents of the possible maximum flooding area and the surely water covered area respectively. The median set represents the averaged water covering area. Extents of the core set and the support set are sensitive to the water area extracted from a single image. Therefore, we estimated the support set at the 0.05 probability values (0.05-level set) and the core set at the 0.95 probability values (0.95-level set), to avoid extremes from seriously noisy images.

To explore the interannual change of the flooding among successive years, we use year as the time unit, and define the RSP as $\{\Gamma^T\} = \{\Gamma^{2000}, \ldots, \Gamma^{2010}\}$. All the observations from images in the year T are realizations of the random set Γ^T. Instead of indicating uncertainty and randomness as in monthly random sets, the covering function of Γ^T, $\Pr(\mathcal{y} \in \Gamma^T)$, reflects the spatial distribution of the flooding frequency at any location $\mathcal{y}$. The core set represents the permanent water covering areas and the median set represents the area covered by water for longer than half of the year. The support set contains all possible water covering areas that appeared at least once in a year.

We group images by a time unit, either month or year, and treat each group as a set. Due to the randomness and noise in the set, they are considered as random sets. By organizing random sets in a time order, we model the random process as an RSP. Then we visualized the spatial extents of random sets by means of their covering functions and representative sets, i.e. the core set, the support set and the median set. Such a visual exploration is used to get a fist impression on the general trend of the dynamic and spatial pattern of the process.

6.3.3 Similarity analysis: spatial pattern of RSP

We propose a similarity analysis to compare any two states in an RSP represented by random sets, in order to detect clusters in the spatial pattern of the process. A Cov-Dist matrix is designed to summarize the spatial distribution of a covering function. We build a distance map as a spatial reference, overlay with the random set and count the number of pixels with different covering probability distributed in each distance range. The distance can be the absolute distance or alternatively a ratio distance between 0 and 1, where a value close to 0 indicates a close distance to the central water channel and a value close to 1 indicates a close distance to the lake boundary. The location of central water channel and lake boundary are shown in Fig. 6.1.

In this study, we divided a covering function map into ten level sets and a distance map into ten categories. Each covering function map is overlapped with the reference distance map and gets a attached Cov-Dist matrix M_t with ten by ten elements. The columns of the matrix are the 0.1, 0.2, ..., 1-level sets of the random set, and the rows are the distance categories 0-0.1, 0.1-0.2, ..., 0.9-1. The matrix element (i, j) is filled with the number of pixels belonging to the i^{th} level set and located in the j^{th} distance category.

The similarity of two random sets, e.g. M_t, M_{t-1}, in an RSP can be measured by the differences or correlation between their Cov-Dist matrices. The difference matrix $\Delta(M_t, M_{t-1}) = |M_t - M_{t-1}|$ summarizes the change size of each level sets at different locations. $corr(M_t, M_{t-1})$ gives a correlation coefficient of two Cov-Dist matrices and is a quantitative measurement of their similarity. By comparing all random sets in an RSP, we can get $corr(M_1, \ldots, M_t)$ as a correlation coefficient matrix and a matrix element (i, j) is filled with $corr(M_i, M_j)$. This correlation coefficient matrix of an RSP can be visualized to find clusters in spatial pattern in the RSP.

6.4 Results

6.4.1 Trend and randomness in the change of lake size

The mean trend in Fig. 6.2(a) shows a clear seasonal pattern and yearly replication of the dynamics of the lake size. The monthly lake sizes were averaged from the 11-year observations (by rows in Table 6.1) for the twelve climatological months. It increases from January, reaches the peak in July and shrinks gradually to its original size until December. These data were fitted by a sine curve ($y = a\sin(2\pi x/b + c) + d$), resulting in a high correlation coefficient ($R_a = 0.97$) and a relatively small RMSE (RMSE = 95.53km^2).

Next, we implemented the curve fitting on the daily water area data for each year to check the generality of the trend. Results show that they follow the sine curve but with different degree of goodness of fit. Among the eleven fitted models, the lowest correlation coefficient is observed in 2002 ($R_a = 0.63$) and the highest correlation in 2006 ($R_a = 0.89$). Fig. 6.2(b) illustrates the daily water area data in 2002 and 2006 and corresponding fitting results. The fitted models have different parameters and confidence widths, indicating noise around data, thus randomness besides trend. The shift of curves in time can be caused by the change of local precipitation like postponed raining season, or by human interference like the impoundments of the reservoir on Poyang Lake. Above results show that the flooding process, indicated by the size of water area, follows a general seasonal dynamic pattern in every year but also shows various randomness. This characteristic triggers us to adopt RSP and random sets for the spatial-temporal modeling.

6.4.2 Flooding process modeling based on monthly random sets

Fig. 6.3 shows the flooding spread process in the 11-year observation period that was modeled by twelve monthly random sets and is demonstrated by their covering functions. To facilitate further comparison on the spatial-temporal distributions of the inundated areas, we extracted the extents of the support set, the core set and the median set, as shown in Fig. 6.4. We observe that during the dry season (November to February), small water areas distribute mainly around the water channel and have similar spatial pattern. In the peak flooding months (July and August), the core sets have the largest extents in the center of the lake and the median sets cover the majority of the lake. Substantial changes happen in the other months, e.g. flooding from March to June and recession in September and October. This temporal pattern of the water area change can also be observed from Fig. 6.5. The water area averaged from 11 years has seasonal variation between 1164km^2 and

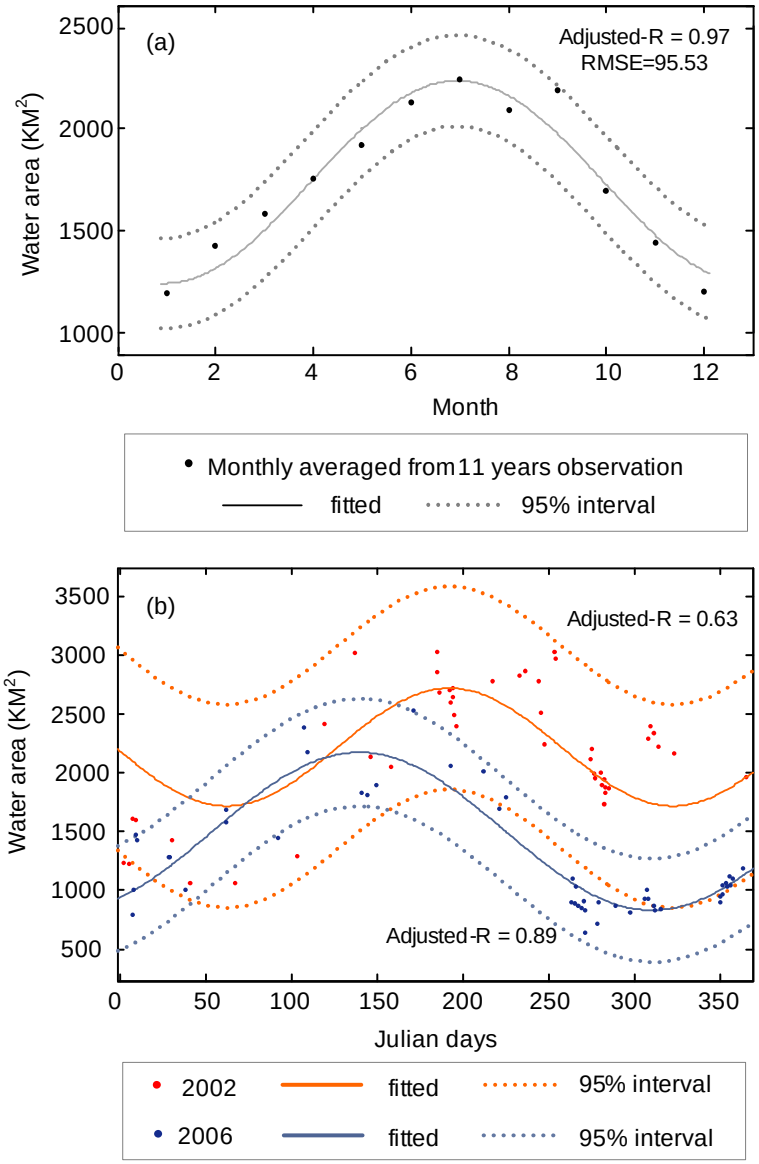

Figure 6.2 Fluctuation of water area in a year, modeled by the sine curve. (a)monthly averaged water area from 2000 to 2010; (b)daily based observations in year 2002 and 2006.

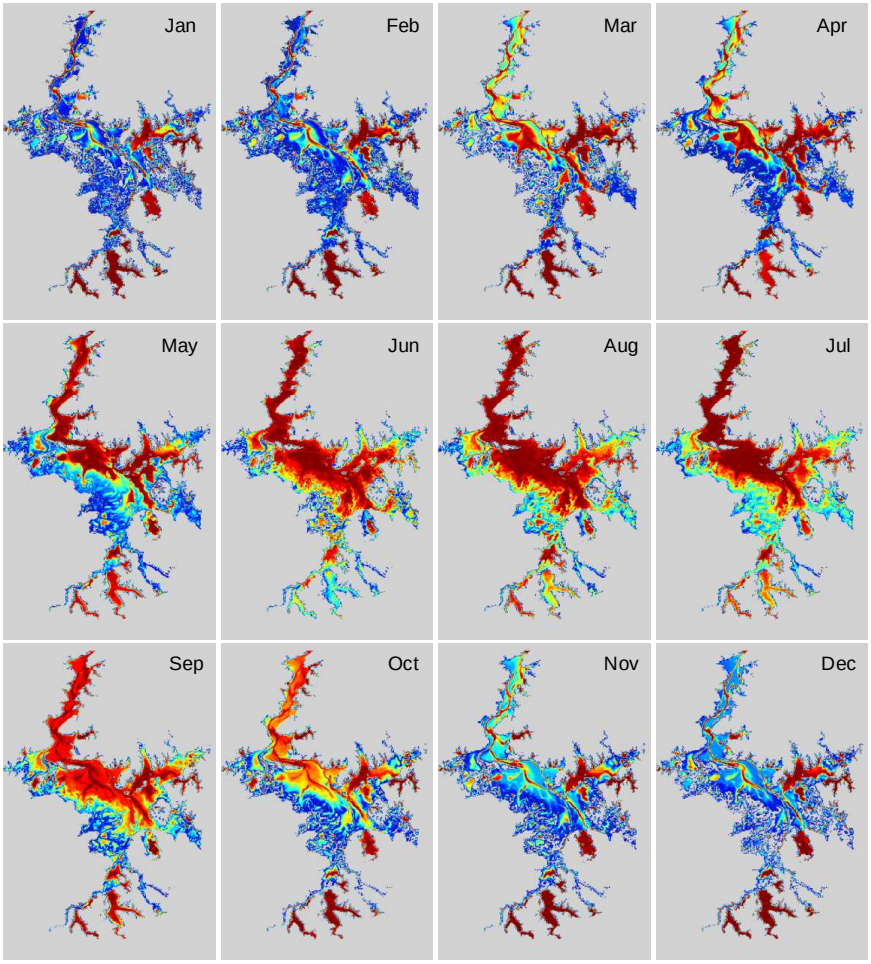

Figure 6.3 Main coverage of water and randomness in observations for every month, represented by covering function of monthly random sets (the legend refers to Fig. 6.6)

2852km^2. In the dry season, the water area is less than 1400km^2, with the average of approximately 1227km^2. The area in the peak flooding months is doubled and becomes larger than 2600km^2. The large difference in water areas between dry and flooding seasons, as well as the large standard deviation 635km^2 of the water areas in twelve months together show the dynamics and the large seasonal variation of the lake.

