A prototype auto-human support system for spatial analysis*

LI Lianfa** and WANG Jinfeng

(State Key Laboratory of Resources & Environmental Information System, Institute of Geographical Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China)

Received November 18, 2005; revised May 8, 2006

Abstract Spatial analysis is a multidisciplinary field that involves multiple influential factors, variation and uncertainty, and modeling of geospatial data is a complex procedure affected by spatial context, mechanism and assumptions. In order to make spatial modeling easier, some scholars have suggested a lot of knowledge from exploratory data analysis (EDA), specification of the model, fitness and diagnosis of the model, to interpretation of the model. Also an amount of software has improved some functionalities of spatial analysis, e.g. EDA by the dynamic link (GeoDa) and robust statistical calculation (R). However, there are few programs for spatial analysis that can automatically deal with unstructured declarative issues and uncertainty in machine modeling using the domain knowledge. Under this context, this paper suggests a prototype support system for spatial analysis that can automatically use experience and knowledge from the experts to deal with complexity and uncertainty in modeling. The knowledge base component, as the major contribution of the system, in support of the expert system shell, codes and stores declarative modeling knowledge, e.g. spatial context, mechanisms and prior knowledge to deal with declarative issues during the modeling procedure. With the open architecture, the system integrates functionalities of other components, e.g. GIS' visualization, DBMS, and robust calculation in an interactive environment. An application case of spatial sampling, design and implementation of spatial modeling with such a system is demonstrated.

Keywords: spatial analysis, automatic support system, modeling of spatial data, artificial inference for modeling.

Spatial analysis has received an extensive theoretical research. Scholars have done a lot of work, e.g. from Kriging's Geostatistics in the 1950s followed by Tobler's Thumb Rule of Geography in 1979, Cressie's classic book, *Statistics for Spatial Data*[1] and Anselin's monograph, *Spatial Economics*[2], to Haining's recent book, *Spatial Data Analysis: Theory and Practice*[3]. Considering complexity of spatial modeling influenced by background knowledge, spatial context and heterogeneous factors, some scholars[1—5] have introduced a lot of knowledge of spatial modeling from exploratory data analysis (EDA), specification of the model, fitness and diagnosis of the model, to re-specification and interpretation of the model. The theories and methods have established a good ground for spatial analysis. Simultaneously, there have been an increasing number of tools for spatial analysis such as WinGSLib[1], S Plus[6], SAGE[7,8], Crime Stat[2], GeoDa and R[3], just to mention a few. This is also exemplified by the growing contents of the software clearinghouse maintained by the U.S.-based Center for Spatially Integrated SocialScience (CSISS, www.csiss.com).

However, the theoretical studies of spatial analysis and the available softwares cannot still satisfy the demand for handling a great amount of spatiotemporal data being produced each day by a burgeoning amount of GISs strong in management and visualization in spatial data[9,10].

What accounts for the disconnection between data and methods is also partly responsible for the phenomenon of "rich-data-but-few-knowledge"[11]. The reasons are multilateral but the major reason is overspecialization, weak interoperation of softwares, and lack of popular understanding of the methods. Many softwares have been developed independently in different ways that has led to less interaction. Further, most programs have been developed for special uses within a certain domain (e.g. SpacSta for econometrics, WinGSLib for mining, GeoStatist- ic+ and SAGE for environmental modeling and Crime Stat for criminalistics) that leads to less universality and overspecialization. Consequently, while these programs

* Supported by the Mimstry of Science and Technology (Grant No. 2002AA135230), and National Natural Science Foundation of China (Grant Nos. 40471111, 70571076)

** To whom correspondence should be addressed. E-mail: lspatial@yahoo.com or lilf@lreis.ac.cn

1) Journal A. Reservoir modeling with GSLIB, 2000

2) NLA(Net Levine & Associates). CrimeStat II. Washington: the National Institute of Justice, 2002

3) Anselin L., Syabri I. and Kho Y. GeoDa: An introduction to spatial data analysis. Geographical Analysis 2006, in press

stand as crisp modeling software, much of knowledge about how to model the practical problem and find the solution lies dormant in scientific papers, modeling code and experts' heads[12].