Three time tracks in Fig. 6.5 show changes in orders of support, median and core sets in the processes of flooding and recession. During

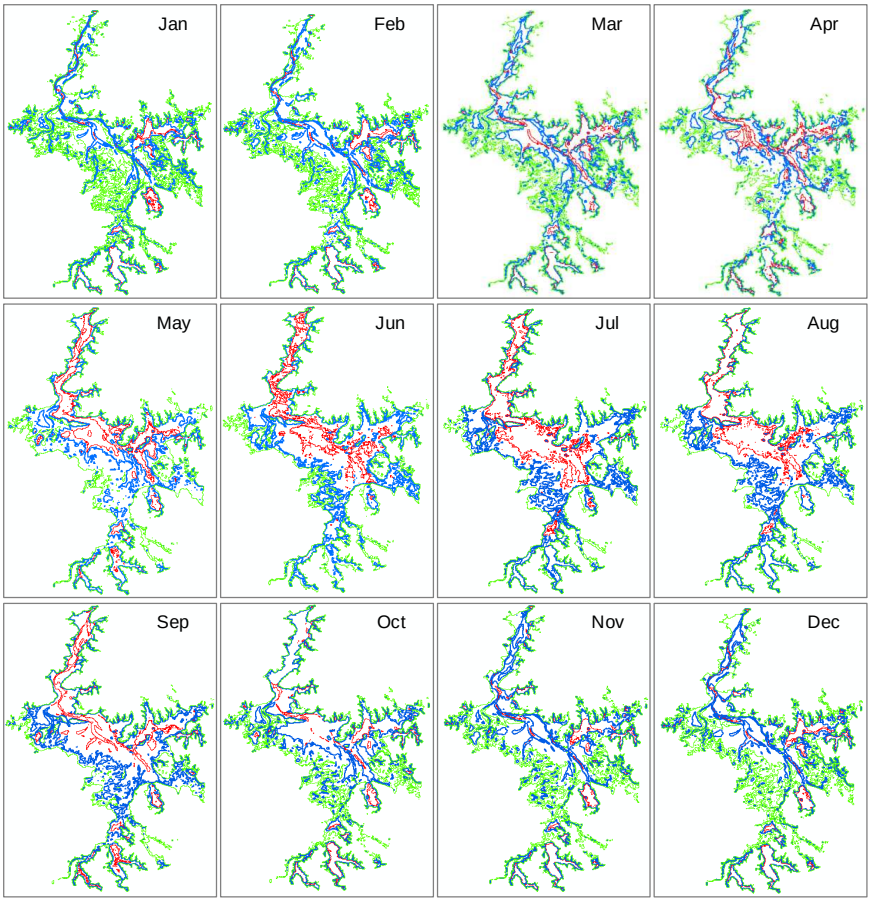

Figure 6.4 The extents of representative sets of monthly random sets (the legend refers to Fig. 6.7)

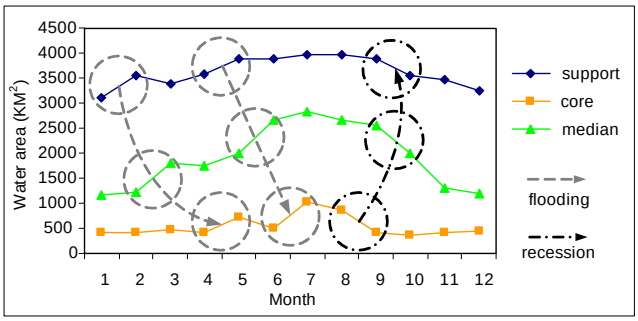

Figure 6.5 The sizes of representative sets of monthly random sets and its temporal trend

flooding, the support is first enlarged, and followed by an increase in the size of the median set, and finally the core set is extended. As flooding continues, therefore, probabilities of inundation increase, resulting in a larger inundation area with a higher certainty. This order is reversed during recession. The spatial change of this pattern is also reflected in Fig. 6.4. The median set first enlarges from February to March (1239km^2 to 1806km^2), followed by a gradual expansion of core set from April to May (429km^2 to 725km^2) around the central water channel. The median set expands again from May to June (2013km^2 to 2683km^2), and is followed by a expansion of the core set from June to July (506km^2 to 1018km^2). During recession, the core set shrinks greatly in a short time from August to September (877km^2 to 426km^2), and is followed by the median set that dramatically reduces its size from September to November (2571km^2 to 1305km^2).

6.4.3 Interannual dynamics of flooding frequency

Covering functions in Fig. 6.6 show the flooding frequency in each year between 2000 and 2010. Generally, covering functions before 2003 have higher values that cover a larger area than after 2003, i.e. the frequently flooded area is larger before 2003. In contrast, the spatial distribution of the flooding frequency in 2010 is different from other years. In Fig. 6.6, the covering function in 2010 shows large yellow areas, indicating a flooding spread over a larger area for a longer time. Whereas the red area only covers a relatively small areas, similar to those in other years. This means large flooding does not expand the size of the permanently inundated area and the lake area remains the same during the dry season.

Spatial distributions of yearly core, median and support sets are

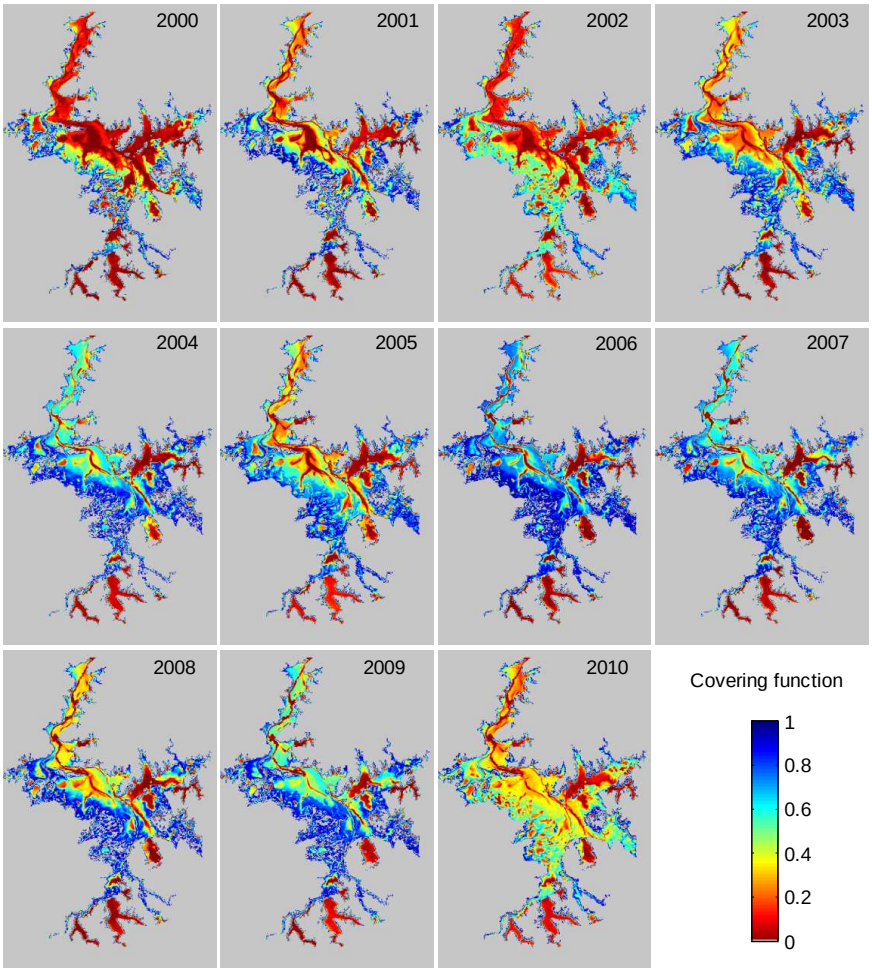

Figure 6.6 Spatial pattern of flooding frequency for every year from 2000 to 2010, represented by covering function of yearly random sets

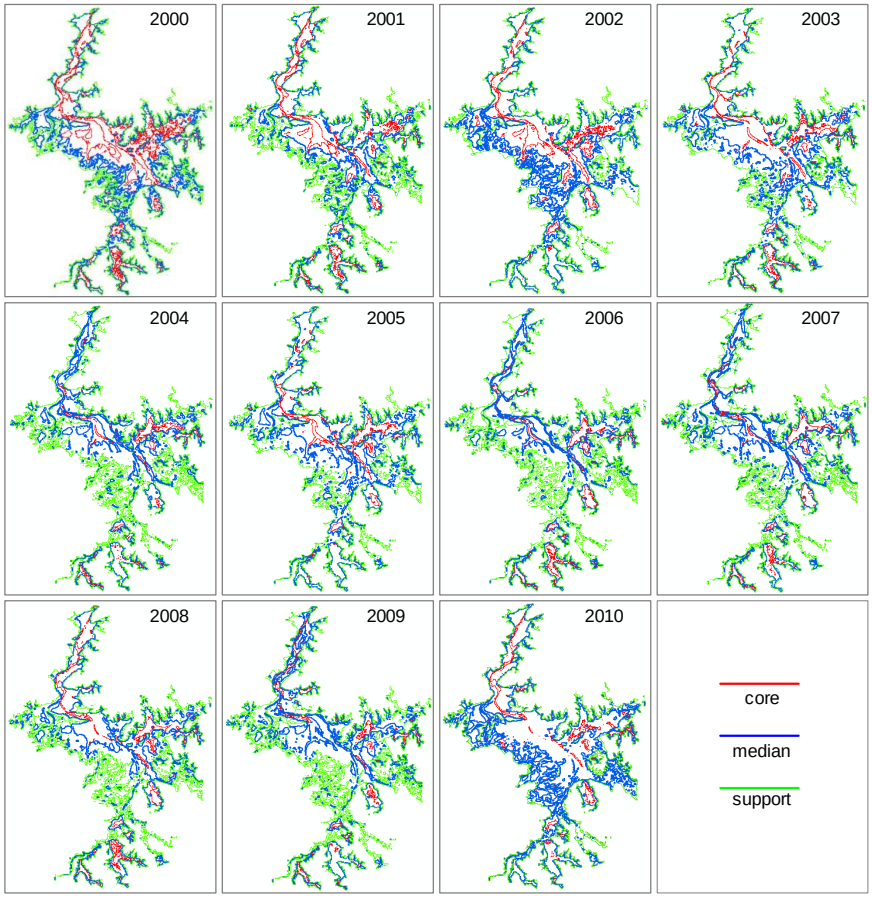

Figure 6.7 The extents of representative sets of yearly random sets

shown in Fig. 6.7. From 2000 to 2003, the core set along the water channel has a size of approximately 488km^2. This size is larger as compared to the other eight years of averaged size 273km^2. The core set shrinks from 385km^2 in 2003 to 192km^2 in 2004, reducing the size of the permanent water areas with almost 50%. The median set also declines in size from 2262km^2 in 2003 to 1306km^2 in 2004. The median set after 2003 almost replaces the previous location of the core set before 2003, except those in 2005 and 2008. The small core set and large median set in 2010 together hints to a different flooding situation in that year. Its median set of size 2851km^2 shows a severe flooding disaster during the wet season, with a long flooding time (longer than half the year) and a large inundation area (covering almost the whole

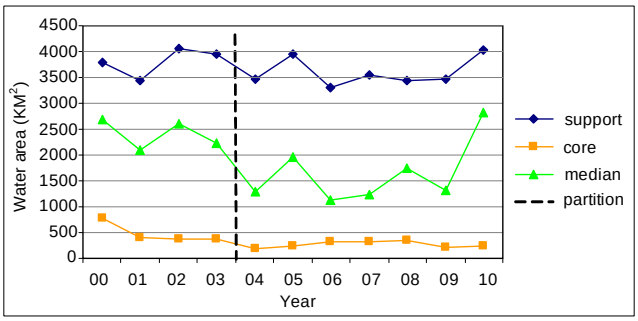

Figure 6.8 The sizes of representative sets of yearly random sets and its temporal trend

lake). The permanent water area of size 245km^2 indicates that the inundation area during the dry season is small, as occur in the years from 2006 to 2009.

The temporal trend of water size change indicated by core sets, support sets and median sets, is shown in Fig. 6.8. These three sets, in particular the median set, show decreasing trends. The year 2003 can be regarded as a partition year by dividing the eleven observation years into two periods. The size of the water area is larger before the partition than after it, with an exception in the year 2010 when the size of the support set and the median set do not follow the trend.

6.4.4 Similarity of the spatial pattern

For every random set, a Cov-Dist matrix summarizes its spatial distribution and $\Delta(\cdot)$ between any two matrices helps to analyze the changes. We use three Cov-Dist matrices from May to July and their $\Delta(\cdot)$ as examples to demonstrate the way of interpretation. Comparing the three Cov-Dist matrices (Fig. 6.9(a)), pixels belonging to the 1-level set, i.e. the first column on the right, increase and expand from low distance value near the water channel to a high distance value near the water boundary. Moreover, a large number of pixels located near the water boundary, i.e. the fist row at the top, have increased probability values. In addition, the cells in light blue color become more clustered around the high distance value and high level set at the upper right conner. These observations can also be seen from Fig. 6.3, but the matrix gives a more concise summary, and it contains more detailed information than the representative sets in Fig. 6.4.

The $\Delta(M_{Jun}, M_{May})$ matrix in Fig. 6.9(b) shows an increase in the number of pixels above the 0.7-level set and a decrease in the number

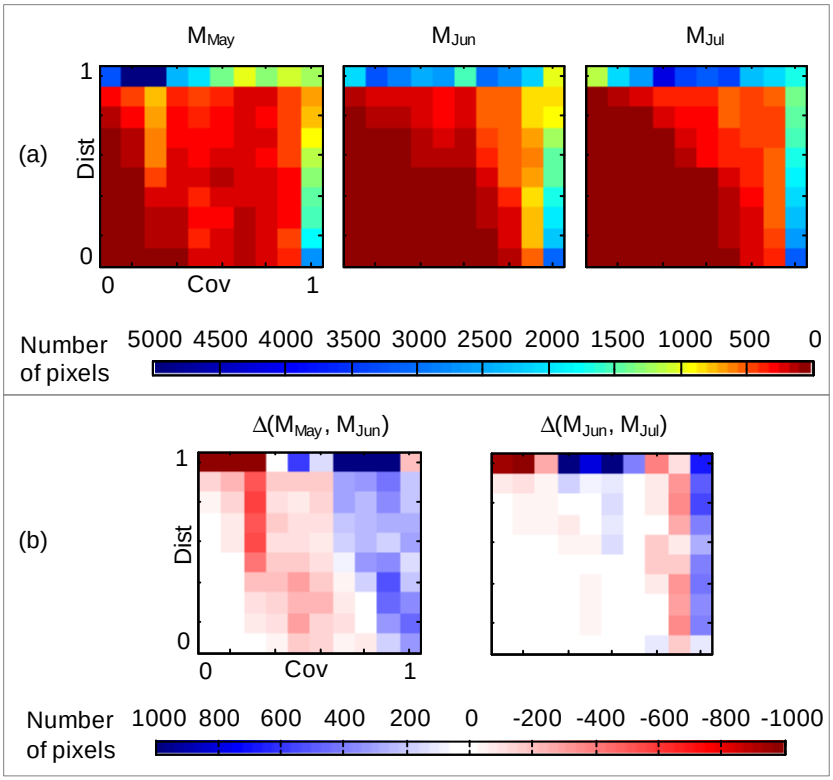

Figure 6.9 Examples of Cov-Dist and $\Delta(\cdot)$ matrix. (a)Cov-Dist matrix of monthly random set of May, June and July; (b) $\Delta(\cdot)$ matrix showing the difference between the three random sets.

of pixels below 0.6-level set. Most changes in the $\Delta(M_{Jul}, M_{Jun})$ matrix, however, happened in the 0.9-level set and the 1-level set, and far from the water channel. The less number of pixels in white in $\Delta(M_{Jun}, M_{May})$ as compared to $\Delta(M_{Jul}, M_{Jun})$ indicate more differences between June and May than between July and June. Values for $corr(M_{Jun}, M_{May}) = 0.82$ and $corr(M_{Jul}, M_{Jun}) = 0.93$ further support this conclusion that the spatial distribution of random set for June has more similarity with that for July.

Correlation coefficient matrices in Fig. 6.10 show the similarity of spatial pattern between any two monthly random sets or yearly random sets. The clusters in Fig. 6.10(a) confirm the seasonal pattern of the flooding. The most similar patterns are found from November to February (dry season), July and August (flooding peak), with high correlation

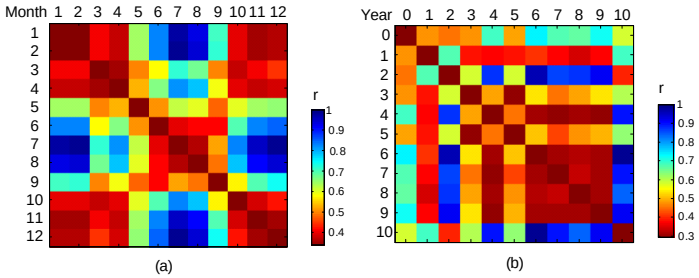

Figure 6.10 The similarity of spatial pattern among monthly random sets (a) and yearly random sets (b).

coefficients ($r \geq 0.97$). Less similar patterns are shown around the dry season months and flooding peak months in March, April and June. May and Sept have relatively low correlation with other months ($r < 0.85$, expect $r = 0.91$ of September and June), which indicate either a dramatic flooding or recession process occur.

In Fig. 6.10(b), the high correlations ($r \geq 0.95$) show that the spatial distributions of flooding frequency are similar in 2003 and 2005, and also that those in 2004, and in the years from 2006 until 2009 are similar. The correlation coefficients of any two years between 2004 and 2009 have high values ($r_{min} = 0.78$, $r_{mean} = 0.92$ and $r_{max} = 0.99$). This high similarity shows a generally regular spatial pattern of flooding frequency among years after 2003 and thus coincides with the previous result that year 2003 is a separator in the eleven years. Besides, 2010 has the lowest correlations with other years ($r_{min} = 0.29$, $r_{mean} = 0.54$ and $r_{max} = 0.88$), which confirms that an extreme flooding event happened that year.

6.5 A management scenario

The spatial distribution of inundation areas in the Poyang Lake wetland is of a particular concern, because it serves the basic source of information for management decisions. Here we present an allocation scenario to demonstrate the virtue of the random set model from the decision maker's perspective. A manager may decide to establish three small islands of size 1km $\times$ 1km each, e.g. for providing shelves for over-winter birds or for constructing automatic hydrological stations. The best available data collected from the last eleven years is the time series of MODIS images which were used in this study. Before allocating the islands, the manager needs information about the flooding situation

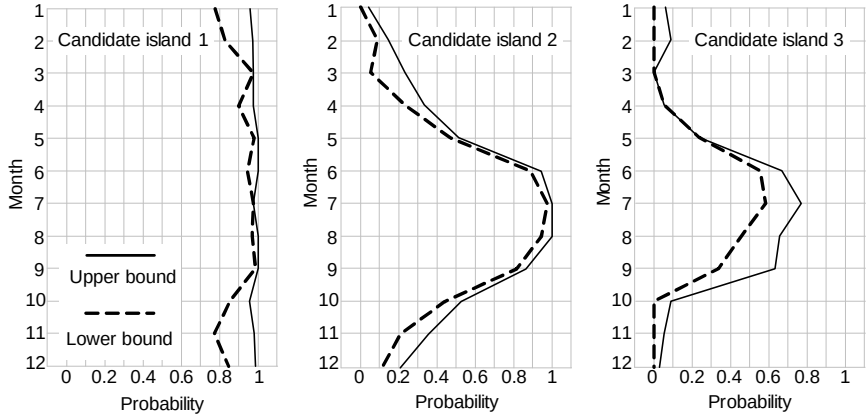

Figure 6.11 Probabilities of being flooded at three candidate islands

at the candidate square areas, including the probability of be attacked by flooding and the probability of be fully occupied by water in every month of a year.

The prior locations of three candidates are shown in Fig. 6.1. Based on the monthly random sets and according to Equations 5 and 7, he is now able to calculate the probability of inundation. The risk of attacking some parts of one of the islands by flooding is shown as the upper bound, and the hazard of occupying the whole island by flooding is shown as the lower bound in Fig 6.11. We observe that candidate locations 1, 2 and 3 have different flooding risk in the twelve months period. Candidate 1 has more risks of being covered by flood than the other candidates during a whole year, all probabilities being above 0.77. Candidate 2 will be attacked by flooding in July and August, and the risks in other months are much lower. Candidate 3 has the safest location against flooding as it is surely not occupied by water during the dry season, whereas in the flooding peak months the probabilities of being attacked by water are less than 0.8.

The above example shows that the upper and lower bound of probabilities helps the manager to explain the certainty of his results instead of a simple 'yes' or 'no'. The results can serve the allocation decision, if the manager should consider the protection against flooding, or reversely want to take the advantage of inundation. In a more realistic project, other factors may also constrain the decision, such as the suitability of soil type as the building base, distance from the residence and the project budget.

6.6 Discussion

A process modeled by an RSP contains two information components: on component described by the spatial pattern and one by temporal pattern. A random set describing individual states of the RSP is built on the spatial information of the process. A time series of random sets records the evolution of the process and temporal information is stored by the time order of the random sets. The random set itself, however, does not contain the order expressed by 'before' and 'after' that is typical for temporal information. For example, the realizations of the random set showing the flooding frequency in a year are from the daily images and have a calendar order. But after they are organized as a random set, it is not possible to distinguish the start or ending time of the flooding. In this sense, the original temporal information is lost. As a consequence, an appropriate time unit needs to be selected when constructing random sets during the RSP process modeling to ensure that the loss of temporal information within the random set is acceptable.

The daily images used in this study were organized in two independent ways, i.e. by monthly random sets and yearly random sets. In doing so, we could analyze both seasonal fluctuations as well as the interannual variation of the inundation area. Note that a monthly random set was estimated from observations for that month over eleven years. It reflects the monthly flooding during the full study period, and therefore does not belong to any specific year. By considering interannual variation and monthly variation within the subcomponents in each year, we may consider to construct a hierarchical structure. With such a hierarchy of random sets in a single RSP, it could be possible to combine several temporal scales within one particular system. For example, the thirty realizations derived from daily images in January 2011 are currently used to estimate the distribution of the monthly random set for January and its median set can act in turn as the realization for the yearly random set at a larger temporal scale. The yearly random set for 2011 can be estimated from the twelve monthly realizations, being the median sets of monthly random sets between January and December 2011. Calculating the median set is one possible way of linking random sets at different scales, i.e. converting a random set in a lower level to a realization of the random set in a higher level. Using other p-level sets is also possible but is left for future study.

Through constructing a random set from a group of images in a specific unit of time, we obtain the frequency of covering, the core and the support extension. The core and support have a meaning that are different from the minimum and maximum inundation areas as derived from single images [like in 39, 139]. They are not equal to real

inundation areas that might be observed from images. Instead, they are the combination of the spatial extents of different inundations: their intersection as the surely water covering area, i.e. the core set, and their union as the possibly inundated area, i.e. the support set. The maximum inundated areas in the eleven years are approximately 2000-3000km^2, whereas the possible water covering areas are approximately 3000-4000km^2. These large differences are caused by different extents of flooding fronts during the flooding season and thus indicate their spatial dynamics of the flooding fronts. Nevertheless, the minimum area and the support set have similar sizes of approximately 400-1200km^2. This relatively small difference indicates that the inundation area in the dry season has a relatively stable extension.

The median set delineates the area with a flooding frequency equal to or larger than 0.5. We adopted it as a representative set to show the mean extent of inundation due to its important mathematical characteristics in the random set theory. It has a practical meaning in the monthly random set that the majority of observations of each pixel is flooded during that month and in yearly random set that the pixel covered by water for longer than half a year. For other application purposes, we can select different p-level sets as meaningful representative sets. In this study, for example, we adopted the 0.05-level set as the support set to avoid that extreme inundation observed from one or two noisy images dominates the covering function.