Some software has done some work to decrease specialization and improve readiness-to-use and interoperation. For instance, Anselin's GeoData does very well in graphic user interface (GUI), exploratory spatial analysis with the dynamic link and interoperation. Another case is the open-source R environment whose robustness in programmable, customable and extensible functionality represents the state-of-the-art spatial statistical computation.

Nevertheless, these improvements have seldom concerned the functionality of unstructured artificial inference that can use the modeling experience and declarative knowledge from the proficient researchers and appliers to guide users to explore relationships among data, identify influential factors and variations, construct the model, diagnose and interpret the model. In fact, many spatial models involve elaborate know-how, i.e. spatial context, mechanisms and assumptions other than procedural reasoning. A typical case is modeling of spatial variation that is an iterative process involving manifold knowledge of different application domains, statistics and uncertainty etc. Without support of suitable helpful auto-human tools, users often need to spend a lot time to learn modeling knowledge before doing it well. Such a lengthy learning curve is time-consuming and not cost effective to applications of methods.

Under the context, this paper presents a prototype human-automated support system for spatial analysis. Compared with the previous softwares, this prototype's major contribution is the support of knowledge base functionality that can code and materialize background knowledge and expert experience of spatial modeling to guide users to model spatial problems. Also this paper suggests the integrative framework of such a system, a good start for us to build an open, automated, scalable and extensible support tool for spatial analysis. In the following parts, Section 1 briefly introduces techniques used and the system architecture, Section 2 describes the major components of the system, Section 3 presents the modeling procedure using the example of spatial regression, Section 4 demonstrates the modeling procedure with an application case of spatial sampling, and Section 5 makes a brief summary.

1 Techniques and system architecture

1.1 The knowledge base system

As an artificial intelligence tool, the expert system (ES) uses knowledge base to store declarative knowledge, including expertise and experience from domain experts that cannot be contained in procedural models, and uses the knowledge and relevant inferences to help solve complicated decision problems.

The expert system has been successfully applied to a wide range of domains, e.g. CRYSALIS for interpretation of the protein's structure, PUFF for diagnosing the lung disease, REACTOR for detecting the nuclear reactor's accident, PROSPECTIOR for interpreting mineral geological stuff and YES/MVS for monitoring the IBM MVS's operation system etc. In recent years, the expert system has been also used to solve some geospatial problems. Leung successfully developed a fuzzy expert system shell to help risk analysis of disasters[10], and Li developed the model base system with limited automated support for the regional resources' modeling[13]. Some RS imagery processing software, say ERDAS, has developed the knowledge base inference module to help information processing and interpretation of images[1)].

As the crucial part of the prototype system, the expert system component is responsible for use of the knowledge base to code expertise of spatial analysis modeling, and artificial inference to solve complex problems involving multiple spatial factors and uncertainties. The inherent artificial inference and explanation facilities of the expert system make easier EDA, selection of the fit model, diagnosis and interpretation of the model in modeling. The artificial inference functionality of the prototype system, as its major virtue, makes it different from the procedural software of spatial analysis, e.g. WinGSLib, SAGE, R and GeoDa.

In this study, the open-source expert system program, CLIPS is adapted for our own shell. CLIPS is the acronym of the C Language Integrated Production System and was developed by the NASA's Johnson Space Center to avoid the high cost and poor integration brought by adopting LISP[2)]. Since 1986,

1) ERDAS 2000. ERDAS user's manual

2) CLIPS, CLIPS reference manual: Advanced programming guide, Vol. 3, 2002

CLIPS has helped the delivery of expert system technology for a wide range of applications, including geosciences[14]. Portability, compatibility, extensibility, capabilities and low-cost of CLIPS make it an ideal option for developing the inbuilt expert system shell.

CLIPS is a forward-chaining language. In CLIPS, knowledge is represented as objects, frames, facts and rules, and reasoning is driven by facts. This expert system's structure is typical and consists of seven essential components.

1) User interface: The mechanism by which the user and the expert system communicate.

2) Fact-list: A global memory for data. For example, a user's demand for interpolation is represented in CLIPS syntax as a fact:

First, the fact template is defined by the keyword "deftemplate":

(**deftemplate** Demand_of_Simple_Model (multislot name) (slot duty) (slot a_domain) (slot xor_spatial_correlation)).

Then, the fact is specified using the template by the key word "deffacts":

(**deffacts** Demand_of_Simple_Model (name: Xin Li) (duty: interpolation) (a_domain: rainfall) (xor_spatial_correlation: YES)).

where the bold keyword, deftemplate is the definition of the fact's template of the demand, and the keyword, deffacts gives the fact's definition. Each fact consists of one or multiple properties (signified by the keyword, slot or multislot).

3) Knowledge base: The set of all the facts, frames, objects and IF-THEN rules.

For instance, consider the following rule of judging the global pattern of spatial points in the Moran's Coefficient (MC) model (No is the total number of the sample points, Δ (delta) is the allowed departure degree):

IF MC ± Δ is greater than -1/(No-1)

THEN the distribution of the spatial points is positively correlative and spatial statistics methodology suggested.

In the CLIPS syntax, this rule is defined by the keyword "defrule":

(**defrule** pattern_of_global_spatial_points (Sum?No&: (integerp?No)) (Delta?delta&: (numberp?delta)) (MoranCo?MC&: (numberp?MC)) (test (>?(MC + delta) -1/(?No-1))) (test (>?(MC-delta) -1/(?No-1)))

⇒ (assert (correlation spatial positive)) (assert (method_type spatial)).

The LHP (left-hand-part) before "⇒" is the pre-conditions required to be satisfied before the RHP (right-hand-part) can be executed. In this rule, the LHP regulates the type of the variables (No, delta and MC) and the condition (MC ± delta > -1/(No-1)) so that their spatial positive correlation can be asserted (assert (correlation...)) and the spatial statistical method is recommended (assert (method_type...)).

The knowledge bases are organized as modules and the ground for making artificial inference.

4) Inference engine: Make inferences by deciding which rules are satisfied by facts, prioritizes the satisfied rules and executes the rule with the highest priority.

5) Agenda: A prioritized list created by the inference engine of instances of rules whose patterns are satisfied by facts in the fact list.

6) Explanation facility: Explain the inference of the problem-solving process for users.

7) Knowledge acquisition facility: Be responsible for acquiring knowledge by manual or machine learning (ML) means. The ML methods include induction, analogue, and decision tree, etc.

1.2 Architecture and techniques for the open system

Besides the ES shell, the prototype system includes other modules which are organically integrated with the shell. From the software engineering's perspective, the whole system should be designed and developed on the open, scalable, extensible and automatic architecture. Fig. 1 briefly illustrates the open architecture of the system whose core consists of three parts, the model base, the model knowledge base and the engine environment.