Visualization of a random set by its covering function and representative sets gives us an intuitive impression on their spatial distributions. For a large number of random sets in an RSP, other quantitative visual exploration methods such as the Cov-Dist matrix may be useful as well. The Cov-Dist matrix takes both the probability information by the covering function axis and spatial information by the distance axis. The detail level of matrix information is controlled by the number of categories along both axes, but also constrained by the detail level of the original covering function map and the distance map. After transformation of the covering function to a Cov-Dist matrix, matrix differencing and correlation calculations allow us to make a quantitative comparison. Visualization of the correlation matrix, further, allows us to find the cluster pattern in the evolution of a process. In this sense, both qualitative and quantitative visualization contribute to the exploration in the image mining.

The Three Gorges Dam, constructed upstream of the Poyang Lake on the Yangzte River was impounded in June 2003, resulting in reduced water flow and water level in the river's downstream [152]. Feng et al. [38] derived the bottom topography of Poyang Lake between 2000 and 2009. They found an overall increasing trend in the lake's mean elevation showing that the lake has become shallower during the last ten years.

Results from our study on the changes in spatial extents of inundation areas coincide with their findings. The similarity of flooding frequency distribution after 2003 gives more evidences of the effects of the Three-Gorges Dam and impoundment of the reservoir on the Poyang Lake wetland since 2003. Further research on the other properties of the Poyang Lake is needed, such as the volume and discharge, in order to gain more understandings about the changes.

6.7 Conclusions

A random spread process model was constructed for monitoring the spatial-temporal pattern of wetland inundation, in which random sets were used to model the spatial extent of each stage. Main trends and random components in the spatial-temporal changes were quantified by basic properties of random sets, including the covering function, the median set and other level sets. The Cov-Dist matrix facilitated the visualization of random sets and related matrix operators helps to explore the similarity of various spatial patterns.

As a representative study area of wetland inundation, changes in the Poyang Lake during last eleven years were observed from more than six hundred multitemporal MODIS images. Substantial seasonal dynamics of inundation extents was found for the monthly random sets. Large differences in the spatial distribution of inundation extents were shown to exist between months from the dry season and those from the flooding season. Distributions in months from the same season showed similar spatial pattern. A unique spatial pattern occurred during those months that a dramatic flooding or recession happened. In addition, a subtle interannual change trend was found for the yearly random sets. Yearly random sets gave detailed information on the spatial distributions of inundation frequency and showed a shrinking trend from 2000 to 2009. 2003 is the partition year in the declining trend and 2010 breaking the trend as an abnormal year. At the end, by calculating the probability bound of being flooded for candidate areas in a management scenario, the RSP model showed its ability of providing diverse spatial-temporal information for decision makers.

Synthesis

7

7.1 Overview

The field of ecology is becoming aware of the importance of studying uncertainty when understanding ecological phenomena, predicting the state of ecosystems and communicating forecasts with policy-makers, managers and the general public. Priorities have been set for estimation, propagation and communication of uncertainty in typical ecological research, since ecological models are affected by all sources of stochasticity [21]. If this is ignored or if it is impossible to quantify, the information content will be reduced and decisions become complex or totally wrong. A low information content is usually caused by inaccurate model inputs, uncertain models and inappropriately estimated parameters [21, 80].

Data play a large role in ecological analysis. Ecological data, however, are observed with measurement errors and often limited by the expense of data collection. Moreover, most ecological phenomena of interest can only be studied by integrating various sources of data such as field samples, remotely sensed images and meteorologic data from weather stations. Such data integration increases the uncertainty of model outputs. For example, ecological niche modeling uses input variables such as climate data, soil type, water depth, and land cover to predict the distribution of species in geographic space on the basis of a mathematical representation of their known distribution in realized ecological niche. The extent to which such modelled results reflect real-world species distributions will depend on a number of factors, including the quality of the available environmental data layers and the nature, complexity, and accuracy of the models used. The ecological model, being an approximation and simplification of reality, is accompanied by model uncertainty and parameter uncertainty. There is no reason to avoid such models, but care should be taken on their quality.

Most ecological studies model uncertainty by representing model inputs, parameters and model outputs as stochastic fields, variables or objects [80]. Ecologists who use soil maps, vegetation maps or land cover maps as model inputs often recognize the limitations of crisp representations and prefer more detailed representations of spatial gradients or spatial patterns of extensional uncertainty [14]. For example, mapping the extents of wetlands is a basic practice for wetland inventory, change tracking and ecosystem modeling. It could suffer, however, from problems in identifying the uncertain boundary between water and vegetation. To extract inundated areas in such a natural landscape from images, we thus need to consider wide transition zones from open water to vegetation when producing input maps for further modeling.

From the users' points of view, wetland managers and decision

makers who are concerned about ecosystem dynamics, are particularly interested in the location, shape and size variance of water bodies and vegetation [97]. Transition zones among different vegetation communities, between water and vegetation are suited for different waterfowl species and serve as important habitats of migratory birds. In addition, frequently flooded shallow waters are an ideal environment for snails growing and thus have high risk of schistosome infection. Identification of these zones, therefore, provides important information for protecting the food habitat of endangered birds like Siberian Crane [186] and for predicting the distribution and preventing abundance of schistosome [188].

Random sets can serve as a robust uncertainty modeling tool to derive and store the probability distribution of uncertain input and output variables in an ecological modeling process. Random sets based approaches show their capabilities in providing the geometric estimation of objects with uncertain boundaries in this thesis. Objects with uncertain extents are desired to be represented by random sets and their probability distributions can be characterized for either stochastic model inputs or outputs. By means of random sets, more accurate maps of wetland land covers and inundation duration can be derived, and our understanding about the ecosystem dynamics are enriched. This information can be important for ecologists who attempt to locate the resources precisely, for example to use it as input in an ecological niche modeling. Whereas managers who are concerned about the forages and habitats of migratory birds, will benefit from this information when they calculate the area of wetland and make a decision on the water level control of the lake. The methods have only been developed and applied in limited wetland applications to address practical problems in this study, their value should be further explored in a wider ecological context and for more diverse situations.

The aim of this thesis was to develop different techniques based on random sets to represent image objects with indeterminate boundaries, quantify their extensional uncertainties, and address uncertainty modeling in a spatial temporal change analysis. To achieve these aims, five research questions and several objectives concerning the wetland application were formulated. Research findings related to these questions and objectives, general conclusions and recommendations for future works are presented in this chapter.

7.2 Research findings

7.2.1 Answers to research questions

1. How should random sets be defined and generated to model uncertain objects?

Random set theory has been viewed as a generalization of other uncertainty theories [58, 59, 101] and the relationships between those theories have been discussed in the past [3, 114, 118, 182]. To the best of our knowledge, however, no research has been done to apply random set theory to uncertainty modeling in geoscience and environmental research yet, though Cressie [24] recommended that it is a sound theory for modeling uncertain objects. Chapter 2 introduced several main concepts of random sets in the context of geo-information science for the purpose of modeling uncertainties of natural entities extracted from remote sensing images. Basic properties of random sets were defined in Chapter 2 and 3. The probability $\Pr(x \in \Gamma)$ that the point x belongs to the random set Γ is called the covering function, $\Pr(x \in \Gamma) = \mathrm{E}I_{\Gamma}(x)$. All the points with positive covering function construct the support set. All the points with covering function equal to or larger than p construct a p-level set of the random set. The 0.5-level and 1-level sets are called median and core sets respectively. The Vorob'ov mean is the p-level set whose area is equal to the accumulated area of the covering function.

Random sets were used to model natural landscape and image objects with uncertain boundaries in Chapter 2 - 6. This was based on the assumption that interpreting larger uncertain boundaries e.g. by digitizing, thresholding or segmentation will result in bigger differences and more randomness in different outputs. The overall set of outputs were modeled as a random set, and its variability was utilized to reflect the degree of uncertainty. Different realizations of a random set were generated by classifying image objects with a random parameter in the thresholding or segmenting algorithms. The distribution of the random parameter and image intensity function together determined the distribution of the random set.

Since the essential part of creating the random set model is to generate realizations that fully characterize its distribution, the probability distribution of the random parameter needs to be carefully determined. The uniform distribution was chosen in Chapter 2 when no emphasis needs to put on special values of the parameter. The normal distribution was chosen in Chapter 3 and 4, based on the assumption that pixel values close to object boundary have a higher probability to be selected by the image interpreters when labeling boundary pixels, than pixel values far away. In Chapter 5, the mixed Gaussian model was used to fit

as a density distribution to the image histogram and to determine the transition interval. Random parameter then followed the distribution of the histogram of the image pixels in the transition interval. In the spatial-temporal modeling in Chapter 6, the distribution of a random set represented the stochastic shape of the dynamic phenomena in a pre-defined time interval. Observations within that time interval or unit were considered as realizations of the random set.

2. How do random sets quantify the extensional uncertainty of image objects?

Extensional uncertainty is the uncertainty in identifying geometric elements that describe the spatial extent of an object [112]. Due to the extensional uncertainties, spatial objects with uncertain boundaries usually face difficulties in determining their spatial extents when extracted from images. To quantify the extensional uncertainty of objects, uncertain boundaries of image objects need to be extracted and modeled first. Based on the basic concepts of random sets introduced in Chapter 2, Chapter 3 has made random sets generation and statistics applicable in image segmentation practices. Results showed that several characteristics of extensional uncertainty of segmented objects can be quantified numerically and spatially by random sets.

Chapter 3 first examined the effect of a small degree of randomness of segmentation parameters on the spatial variance of extracted image objects and modeled the extracted objects by random sets. By comparing the statistical parameters (Γ_c, Γ_m, Γ_s and Γ_{var}) of random sets and the spatial distribution of centroids, randomness of segmentation parameters has shown different effects on extracted features when objects have different extensional uncertainties. Several indices, including n_ϵ, SD, CV and $Ratio_{ASM}$, were proposed to summarize the degree of extensional uncertainty numerically.

The number of iterations n_ϵ to achieve a stable covering function is a good indicator of boundary sharpness. Objects with relative clear boundaries have covering functions that become stable rapidly as the number of iterations increases. In contrast, the convergence curve for an object with a gradual transitional boundary stabilizes slowly, resulting in a large n_ϵ. The sum of the variances SD and the coefficient of variation CV have positive relations with the degree of uncertainty. But objects with the same value of SD can have very different CV values since CV takes object area into account. Therefore, n_ϵ and SD were used for indicating the absolute amount of uncertainty, and CV for the relative amount of uncertainty per area unit. The asymmetry property of covering function reflects the uneven gradients in different positions

or along different directions of the broad object boundary. This property was quantified by the asymmetry index $Ratio_{ASM}$. The $Ratio_{ASM}$ is equal to zero for objects that have perfect symmetric boundaries. Whereas the value of $Ratio_{ASM}$ increases as the asymmetrical degree increases.

3. How can the accuracy of classified uncertain objects that are represented by random sets be assessed?

The validation of random sets in Chapter 3 relies only on sensitivity analysis and is partial supported from temporal changes analysis. Direct accuracy assessment is lacking due to the absence of detailed and synchronous field data. In Chapter 4, a sampling plan was designed specifically for a) investigating the zonal pattern of wetland grassland and b) assessing the accuracy of a random set model. Since the statistical distribution of the random set is expected to be close to the distribution of some variables collected in the field, statistical methods were utilized for measuring and testing the modeling results of random sets.

Chapter 4 first explored the corresponding measurable variables which were collected on the ground for validating the uncertain image objects modeled by random sets. Regression analysis and statistical distance (Kullback-Leibler divergence and total variation distance) calculation were applied to examine which variables are correlated to the covering function of random set and whether their probability distributions are similar. The variable with the lowest distance has the probability distribution most similar to that of the covering function. In addition, the samples were divided to four groups according to the 0-25, 25-50, 50-75 and 75-100 percentiles of the covering function and measured variables respectively, and the number of samples falling in each corresponding group was compared by Pearson's chi-square test. The result showed that *Carex* volume has the most significant relationship with covering function, followed by *Carex* coverage. The distribution of the covering function, however best reflects the distribution of *Carex* volume based on their closest statistical distance. Therefore, *Carex* volume and coverage can be the corresponding measurements on the ground, by which the covering function of random sets can be quantified and interpreted adequately.