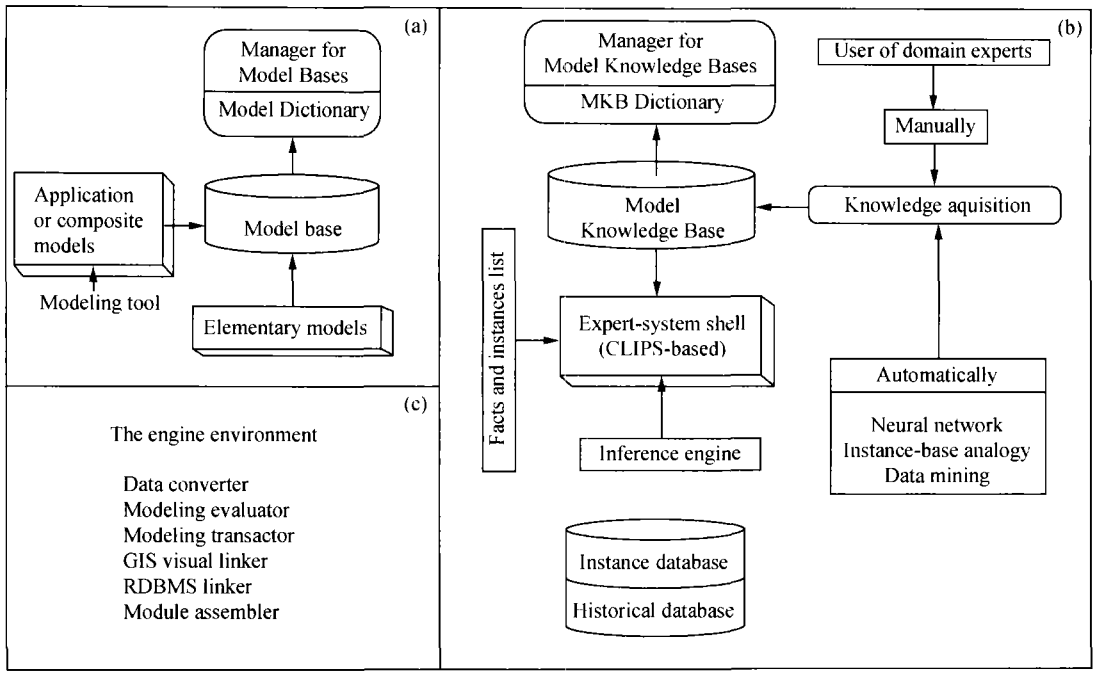


Fig. 1. The architecture of the prototype system.

In the architecture, the model base includes elementary and application models and the Manager for Model Bases (MMB) is responsible for management of the models through the model dictionary (Fig. 1 (a)); the model knowledge base contains the knowledge about models and modeling, and the Manager for Model Knowledge Bases (MMKB) is responsible for management of the knowledge base (Fig. 1(b)); the engine environment is responsible for providing data conversion, connection with GIS components or other database systems, transaction of modeling etc. (Fig. 1(c)). The environment's major role is to ensure coordination and interactions between modules.

The component technology, an industrial standard protocol of software development is the essential technology for development of the open system. Component-based programs are fully used to improve efficiency of developing the system from scratches[8,15]. For instance, component-compatible GIS such as MapObjects is used for visual exploration and representation of results; standard database management systems, e.g. Oracle, are used to manage the aspacial data and the dictionaries of the model and knowledge bases; the software with robust calculation functionality (e.g. MATLAB) serves as calculation support; the open codes of the ES software, CLIPS is adapted into an independent module with universal interfaces to other modules. The UML tool is used to help design component modules[16].

2 Major components of the system

2.1 Model base

The model base includes all models. The Manager for Model Bases (MMB) is responsible for creating, storing, retrieving, running and managing the model base (Fig. 1(a)). Models in the base are categorized into elementary and application types. In the model base, all models are packaged within the model library with each model contained in a binary component file (for Microsoft's COM, the file type is dll or ocx). As an effective way of retrieving the meta-data of models, the Model Dictionary (MD) manages the models as the MMB's core.

The Model Dictionary contains key items storing the meta-data for each model (Table 1). Among these items, "elementary or not" indicates the model's type (elementary or application), "Applicable domain" and "Assumption" suggest the application domain and pre-conditions respectively, "Location" identifies the address (path, file and component name etc.) where the model is stored, and "Index of interfaces" illustrates related interfaces of this model component. Table 1 also gives a specific example of the Moran Coefficient (MC) model.

Table 1. Items in the Model Dictionary and MC case

Items	Moran coefficient model
Model name	MC
Purpose	To measure spatial dependence at the global scale
Algorithm description	Measure the degree of spatial dependence
Applicable domain	Geospatial features with the spatial correlation
Applicable assumption	The objects distributed randomly in the space
Elementary or not	Yes
Input format	Text format: Matrix X and C; n
Output format	Text format: coefficient—MC
Index of rules	MC_App_Rule1, MC_App_Rule2, ...
Index of models	None
Location	COM module: lib\spbasics.dll
Index of interfaces	IMC Emodel(features and methods)
Author	Li X.
Version No.	1.0
Time	2004-01-10
Memo	Reference: Cressie 1991[1].