To measure the thematic accuracy of the random set, the covering function was thresholded to obtain several important p-level sets such as support, mean, median and core, and those sets were validated using confusion matrices and the k statistic. Although the *Carex* volume explains a little more about the proportion of the variance of the covering function, due to practical reasons, only the *Carex* coverage variable was

used to report on the classification accuracy assessment. The accuracy of the random set varies from good to moderate according to the k coefficients derived from the confusion matrixes of support, mean, median, and core sets. The support set has a higher accuracy than the core set after making some adjustments, whereas the mean set has a better performance than the median set.

4. Does the random set model give more information than a crisp data model on the spatial dynamics of uncertain phenomena?

Use of hard classification and crisp data models simplifies the extraction and representation of wetlands. Artificially imposed crisp boundaries of inundation areas lead to a stepwise and imprecise map of water covering duration from monthly collected images, like in [73]. In Chapter 5, we differentiated transition zones from open waters and non-water areas, and modeled the wetland inundation extents (consisting of open waters and transition zones) as random sets. The extraction of transition zones offered extra information for locating the inundation extents and for calculating the areas of inundation. Further, the water regime was investigated by accumulating a series of random sets during one year and determining the water covering days (WCD) at the pixel level.

The results in Chapter 5 revealed that random set based WCDs (WCD-RS) outperformed crisp set based WCDs (WCD-ODF). WCD-RS better matched the underwater elevations and had a stronger correlation with the DEM than WCD-ODF. The WCD-WL, which was derived from the DEM and the daily water level, was also compared with WCD-RS and WCD-ODF. Results showed that WCD-RS had a stronger positive relationship with WCD-WL than WCD-ODF.

Detailed spatial configurations of WCDs become detectable through random set modeling. We observed stepwise changes in WCD-ODF, whereas WCD-RS were smoother and reflected more details in water covering change. The smoothing effects could be explained by utilizing a detailed characterization of the current status to reflect flooding and recession processes in the past. For example, a place classified as open water should have a larger probability to be covered by water during the whole observation interval (days from current image to previous image) than a place classified as a transition zone. Those classified as transition zones should be covered by water for a longer time than those classified as vegetation. Finally, we reached the conclusion that random sets contributed to the WCD mapping through smoothing the WCD variation captured by limited snapshots, and through providing more details than using a crisp method.

5. How should a random spread process be built based on random sets, that allows one to include uncertainty in spatial temporal modeling?

In Chapter 6, a random spread process model (RSP) was constructed for monitoring the spatial-temporal pattern of wetland inundation, in which random sets were used to model the spatial extent of each stage. The wetland inundation was observed with a large amount of images during a period of eleven years. The RSP was adopted to model the periodicity, the trend and the random components in the spatial-temporal change of the inundation extent.

A series of images was grouped by a time unit, either a month or a year, and each group was treated as a set. Due to the randomness and noise in the set, they were considered as random sets. By organizing random sets in a time order, the random process was modeled as an RSP. Main trends and random components in the spatial-temporal changes were quantified by basic properties of random sets, including the covering function, the median set and other level sets. The Cov-Dist matrix and related operators summarized and visualized the spatial pattern and quantified the similarity of different stages in the process.

7.2.2 Findings related to the wetland application

6. Wetland grassland and vegetation patches around Banghu lake

PLNNR is characterized by wetlands grassland and lakes. Field work were carried out in the area surrounding Banghu lake within PLNNR where the wetlands grassland has the largest area and show obvious transition zones. Three typical zones were found along transects in the field work from the river bank to the lake bank. The first zone is on the river bank. Flowered *Miscanthus* of 1-2m height appears at high elevations near the river bank, often mixed with *Cynodon*, *Carex*, *Polygonum*, and human planted poplar. Some of the flowered *Miscanthus* also have green leaves at lower height, and some shorter *Miscanthus* are not flowered. In areas where human activities are intensive, the *Miscanthus* has usually been harvested to approximately 0.5m. The second zone is *Carex* dominant, approximately 500m across in horizontal distance. The height of *Carex* ranges from 0.3 to 0.6m, and it thrives in the autumn, with very high density. Toward the lake bank, the height and density of *Carex* decreases and *Polygonum*, *Artemisia* and *Eleocharis* appear and are mixed together. The third zone is near the lake bank where elevation changes gradually. The indicators of low elevation are high soil moisture and plant such as *Cardamine* and young *Carex*. On

the bank, we found wet soil, *Cardamine*, dead and dry *Potamogeton* and *Vallisneria* covering to the soil, and shooting up *Carex* with very low density. The farthest place we reached on the lake bank is covered by 10cm of shallow water with dead *Potamogeton* and *Vallisneria* beneath. Bird droppings and feathers were frequently found and birdcalls were very evident.

In conventional crisp segmentation, the extracted *Carex* patch is represented by a crisp boundary. The area inside the boundary is regarded as dominated by *Carex* and the area outside is not. This hard partition of continuously vegetated area ignores the extensional uncertainty, if looking at the original Landsat TM or HJ image where pixel values change gradually from one class to the other or in the field where changes between different types of vegetation dominance are smooth. In Chapter 3, the random set model was applied for representing uncertain boundary of *Carex* patches, and an accuracy assessment on the modeling results was performed in chapter 4. Results show that modeling the extensional uncertainty offers detailed information on the spatial distribution of *Carex* patches, especially for patches with uncertain boundaries. Several characteristics of extensional uncertainty of *Carex* patches can be quantified numerically and spatially by random sets. *Carex* volume and coverage can be the corresponding measurements on the ground, by which the covering function of random sets can be quantified and interpreted adequately.

7. Changes of wetland inundation extents in PLNNR

Each year, during late spring and early summer (May and June), the five upstream rivers of Poyang Lake are flooded due to concentrated rainfall. As a consequence, the water level of Poyang Lake swells up without reaching the peak yet. From July until early September, the water supply from the five tributaries decreases, whereas the Yangtze River has its highest level, which causes the water level to rise to its peak, and the nine lakes in PLNNR merge into one single water body. From late September until November, the water level recedes towards its lowest level after the recession of the Yangtze River. Lakes in PLNNR are disconnected by then and have different water levels throughout the winter (from December to February).

In Chapter 5, annual variation of wetland inundation extents in PLNNR was monitored from a series of images in 2004 and 2009. Transition zones between wetland vegetation and open water were identified and inundation areas were modeled as random sets. Transition zones have the largest area and occupy almost half of the PLNNR immediately after flood recession in early October. In other months when the water

level is stable, the proportions of transition zone are still approximately twenty percent and can hardly be ignored in a classification. Temporal profiles of p-level sets and two indices SD and CV, showed the seasonal dynamics of inundated areas and related extensional uncertainty. The size of the uncertain areas decreases from April to July or August, it jumps to its highest value in early October, and decreases afterwards.

A clear distinction exists between the narrow transition zones separating rivers and vegetation and broad transition zones between lakes and vegetation. The lake bank of Dachahu and Banghu has a wide and vague transition zone, followed by Dahuchi, whereas Shahu has a relatively narrow transition zone to surrounding wetland vegetation.

The wetland inundation in PLNNR experienced great changes between 2004 and 2009. Drought with different severities occurred in nearly half of the PLNNR, whereas only a small part of the areas was covered by water for a longer time. The most severe and the largest areas of drought happened in Banghu. The permanent water covering areas shrink to the small corner in the southeast of Banghu, whereas the other places have less water covering days in 2009. Dahuchi experienced less modification and a decreasing WCD mainly on the west bank and south bank. In contrast, Shahu has larger areas covered by water during the whole year 2009 and its lake bank becomes more crisp. The wetland in Dachahu appears more fragmented in 2009. Some parts have less WCD and some have more, which makes the previous continuous shore break into patches.

8. Spatial-temporal dynamics in the Poyang Lake

The size of the lake area in Poyang Lake is not well known. In most of studies on Poyang Lake, it is only mentioned briefly as "no more than 1000km^2 in low-water seasons to approximately 4000km^2 in high-water seasons" [79, 128, 153, 184]. Studies which concern the lake size change either adopt a different region boundary for calculating the lake size as there is no unique and official boundary definition of Poyang Lake [38], or their conclusions are derived from images at a different temporal scale such as images within a year e.g. 2000 [73]. In order to address the inherent uncertainty in these estimates, a random spread process model was implemented in Chapter 6 to monitor spatial-temporal changes in the Poyang Lake from more than six hundred multitemporal MODIS images during last eleven years.

Substantial seasonal dynamics of inundation extents was found for the monthly random sets. In the peak flooding months (July and August), the core sets have the largest extents in the center of the lake and the median sets cover the majority of the lake. During the dry season

(November to February), small water areas distribute mainly around the water channel. The spatial pattern of water distribution are similar within these two seasons respectively. Less similar patterns were shown around the dry season months and flooding peak months in March, April and June. May and Sept had relatively low correlation with other months, which indicated either a dramatic flooding or recession process occur.

A subtle interannual change trend was found for the yearly random sets. Yearly random sets gave detailed information on the spatial distributions of inundation frequency and showed a shrinking trend from 2000 to 2009. Covering functions before 2003 have higher values that cover a larger area than after 2003, i.e. the frequently flooded area is larger before 2003. The spatial distributions of flooding frequency were similar in 2003 and 2005, and also those in 2004, 2006 until 2009 were similar. The correlation coefficients of any two years between 2004 and 2009 had high values. This high similarity shows a generally regular spatial pattern of flooding frequency among years after 2003 and thus coincides with the result that year 2003 is a separator in the eleven years. Besides, 2010 has the lowest correlations with other years, which confirms that an extreme flooding event happened that year.

7.3 General conclusion

Random sets provide a general framework for describing uncertainties in spatial and temporal modeling of natural landscape from remote sensing images. Extensional uncertainty of segmented image objects can be represented by random sets, with the aim of summarizing the essential properties of their uncertain boundaries by statistic characteristics of random sets and derived indices. Random sets identify uncertain extents of inundated wetlands in a nondeterministic manner and its derived dynamic modeling resolves the deficiencies of exploring water regime pattern by deterministic methods. By modeling wetland inundation as a random spread process, periodicity, trend and random components of the inundation dynamics can be captured and characterized. In sum, we conclude that the random set model enriches spatial and spatial-temporal modeling of phenomena which are uncertain in space and dynamic in time.

7.4 Recommendations

This study proposed random sets as an uncertainty modeling tool, and shown its capabilities in several applications. Many aspects of the

method has not been fully addressed yet, and further investigations are needed.

Firstly, the distribution of a random set is estimated from its realizations. In this research, these realizations were obtained during the image object extraction, i.e. by slicing or segmentation, to model the uncertainties in image interpretation and object identification. They were also derived from a group of images that represents the status of underlying phenomena in a time interval, to model the uncertainties in spatial-temporal change and tracking. Actually, the ways of getting realizations are not limited by the above mentioned methods and many others are possible. For example, in image fusion, different realizations can be from images acquired from different sensors with various resolutions.

Secondly, the random set method, until now, has been designed for analyzing individual objects, so whether one object is overlapped with another does not influence the uncertainty quantification procedure for each object. To deal with the overlapping situation, we may turn to studying the topological relationship of uncertain objects, which is out of the scope of this study. As an alternative, we can also investigate how random set methods can be combined with other object extraction methods which consider space partition such as split-and-merge segmentation. In those cases, other types of uncertainties such as existential uncertainty will also appear, and more classes besides object and background are then to be determined.

Thirdly, random sets identified uncertain extents of inundated wetlands in a nondeterministic manner and its derived dynamic modeling resolved the deficiencies of exploring water regime pattern by deterministic methods. This random set model is worthy of further investigations into other wetlands or natural landscapes with uncertain extents. The spectral information about inundated areas was derived from a multispectral index NDVI, in order to apply the random set model to a single input band. To implement the random set model on multiple bands, we propose to first run the random set model on each of them separately and then combine the resulting random sets by logical operators. Another option is to first combine all input bands by logical operators and then model the collection of results by random sets. At this stage the best way to proceed is unknown and further exploration is required.