The Moran coefficient model measures spatial autocorrelation at global scale. In the model where n denotes the total number of spatial points, $MC \pm \Delta = -1/(n-1)$ (Δ is the allowed random departure degree) indicates a random map pattern, $MC \pm \Delta > -1/(n-1)$ indicates a positive correlation pattern and $MC \pm \Delta < -1/(n-1)$ indicates a negative correlation pattern. The MC value is an indicator for spatial dependence and for whether classical or spatial statistical models should be adopted.

Each model is developed in support of other models or the calculation software, say MatLab. This way decreases the repeatable work in a certain degree, thus saving time. Also each new model needs to register itself in the model dictionary and the model knowledge base needs to input relevant rules if necessary.

The duties of spatial analysis are categorized into several types, i.e. exploration of spatial correlation, modeling of spatial variation, identifying spatial outlier or hot spots and spatial sampling etc., as listed on the left side of the dotted line in Fig. 2. The right side of the line listed the referred methods. The model and knowledge bases are organized as modules according to the category in Fig. 2.

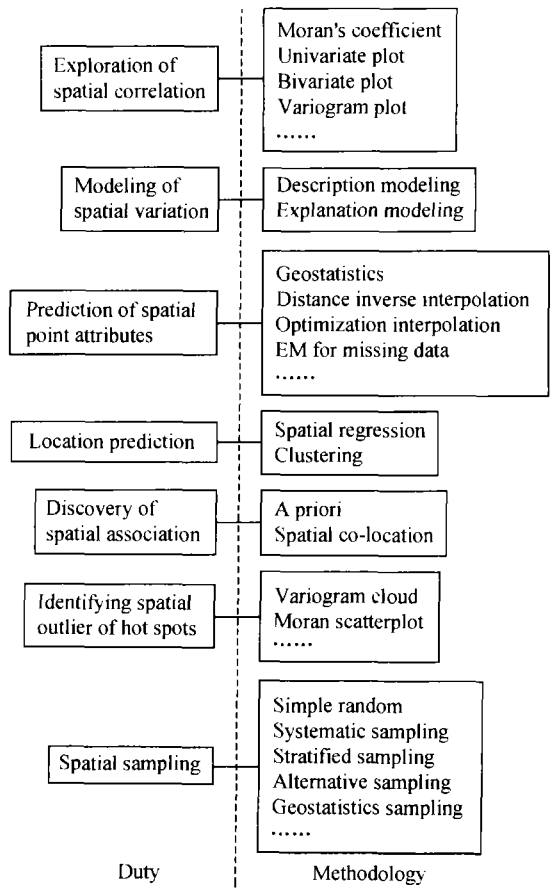


Fig. 2. The duties and methodologies of modeling

2.2 Model knowledge base

In the ES shell, knowledge is represented as frames, objects, facts or rules: meta-information about every model or algorithm is stored as objects whose content is similar to Table 1; factual knowledge is stored as facts or frames; conditional judgments and decision-making knowledge during modeling are often categorized, organized and implemented logically according to the classification or decision trees (Figs. 3—5). Reasoning is based upon IF-THEN rules and Rete's pattern match is the major inference algorithm. Models are executed as external functions (for structured models) or rule modules (for unstructured models involving declarative inference).

There are 4 types of knowledge with their different functionalities in modeling.

1) Knowledge for guidance of users in modeling of geospatial data

The type of knowledge is knowledge about modeling and models, i.e. prior research, relevant conclusions, experience, assumptions and even personal

opinions and is intended to guide users to analyze the problem, construct the model, estimate the parameters, examine the result and improve it. These rules are related with EDA and model specification in modeling. During the guidance, the system first collects some information from users, mines relations among variables from the data set using the induction algorithm, explores spatial dependence and specifies the model.