Fourthly, the random spread model developed in the present study did not contain a hierarchical structure. If considering interannual variation and monthly variation within the subcomponents in each year, we may construct a hierarchical structure of random sets in a single RSP, which could be possible to combine several temporal scales within one particular system. Calculating the median set is one possible way of linking random sets at different scales, i.e. converting a random set in

a lower level to a realization of the random set in a higher level. Using other p-level sets is also possible but is left for future study.

Lastly, the random spread model and random sets serves as descriptive statistic tools and their predicting functions have not been explored yet in this thesis. This suggests the direction of further research: that a fully developed random set model should able to do inference statistic. It should provide answers to questions like: what is the most likely uncertain extents of the object in the future time and how likely would that be? Other statistic approaches like Bayesian methods maybe helpful.

Bibliography

[1] O. Ahlqvist, J. H. D. Keukelaar, and K. Oukbir. Rough classification and accuracy assessment. *International Journal of Geographic Information Science*, 14:475–496, 2000.

[2] O. Ahlqvist, J. H. D. Keukelaar, and K. Oukbir. Rough and fuzzy geographical data integration. *International Journal of Geographic Information Science*, 17:223–234, 2003.

[3] D. Alvarez. On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43(3):241–267, 2006.

[4] C. Arnot, P. Fisher, R. Wadsworth, and J. Wellens. Landscape metrics with ecotones: pattern under uncertainty. *Landscape Ecology*, 19:181–195, 2004.

[5] P. M. Atkinson and G. M. Foody. Uncertainty in remote sensing and GIS: Fundamentals. In G. M. Foody and P. M. Atkinson, editors, *Uncertainty in Remote Sensing and GIS*, chapter 1, pages 1–18. John Wiley & Sons, 2002.

[6] A. Baddeley and I. Molchanov. Averaging of random sets based on their distance functions. *Journal of Mathematical Imaging and Vision*, 8:79–92, 1998.

[7] O. Barndorff-Nielsen, W. Kendall, and M. Lieshout, editors. *Stochastic Geometry: Likelihood and Computation*. Chapman & Hall/CRC, 1999.

[8] M. Baudin. Multidimensional point processes and random closed sets. *Journal of Applied Probability*, 21:173–178, 1984.

[9] L. Bejaoui, F. Pinet, Y. Bedard, and M. Schneider. Qualified topological relations between spatial objects with possible vague shape. *International Journal of Geographical Information Science*, 23(7): 877–921, 2009.

[10] K. Beutner III, G. Prasad, E. Fletcher, C. DeCarli, and O. Carmichael. Estimating uncertainty in brain region delineations. In D. M. K.

Prince, J.L. Pham, editor, *Proceedings of Information Processing in Medical Imaging*, volume 5636 of *Lecture Notes in Computer Science*, pages 479–490, 2009.

[11] T. Bittner and J. G. Stell. Rough sets in approximate spatial reasoning. In *Second International Conference on Rough Sets and Current Trends in Computing*, volume Lecture Notes in Computer Science (2005), pages 445–453. Berlin: Springer-Verlag, 2000.

[12] J. Bolton and P. Gader. Random set framework for context-based classification with hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 47:3810–3821, 2009.

[13] B. A. Bradley, R. W. Jacob, J. F. Hermance, and J. F. Mustard. A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data. *Remote Sensing of Environment*, 106:137–145, 2007.

[14] D. G. Brown. Classification and boundary vagueness in mapping presettlement forest types. *International Journal of Geographical Information Science*, 12:105–129, 1998.

[15] R. G. Bryant and M. P. Rainey. Investigation of flood inundation on playas within the Zone of Chotts, using a time-series of AVHRR. *Remote Sensing of Environment*, 82:360–375, 2002.

[16] P. Burrough and A. Frank, editors. *Geographical Objects with Indeterminate Boundaries*. Taylor & Francis, 1996.

[17] M. T. Casanova and M. A. Brock. How do depth, duration and frequency of flooding influence the establishment of wetland plant communities? *Plant Ecology*, 147:237–250, 2000.

[18] G. Chander, D. L. Helder, B. L. Markham, J. D. Dewald, E. Kaita, K. J. Thome, E. Micijevic, and T. A. Ruggles. Landsat-5 TM reflective-band absolute radiometric calibration. *IEEE Transactions on Geoscience and Remote Sensing*, 42(12):2747–2760, 2004.

[19] H. Chen and Y. Zhao. Evaluating the environmental flows of China's Wolonghu wetland and land use changes using a hydrological model, a water balance model, and remote sensing. *Ecological Modelling*, 222:253–260, 2010.

[20] T. Cheng and M. Molenaar. Objects with fuzzy spatial extent. *Photogrammetric Engineering & Remote Sensing*, 65:797–801, 1999.

[21] J. Clark, S. Carpenter, M. Barber, S. Collins, A. Dodson, J. Foley, D. Lodge, M. Pascual, R. Pielke, Jr., W. Pizer, C. Pringle, W. Reid, K. Rose, O. Sala, W. Schlesinger, D. Wall, and D. Wear. Ecological forecasts: An emerging imperative. *Science*, 293:657–660, 2001.

[22] E. Clementini and P. Di Felice. An algebraic model for spatial objects with indeterminate boundaries. In P. Burrough and A. Frank, editors, *Geographic objects with indeterminate boundaries*, pages

171–187. Taylor&Francis, 1996.

[23] A. Cohn and N. Gotts. *Geographic objects with indeterminate boundaries*, chapter The 'egg-yolk' representation of regions with indeterminate boundaries, pages 171–187. Taylor&Francis, 1996.

[24] N. Cressie. *Statistics for spatial data*, chapter 9: modeling objects, pages 725–803. Wiley-Interscience, 1993.

[25] N. Cressie and F. L. Hulting. A spatial statistical analysis of tumor growth. *Journal of the American Statistical Association*, 87:272–283, 1992.

[26] M. Crosetto and S. Tarantola. Uncertainty and sensitivity analysis: tools for gisbased model implementation. *International Journal of Geographical Information Science*, 15:415–437, 2001.

[27] M. Crosetto, J. A. M. Ruiz, and B. Crippa. Uncertainty propagation in models driven by remotely sensed data. *Remote Sensing of Environment*, 76:373–385, 2001.

[28] X. Dai, Z. Guo, L. Zhang, and D. Li. Spatio-temporal exploratory analysis of urban surface temperature field in shanghai, china. *Stochastic Environmental Research and Risk Assessment*, 24:247–257, 2010.

[29] A. Davranche, G. Lefebvre, and B. Poulin. Wetland monitoring using classification trees and SPOT-5 seasonal time series. *Remote Sensing of Environment*, 114:552–562, 2010.

[30] S. de Bruin. Modelling positional uncertainty of line features by accouting for stochastic deviations from straght line segments. *Transactions in GIS*, 12:165–177, 2008.

[31] P. Diggle. Binary mosaics and the spatial pattern of heather. *Biometrics*, 31:531–539, 1981.

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

[33] X. Ding and X. Li. Monitoring of the water-area variations of Lake Dongting in China with ENVISAT ASAR images. *International Journal of Applied Earth Observation and Geoinformation*, 13: 894–901, 2011.

[34] D. Dubois and H. Prade. Random sets and fuzzy interval analysis. *Fuzzy Sets and Systems*, 42:87–101, 1991.

[35] J. L. Dungan. Toward a comprehensive view of uncertainty in remote sensing analysis. In G. M. Foody and P. M. Atkinson, editors, *Uncertainty in Remote Sensing and GIS*, chapter 3, pages 25–35. John Wiley & Sons, 2002.

[36] I. Epifanio and G. Ayala. A random set view of texture classification.

IEEE Transactions on Image Processing, 11:859–867, 2002.

[37] I. Epifanio and P. Soille. Morphological texture features for unsupervised and supervised segmentations of natural landscapes. *IEEE Transactions on Geoscience and Remote Sensing*, 45:1074–1083, 2007.

[38] L. Feng, C. Hu, X. Chen, R. Li, L. Tian, and B. Murch. MODIS observations of the bottom topography and its inter-annual variability of Poyang Lake. *Remote Sensing of Environment*, 115:2729–2741, 2011.

[39] L. Feng, C. Hu, X. Chen, L. Tian, X. Cai, and W. Gan. Assessment of inundation changes of Poyang Lake using MODIS observations between 2000 and 2010. submitted for publication.

[40] P. Fisher. Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113:7–18, 2000.

[41] P. Fisher. Remote sensing of land cover classes as type 2 fuzzy sets. *Remote Sensing of Environment*, 114:309–321, 2010.

[42] P. Fisher, T. Cheng, and J. Wood. Higher order vagueness in geographical information: Empirical geographical population of type n fuzzy sets. *Geoinformatica*, 11(3):311–330, 2007.

[43] G. M. Foody. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80:185–201, 2002.

[44] G. M. Foody and P. M. Atkinson, editors. *Uncertainty in remote sensing and GIS*. John Wiley and Sons, 2002.

[45] G. M. Foody and P. M. Atkinson. Current status of uncertainty issues in remote sensing and GIS. In G. M. Foody and P. M. Atkinson, editors, *Uncertainty in Remote Sensing and GIS*, chapter 17, pages 287–302. John Wiley & Sons, 2002.

[46] M.-J. Fortin and G. Edwards. Delineation and analysis of vegetation boundaries. In C. Hunsaker, M. Goodchild, and M. Friedl, editors, *Spatial uncertainty in ecology*. Springer, 2001.

[47] M.-J. Fortin, R. Olson, S. Ferson, L. Iverson, C. Hunsaker, G. Edwards, D. Levine, K. Butera, and V. Klemas. Issues related to the detection of boundaries. *Landscape Ecology*, 15:453–466, 2000.

[48] N. Friel and I. Molchanov. A new thresholding technique based on random sets. *Pattern Recognition*, 32:1507–1517, 1999.

[49] M. Gahegan and M. Ehlers. A framework for the modelling of uncertainty between remote sensing and geographic information systems. *ISPRS Journal of Photogrammetry &. Remote Sensing*, 55: 176–188, 2000.

[50] A. Gallego and A. Simo. Random closed set models: estimating and simulating binary images. *Image Analysis & Stereology*, 22:

133–145, 2003.

[51] F. Gao, J. Morisette, R. Wolfe, G. Ederer, J. Pedelty, E. Masuoka, R. Myneni, B. Tan, and J. Nightingale. An algorithm to produce temporally and spatially continuous MODIS-LAI time series. *IEEE Geoscience and Remote Sensing Letters*, 5:60–64, 2008.

[52] M. S. Gilmore, E. H. Wilson, N. Barrett, D. L. Civco, S. Prisloe, J. D. Hurd, and C. Chadwick. Integrating multi-temporal spectral and structural information to map wetland vegetation in a lower Connecticut River tidal marsh. *Remote Sensing of Environment*, 112:4048–4060, 2008.

[53] M. Glemser and D. Fritsch. Data uncertainty in a hybrid gis. In *International Archive of Photogrammetry and Remote sensing*, volume 32, pages 180–187, 1998.

[54] M. Goodchild and S. Gopal, editors. *Accuracy of Spatial Databases*. London, Taylor&Francis, 1989.

[55] M. F. Goodchild and G. J. Hunter. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11:299–306, 1997.

[56] I. Goodman. Fuzzy sets as equivalence classes of random sets. In R. Yager, editor, *Fuzzy Sets and Possibility Theory: Recent Developments*. Permagon Press, 1982.

[57] I. Goodman and H. Nguyen. *Uncertainty Models for Knowledge Based Systems*. North-Holland, 1985.

[58] I. Goodman, R. Mahler, and H. Nguyen. *Mathematics of Data Fusion*, volume 37 of *SeriesB: mathematical and statistical methods*. Kluwer Academic Publishers, 1997.

[59] J. Goutsias, R. Mahler, and H. Nguyen, editors. *Random Sets Theory and Applications*. The IMA Volumes in Mathematics and its applications. Springer, 1997.

[60] S. Guptill and J. Morrison, editors. *Elements of spatial data quality*. Elsevier Science Ltd, 1995.

[61] R. B. Gurung, F. J. Breidt, A. Dutin, and S. M. Ogle. Predicting enhanced vegetation index (EVI) curves for ecosystem modeling applications. *Remote Sensing of Environment*, 113:2186–2193, 2009.

[62] J. Hall and J. Lawry. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. *Reliability Engineering & System Safety*, 85:89–101, 2004.

[63] A. Hansen and F. di Castri, editors. *Landscape Boundaries: Consequences for Biotc Diversity and Ecological Flows*. Springer-Verlag, 1992.

[64] A. Harris and R. G. Bryant. A multi-scale remote sensing approach for monitoring northern peatland hydrology: Present possibilities and future challenges. *Journal of Environmental Management*, 90: 2178–2188, 2009.

[65] K. Hestir, H. Nguyen, and G. Rogers. A random set formalism for evidential reasoning. In I. Goodman, M. Gupta, H. Nguyen, and G. Roger, editors, *Conditional Logic in Expert Systems*, pages 309–344. North-Holland, 1991.

[66] G. Heuvelink. *Error propagation in environmental modelling with GIS*. London, Taylor&Francis, 1998.

[67] G. Heuvelink, P. Burrough, and A. Stein. Developments in analysis of spatial uncertainty since 1989. In P. Fisher, editor, *Classics from IJGIS : 20 years of the International Journal of Geographical Information Science and Systems*, pages 91–95. Boca Raton : CRC, 2006.

[68] G. Heuvelink, J. D. Brown, and E. Van Loon. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, 21(5):497–513, 2007.

[69] R. A. Hill, K. Granica, G. M. Smith, and M. Schardt. Representation of an alpine treeline ecotone in SPOT 5 HRG data. *Remote Sensing of Environment*, 110:458–467, 2007.

[70] J. N. Hird and G. McDermid. Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sensing of Environment*, 113:248–258, 2009.

[71] C. Hu. A novel ocean color index to detect floating algae in the global oceans. *Remote Sensing of Environment*, 113:2118–2129, 2009.

[72] Q. Hu, S. Feng, H. Guo, G. Chen, and T. Jiang. Interactions of the Yangtze river flow and hydrologic processes of the Poyang Lake, China. *Journal of Hydrology*, 347:90–100, 2007.

[73] F. Hui, B. Xu, H. Huang, Q. Yu, and G. P. Modelling spatialtemporal change of Poyang Lake using multitemporal Landsat imagery. *International Journal of Remote Sensing*, 29(20):5767–5784, 2008.

[74] G. Jacquez, S. Maruca, and M.-J. Fortin. From fields to objects: a review of geographic boundary analysis. *Journal of Geographical Systems*, 2:221–241, 2000.

[75] S. K. Jain, R. D. Singh, J. M. K., and A. K. Lohani. Delineation of flood-prone areas using remote sensing techniques. *Water Resources Management*, 19:333–347, 2005.

[76] H. Jankowski and L. I. Stanberry. Expectations of random sets and their boundaries using oriented distance functions. *Journal of*

Mathematical Imaging and Vision, 36:291–303, 2010.

[77] J. Jensen, S. Narumalani, O. Weatherbee, and H. Mackay. Measurement of seasonal and yearly cattail and waterlily changes using multidate SPOT panchromatic data. *Photogrammetric Engineering & Remote Sensing*, 59(4):519Ű525, 1993.

[78] B. Jiang, S. Liang, J. Wang, and Z. Xiao. Modeling MODIS LAI time series using three statistical methods. *Remote Sensing of Environment*, 114:1432–1444, 2010.

[79] L. Jiang, K. M. Bergen, D. G. Brown, T. Zhao, Q. Tian, and S. Qi. Land-cover change and vulnerability to flooding near Poyang Lake, Jiangxi Province, China. *Photogrammetric Engineering & Remote Sensing*, 74:775–786, 2008.

[80] D. Karssenberg and K. de Jong. Dynamic environmental modelling in GIS: 2. modelling error propagation. *International Journal of Geographical Information Science*, 19:623–637, 2005.

[81] D. G. Kendall. Foundations of a theory of random sets. In E. F. Harding and D. G. Kendall, editors, *Stochastic Geometry*, pages 322–376. Wiley, 1974.

[82] H. T. Kiiveri. Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, 11:33–52, 1997.

[83] K. Kindscher, A. Fraser, M. Jakubauskas, and D. Debinski. Identifying wetland meadows in Grand Teton National Park using remote sensing and average wetland values. *Wetlands Ecology and Management*, 5:265–273, 1998.

[84] G. Klir. *Uncertainty and information-foundations of generalized information theory*. John Wiley & Sons, 2006.

[85] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy logic: Theory and Application*. New Jersey: Prentice-Hall, 1995.

[86] J. Kovitz and G. Christakos. Assimilation of fuzzy data by the bme method. *Stochastic Environmental Research and Risk Assessment*, 18:79–90, 2004.

[87] R. Kruse, E. Schwencke, and J. Heinsohn. *Uncertainty and Vagueness in Knowledge-Based Systems*. Springer-Verlag, 1991.

[88] J. Lein. Applying evidential reasoning methods to agricultural land cover classification. *International Journal of Remote Sensing*, 24(21):4161–4180, 2003.

[89] K. Liu and W. Shi. Computing the fuzzy topological relations of spatial objects based on induced fuzzy topology. *International Journal of Geographical Information Science*, 20:857–883, 2006.

[90] K. Liu and W. Shi. Quantitative fuzzy topological relations of

spatial objects by induced fuzzy topology. *International Journal of Applied Earth Observation and Geoinformation*, 11(1):38–45, 2009.

[91] I. Lizarazo and P. Elsner. Fuzzy segmentation for object-based image classification. *International Journal of Remote Sensing*, 30: 1643–1649, 2009.

[92] K. Lowell. Fiat boundaries: some implications for interpretation, decision-support, and multi-temporal analysis. *Environmental and Ecological Statistics*, 15:369–383, 2008.

[93] S. Lu, B. Wu, N. Yan, and H. Wang. Water body mapping method with HJ-1A/B satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, doi: 10.1016/j.jag.2010.09.006, 2010.

[94] A. Lucieer and A. Stein. Existential uncertainty of spatial objects segmented from satellite sensor imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40:2518–2521, 2002.

[95] J. Lund and M. Rudemo. Models for point processes observed with noise. *Biometrika*, 87:235–249, 2000.

[96] R. L. Lunetta, F. K. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment*, 105:142–154, 2006.

[97] J. Lyon. *Wetland landscape characterization: techniques and applications for GIS, mapping, remote sensing and image analysis*. Ann Arbor Press, 2001.

[98] R. Mahler. Combining ambiguous evidence with respect to ambiguous a priori knowledge, II: Fuzzy logic. *Fuzzy Sets and Systems*, 75:319–354, 1995.

[99] R. Mahler. Combining ambiguous evidence with respect to ambiguous a priori knowledge, I: Boolean logic. *IEEE Transactions on Systems, Man and Cybernetics - Part A*, 26:27–41, 1996.

[100] R. Mahler. Random sets: Unification and computation for information fusion - a retrospective assessment. In *The 7th International Conference on Information Fusion*, volume 1, pages 1–20, 2004.

[101] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House,INC, 2007.

[102] J.-M. Martinez and T. Toan. Mapping of flood dynamics and spatial distribution of vegetation in the Amazon floodplain using multitemporal SAR data. *Remote Sensing of Environment*, 108:209–223, 2007.

[103] P. M. Mather. Land cover classification revisited. In P. M. Tate and N. J. Tate, editors, *Adances in remote sensing and GIS analysis*, pages 7–16. Chichester: Wiley, 1999.

[104] G. Matheron. *Random Sets and Integral Geometry*. New York: Wiley, 1975.

[105] S. K. Mcfeeters. The use of normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17:1425–1432, 1996.

[106] H. J. Miller. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94:284–289, 2004.

[107] A. Moghaddamzadeh and N. Bourbakis. A fuzzy region growing approach for segmentation of color images. *Pattern Recognition*, 30:867–881, 1997.

[108] I. Molchanov. Averaging of random sets and binary images. *CWI Quarterly*, 11(4):371–384, 1998.

[109] I. Molchanov. Random closed sets: results and problems. In O. Barndorff-Nielsen, W. S. Kendall, and M. N. M. Van Lieshout, editors, *Stochastic geometry: likelihood and computation*, pages 286–331. Chapman and Hall/CRC, 1999.

[110] I. Molchanov. *Theory of Random Sets*. Springer, 2005.

[111] I. Molchanov. Random closed sets. In M. Bilodeau, F. Meyer, and M. Schmitt, editors, *Space, Structure, and Randomness*, volume LNS183, pages 135–149, 2005.

[112] M. Molenaar. *An introduction to the theory of spatial object modeling for GIS*. Research Monographs in GIS series. Taylor & Francis, 1998.

[113] H. Nguyen. *An Introduction to Random Sets*. Chapman & Hall/CRC, 2006.

[114] H. Nguyen and B. Wu. Random and fuzzy sets in coarse data analysis. *Computational Statistics & Data Analysis*, 51(1):70–85, 2006.

[115] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

[116] M. Nixon and A. Aguado. *Feature extraction and image processing*. Newnes, 2002.

[117] J. Nuñez-Garcia and O. Wolkenhauer. Random set system identification. *IEEE Transactions on Fuzzy Systems*, 10(3):287–296, 2002.

[118] M. Oberguggenberger and W. Fellin. Reliability bounds through random sets: Non-parametric methods and geotechnical applications. *Computers and Structures*, 86(10):1093–1101, 2008.

[119] M. Ortner, X. Descombes, and J. Zerubia. Building outline extraction from digital elevation models using marked point processes. *International Journal of Computer Vision*, 72(2):107–132, 2007.

[120] M. Ortner, X. Descombes, and J. Zerubia. A marked point process of rectangles and segments for automatic analysis of digital elevation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):105–119, 2008.

[121] Y. O. Ouma and R. Tateishi. A water index for rapid mapping of shoreline changes of five East African Rift Valley lakes: an empirical analysis using Landsat TM and ETM+ data. *International Journal of Remote Sensing*, 27(15):3153–3181, 2006.

[122] S. Ozesmi and M. Bauer. Satellite remote sensing of wetlands. *Wetlands Ecology and Management*, 10:381–402, 2002.

[123] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[124] Z. Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11:341–356, 1982.

[125] D. Pfoser and N. Tryfona. Capturing fuzziness and uncertainty of spatiotemporal objects. In A. Caplinskas and J. Eder, editors, *Advances in Databases and Information Systems*, Lecture Notes in Computer Science, pages 112–125. Springer-Verlag, 2001.

[126] M. Philibert, M.-J. Fortin, and F. Csillag. Spatial structure effects on the detection of patches boundaries using local operators. *Environmental and Ecological Statistics*, 15:447–467, 2008.

[127] B. Poulin, A. Davranche, and G. Lefebvre. Ecological assessment of *Phragmites australis* wetlands using multi-season SPOT-5 scenes. *Remote Sensing of Environment*, 114:1602–1609, 2010.

[128] S. Qi, D. G. Brown, Q. Tian, L. Jiang, T. Zhao, and K. M. Bergen. Inundation extent and flood frequency mapping using LANDSAT imagery and digital elevation models. *Giscience & Remote Sensing*, 46:101–127, 2009.

[129] P. Quinio and T. Matsuyama. Random closed sets: A unified approach to the representation of imprecision and uncertainty. In R. Kruse and P. Siegel, editors, *Symbolic and Quantitative Approaches to Uncertainty*. Springer-Verlag, 1991.

[130] K. J. Ranson, G. Sun, V. I. Kharuk, and K. Kovacs. Assessing tundra-taiga boundary with multi-sensor satellite data. *Remote Sensing of Environment*, 93:283–295, 2004.

[131] E. Renshaw, C. Comas, and J. Mateu. Analysis of forest thinning strategies through the development of space-time growth-interaction simulation models. *Stoch Environ Res Risk Assess*, 23 (3):275–288, 2009.

[132] E. Ricciardelli, F. Romano, and V. Cuomo. A technique for classifying uncertain mod35/myd35 pixels through meteosat second generation-spinning enhanced visible and infrared imager obser-

vations. *IEEE Transactions on Geoscience and Remote Sensing*, 40 (4):2137–2149, 2010.