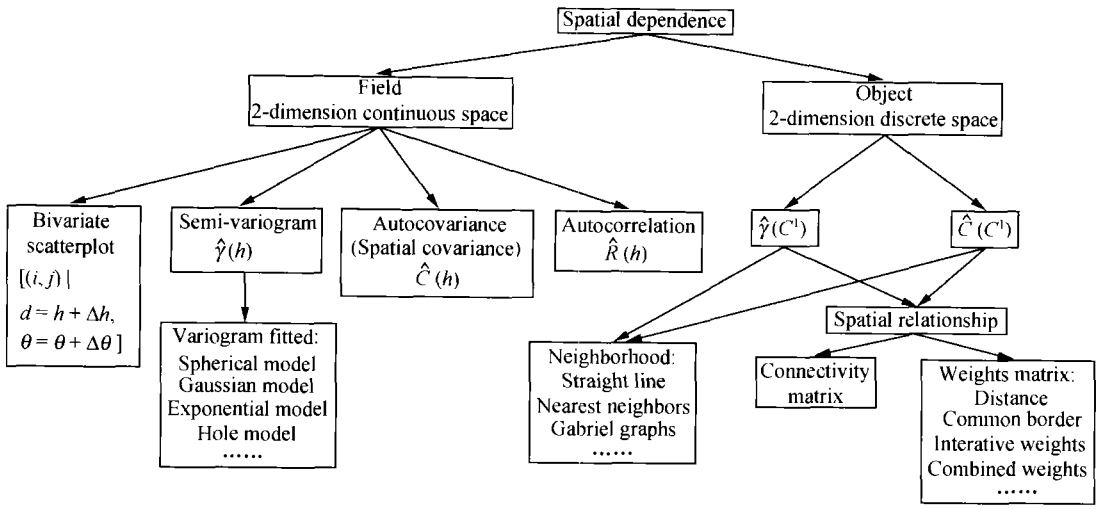


Fig. 3. The classification tree of spatial dependence analysis.

The modeling purposes are categorized into the types listed in Fig. 1. For instance, analysis of spatial dependence and specification of the model are carried out in two steps in many modeling duties. Fig. 3 makes a summary for analysis of spatial dependence according to relevant theories[7,17]. Based on the similar but more specific tree, the system can automatically guide users to locate the very algorithm for their dependence's modeling according to their goal and object under study. Model specification is to construct the model according to the user's purpose and the EDA's output.

2) Rules for automation of the model's judgment

Modeling of geospatial data often involves the model's condition judgment which can be easily handled by the IF-THEN rule and the rule's pattern match and inference in the knowledge system obviously improves the effect and automation level of modeling for this kind conditional judgment. Section 1.1 while introducing the concepts of rules in CLIPS illustrates the case of the judgment rule in the Moran's Coefficient.

3) Knowledge for diagnosis

The diagnosis is to assess the modeling effect. There are three types of diagnoses: fitness, substantive meaning and optimization. The fitness diagnosis measures how the model fits the data; the diagnosis of substantive meaning measures interpretability of the model, i.e. how well the model explains the phenomenon (e.g. in spatial regression of the infection, every predicator and its weight coefficient's practical meanings for the problem); the optimization diagnosis is for measuring advantages of the model over others. Fig. 4 summarizes the types of diagnosis and methods in a logical tree as the ground for constructing the diagnosis knowledge base.

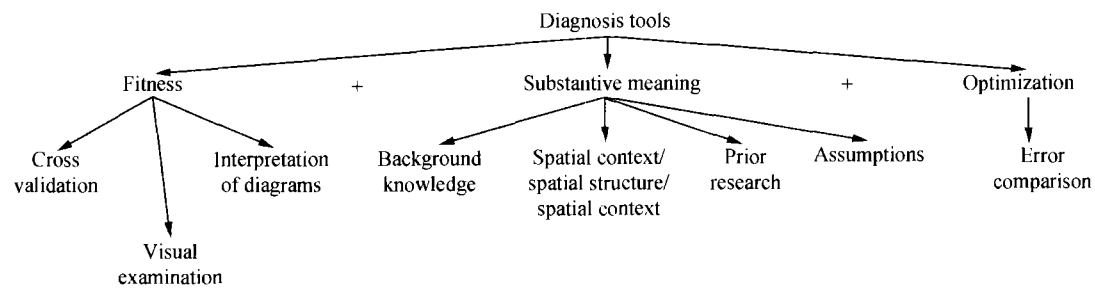


Fig. 4. Three components of the diagnosis tool.

4) Knowledge for uncertain and fuzzy inference

Uncertainty and fuzziness are two features of geospatial data that have the implication for spatial modeling. Although the original CLIPS doesnot support fuzzy and uncertain inference, the embedded shell adapted from CLIPS supports the inference indirectly to some extent by incorporating uncertainty or fuzzy information, that is adding the slot of Certainty Factor (CF) and fuzzy slots and executing related calculations by calling external functions[14].

Uncertainty of the data and models is propagated during the modeling procedure using the Certainty Factor slots and MIN-MAX calculation mechanism. A typical case is the Bayesian probability network that uses uncertainty represented as probability in the model's supportive components to induce the model's summary uncertainty. Fig. 5 is a probability-based uncertainty inference web for prediction of whether a site is a High-intensity Crime Area (HCA) according to contributions of regional, local, neighborhood and individual evidence. Different from spatial regression, this model is based on the probability inference. There are similar applications of this method in geosciences, say recognition of the mineral type at a site[14,18].

Fuzzy inference is implemented in the fuzzy slot. A fuzzy spatial modeling of the natural disaster's risk has been done in our initial system[15].

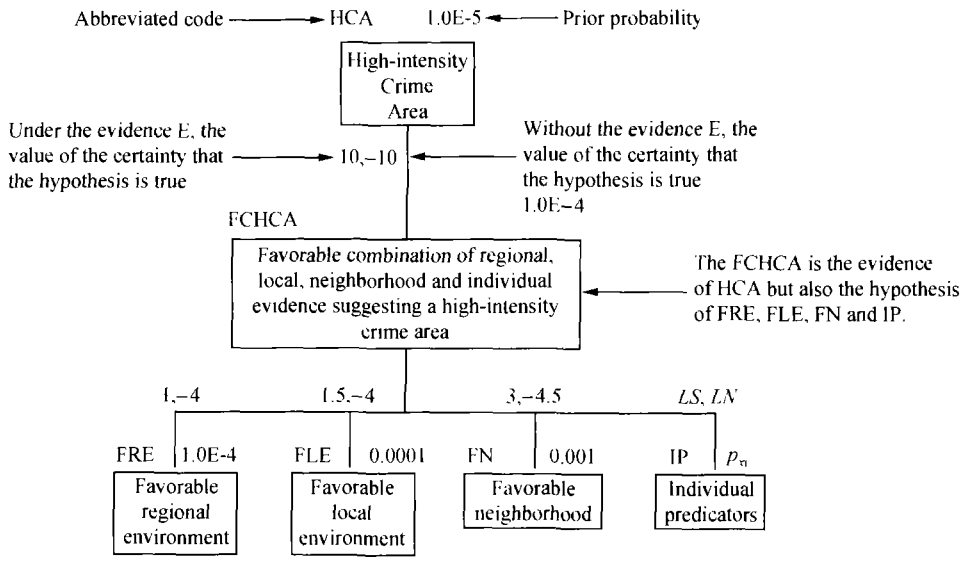
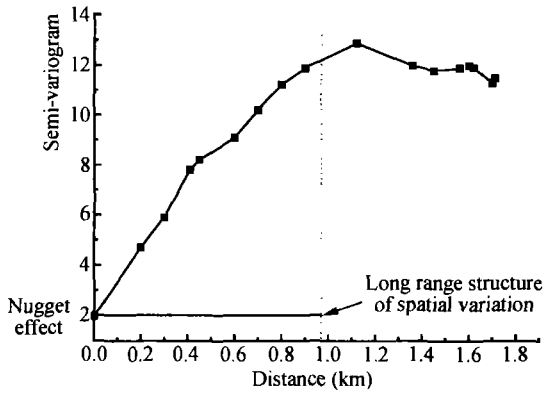


Fig. 5. A probability inference web for prediction of the attribute value (HCA) at a spatial site.