[133] B. Ripley. Locally finite random sets: foundations for point process theory. *Annals of Probability*, 4:983–994, 1976.

[134] V. Robinson. A perspective on the fundamentals of fuzzy sets and their use in geographic information systems. *Transactions in GIS*, 7(1):3–30, 2003.

[135] J. W. Rouse, R. H. Hass, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains wih ERTS. In *Pages 309-317 in 3rd ERTS symposium*. U.S. Government Printing Office, Washington, DC., 1973.

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

[137] J. Russ. *The Image Processing Handbook*. Taylor & Francis, 2007.

[138] P. K. Saha and J. K. Udupa. Optimum image thresholding via class uncertainty and region homogeneity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:689–706, 2001.

[139] T. Sakamoto, N. Van Nguyen, A. Kotera, H. Ohno, N. Ishitsuka, and M. Yokozawa. Detecting temporal changes in the extent of annual flooding within the Cambodia and the Vietnamese Mekong Delta from MODIS time-series imagery. *Remote Sensing of Environment*, 109:295–313, 2007.

[140] C. Scheidt and J. Caers. representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, 41:397–419, 2008.

[141] J. Schiewe and M. Ehlers. Modeling uncertainty in high resolution remotely sensed scenes using a fuzzy logic approach. In T. Blachke, S. Lang, and G. Hay, editors, *Object-based Image Analysis Spatial Concepts for Knowledge-Driven Remote Sensing Application*, Lecture Notes in Geoinformation and Cartography, pages 755–768, 2008.

[142] T. Schmid, M. Koch, and J. Gumuzzio. Multisensor approach to determine changes of wetland characteristics in semiarid environments (Central Spain). *IEEE Transactions on Geoscience and Remote Sensing*, 43(11):2516–2525, 2005.

[143] M. Schneider. *Geographic Objects with Indeterminate Boundaries*, chapter Modeling Spatial Objects with Undetermined Boundaries Using the Realm/ROSE Approach, pages 141–152. Taylor and Francis, 1996.

[144] M. Schneider. Uncertainty management for spatial data in databases: fuzzy spatial data types. In *SSD' 99: Proceedings of the*

6th International Symposium on Advances in Spatial Databases, volume Lecture Notes in Computer Science(1651), pages 330–351. Berlin: Springer-Verlag, 1999.

[145] M. Schneider. Metric operations on fuzzy spatial objects in databases. In *8th ACM Symposium on Geographic Information Systems (ACM GIS)*, pages 21–26. New York: ACM Press, 2000.

[146] M. Schneider. Fuzzy topological predicates, their properties, and their integration into query languages. In *9th ACM Symposium on Geographic Information Systems (ACM GIS)*. New York: ACM Press, 2001.

[147] M. Schneider. Design and implementation of finite resolution crisp and fuzzy spatial objects. *Data & Knowledge Engineering*, 44:81–108, 2003.

[148] H. Schweiger and G. Peschl. Basic concepts and applications of random sets in geotechnical engineering. In D. Griffiths and G. Fenton, editors, *Probabilistic Methods in Geotechnical Engineering*, pages 113–126. Springer, 2007.

[149] J. Serra. The boolean model and random sets. *Computer Vision, Graphics and Image Processing*, 12:99–126, 1980.

[150] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, INC, 1982.

[151] J. Serra. The random spread model. In *8th International Symposium on Mathematical Morphology*, 2007.

[152] D. Shankman, B. D. Keim, and J. Song. Flood frequency in China's Poyang Lake region: Trends and teleconnections. *International Journal of Climatology*, 26:1255–1266, 2006.

[153] D. Shankman, L. Davis, and J. de Leeuw. River management, landuse change, and future flood risk in China's Poyang Lake region. *International Journal of River Basin Management*, 7:423–431, 2009.

[154] W. Shi. A generic statistical approach for modelling error of geometric features in gis. *International Journal of Geographical Information Science*, 12:131–143, 1998.

[155] W. Shi. *Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses*. Taylor & Francis, 2009.

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

[157] W. Shi and H. Wu. A probabilistic paradigm for handling uncertain objects in gis by randomized graph algebra. *Progress in Natural Science*, 13:648–657, 2003.

[158] W. Shi, P. F. Fisher, and M. F. Goodchild, editors. *Spatial data quality*. Taylor&Francis., 2002.

[159] W. Shi, K. Liu, and C. Huang. A fuzzy-topology-based area object extraction method. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1):147–154, 2010.

[160] J. Silván-Cárdenas, L. Wang, and F. Zhan. Representing geographical objects with scale-induced indeterminate boundaries: A neural network-based data model. *International Journal of Geographical Information Science*, 23:295–318, 2009.

[161] D. Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. In G. Dutton, editor, *Harvard Papers on Geographic Information Systems*, volume 6, pages 1–17. Addison-Wesley, 1978.

[162] S. V. Stehman and R. L. Czaplewski. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, 64:331–344, 1998.

[163] A. Stein. Modern developments in image mining. *Science in China Series E: Technological Sciences*, 51:13–25, 2008.

[164] A. Stein, F. van der Meer, and B. Gorte, editors. *Spatial statistics for remote sensing*. Springer, 1999.

[165] A. Stein, W. Shi, and W. Bijker, editors. *Quality aspects in spatial data mining*. CRC Press, 2008.

[166] A. Stein, P. Budde, and M. Z. Yifru. Stereology for multitemporal images with an application to flooding. In G. Navratil, editor, *Research Trends in Geographic Information Science*, pages 135–150. Springer-Verlag, 2009.

[167] A. Stein, N. Hamm, and Q. Ye. Handling uncertainties in image mining for remote sensing studies. *International Journal of Remote Sensing*, 30(20):5365–5382, 2009.

[168] D. Stoyan. Set-valued means of random particles. *Journal of Mathematical Imaging and Vision*, 7:111–121, 1997.

[169] D. Stoyan and H. Stoyan. *Fractals, random shapes and point fields*. Wiley, Chichester, 1994.

[170] D. Stoyan, W. Kendall, and J. Meche. *Stochastic Geometry and Its Applications*. John Wiley & Sons, second edition edition, 1995.

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

[172] O. Tobias and R. Seara. Image segmentation by histogram thresholding using fuzzy sets. *IEEE Transactions on Image Processing*, 11(12):1457–1465, 2002.

[173] F. Tonon, A. Bernardini, and A. Mammino. Reliability analysis of rock mass response by means of random set theory. *Reliability Engineering & System Safety*, 70:263–282, 2000.

[174] F. Tonon, A. Bernardini, and A. Mammino. Determination of parameters range in rock engineering by means of random set theory. *Reliability Engineering & System Safety*, 70:241–261, 2000.

[175] V. Uddameri and V. Honnungar. Combining rough sets and gis techniques to assess aquifer vulnerability characteristics in the semi-arid south texas. *Environmental Geology*, 51(6):931–939, 2007.

[176] R. Umamaheshwaran, W. Bijker, and A. Stein. Image mining for modeling forest fires from meteosat images. *IEEE Transactions on Geoscience and Remote Sensing*, 45:246–253, 2007.

[177] D. van de Vlag and A. Stein. Modeling dynamic beach objects using spatio - temporal ontologies. *Journal of Environmental Informatics*, 8(1):22–33, 2006.

[178] D. van de Vlag and A. Stein. Incorporating uncertainty via hierarchical classification using fuzzy decision trees. *IEEE Transactions on Geoscience and Remote Sensing*, 45:237–245, 2007.

[179] Vorob'ov. Random set models of fire spread. *Fire Technology*, 32 (2):137–173, 1996.

[180] F. Wang and G. Hall. Fuzzy representation of geographical boundaries in gis. *International Journal of Geographical Information System*, 10:573–590, 1996.

[181] S. Willard. *General Topology*. Alberta: Addison-Wesley, 1970.

[182] O. Wolkenhauer. Qualitative uncertainty models from random set theory. In X. Liu, P. Cohen, and M. Berthold, editors, *Advances in Intelligent Data Analysis Reasoning about Data*, Lecture Notes in Computer Science, pages 609–620. Springer-Verlag, 1997.

[183] C. Woodcock and S. Gopal. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, 14:153–172, 2000.

[184] G. Wu, J. de Leeuw, A. K. Skidmore, H. H. T. Prins, and Y. Liu. Concurrent monitoring of vessels and water turbidity enhances the strength of evidence in remotely sensed dredging impact assessment. *Water Research*, 41:3271–3280, 2007.

[185] J. Wu and G. Shao. On the accuracy of landscape pattern analysis using remote sensing data. *Landscape Ecology*, 23:505–511, 2008.

[186] Y. Wu and W. Ji. *Study on Jiangxi Poyang Lake national nature reserve*. Forest Publishing House, 2002.

[187] H. Q. Xu. Modification of normalised difference water index (NDWI)

to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033, 2006.

[188] G. J. Yang, P. Vounatsou, X. N. Zhou, J. Utzinger, and M. Tanner. A review of geographic information system and remote sensing with applications to the epidemiology and control of schistosomiasis in China. *Acta Tropica*, 96:117–129, 2005.

[189] Q. Ye, S. Kang, F. Chen, and J. Wang. Monitoring glacier variations on geladandong mountain, central tibetan plateau, from 1969 to 2002 using remote sensing and gis technologies. *Journal of Glaciology*, 52:537–545, 2006.

[190] Q. Ye, F. Chen, A. Stein, and Z. Zhong. Use of a multi - temporal grid method to analyze changes in glacier coverage in the Tibetan plateau. *Progress in Natural Science*, 19:861–872, 2009.

[191] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[192] L. A. Zadeh. Discussion: Probability theory and fuzzy logic are complementary rather than competitive. *Technometrics*, 37:271–276, 1995.

[193] B. F. Zhan. Approximate analysis of topological relations between geographic regions with indeterminate boundaries. *Soft Computing*, 2:28–34, 1998.

[194] Q. Zhan, M. Molenaar, K. Tempfli, and W. Shi. Quality assessment for geo-spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, 26:2953–2974, 2005.

[195] J. Zhang and M. Goodchild. *Uncertianty in Geographical Information*. London, Taylor&Francis, 2002.

[196] X. Zhao, X. Chen, and A. Stein. Random sets model for spatial objects with uncertain boundaries. In *Proceedings of the 12th AGILE internationational conference on Geographic Information Science*, Hanover, Germany, 2009.

[197] X. Zhao, A. Stein, and X. Chen. Application of random sets to model uncertainties of natural entities extracted from remote sensing images. *Stochastic Environmental Research and Risk Assessment*, 24:713–723, 2010.

[198] X. Zhao, A. Stein, X. Chen, and X. Zhang. Quantification of extensional uncertainty of segmented image objects by random sets. *IEEE Transactions on Geoscience and Remote Sensing*, 49: 2548–2557, 2011a.

[199] X. Zhao, A. Stein, and X. Chen. Monitoring the dynamics of wetland inundation by random sets on multi-temporal images. *Remote Sensing of Environment*, 115:2390–2401, 2011b.

Author's Biography

Xi Zhao was born on the 29^{th} of April in Hubei province, China. From 2001 to 2005, she studied in the School of Resource and Environmental Sciences, Wuhan University, China, and obtained Bachelor degree on Geographical Information System there. From 2005 to 2007, she attended a joint master's program between Wuhan University and the International Institute for Geo-Information Science and Earth Observation (ITC) in the Netherlands. She received her M.Sc. degree with distinction in geo-information science and earth observation, with a specialization in environmental systems analysis and management. From 2007 onwards, she has undertaken the doctoral study with a joint PhD scholarship offered by ITC, the Netherlands, and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China. Her research interest is in uncertainty modeling, image mining and spatial-temporal change detection.

ITC dissertations

A complete list of ITC dissertations is online on the ITC website: www.itc.nl/research/phd/phd_graduates.aspx.

This dissertation has number 203.

Cover design: Xiang Zhang

The "word cloud" on the back page was generated from the thesis introduction, showing the relative frequency of word occurence.