5) Knowledge for interpretation of the model

Spatial modeling is an interpretable issue that means that the output needs to account for a certain social, economic or physical spatial phenomenon. The background knowledge including spatial scales, spatial process and mechanism is important for interpretability of the model and it is necessary for the model to give an acceptable substantive explanation for the problem under study. Due to a great amount of unstructured knowledge in interpretation, the ES shell of the prototype system is a natural facility for interpretation of the output. In the system, there are three types of interpretations: the first is inference explanation that is provided by the ES shell's inherent facility, the second is interpretation of the numeric values that can be done with common IF-THEN rules incorporating prior knowledge and the third one is interpretation of the diagrams. As for the third type of interpretation, pattern recognition of the diagram is required to identify the critical, turning, extreme or unique points, the trend of the curve, the threshold and the dispersion degree of the diagram that are crucial for the interpretation. Techniques of the diagram's interpretation include recognition of key points, wavelet transform and clustering etc. Fig. 6 is a case of interpreting the reason for the long-range spatial variation of a certain mineral, Co according to its prior knowledge.

The boundary between the above types of knowledge is not strict but closely related. Knowledge for guidance often involves ones for diagnosis and interpretation.



The semi-variogram of the Co concentration (mg/kg) in a certain region. The relevant prior knowledge is that the long range structure of spatial variation in the concentration of metals is often affected by regional changes in geology, say rock type.

Interpretation rule of the plot

LHS: The spatial variation of the Co in the region is a long structure

⇒

RHS: In this region, the spatial variation of the Co closely relates with the geological changes.

Fig. 6. The interpretation rules of the variogram.

2.3 The engine environment

The engine environment contains a set of support modules, each having its COM objects and interfaces for its correspondent functionality. Table 2 lists the major modules.

Table 2. Major functionality modules in the engine environment

Name	Role description in the environment
Data converter	To convert data formats between standard DBMS, e.g. Oracle, GIS and other modules
GIS linker	To call GIS interfaces to manage geospatial data for EDA and representing results.
MMB	To manage and maintain the model bases through the Model Dictionary
MKB	To manage the knowledge base and implement artificial reasoning
UGI	To provide a set of public classes and functions to support UGI development.
Diagnosis tool	To evaluate the modeling output, diagnose the problem and prepare for improvement
Modeling transactor	To use procedural models to accomplish intended calculations
Modeling tool	To construct composite models using elementary or other composite models
Interpretation tool	To interpret the modeling output according to prior knowledge, context and pattern
Advisor for spatial modeling	To provide a set of integral functionalities with help of other modules to help users identify their problems and implement modeling

The central control program or other systems can call any of these modules through their interfaces to implement certain functionality. The component-based engine environment can seamlessly integrate each component together within a single software environment while having interaction and scalability across systems.

3 Modeling procedure

3.1 Design criteria of modeling

Haining summarizes several design criteria for geospatial modeling[3]: fitness-for-purpose, robustness, parsimony and uncorrelated residuals. "Fitness-for-purpose" means that the model should be designed to solve the problem under study and give a reasonable explanation to the answers. Robustness means that there is no serious correlation among the predicators and parameters are interpretable in theory. Parsimony means that where there is a choice, always choose the simple one. Uncorrelated residuals mean that the residual in the model should be free of spatial correlation and distributed randomly. The criteria are the directives for the iterative geospatial modeling in our prototype system.

3.2 Modules and procedure of modeling

Fig. 7(a) represents a brief modeling procedure that involves major modules (Fig. 7(b)) and (knowledge and model) bases (Fig. 7(c)). In Fig. 7, the module, Advisor for Spatial Modeling (ASM) is to guide users in the modeling. In support of other modules and bases, ASM initiates the modeling from EDA, model specification, diagnosis, re-specification and interpretation:

(1) User information collection and EDA

The first step is intended to identify the nature of the problem, i.e. purpose and duty (Fig. 2), methods (classical or spatial statistics), data and demand for the accuracy etc. Rule-based knowledge, i.e. theoretical and substantive considerations, prior research and assumption (see Section 2.2) is stored in the EDA knowledge base and used to help users improve their understanding of the problem, find relevant variables and do preparations for modeling.

(2) Model specification

This step specifies a general model for the problem under study. This step needs to make certain the

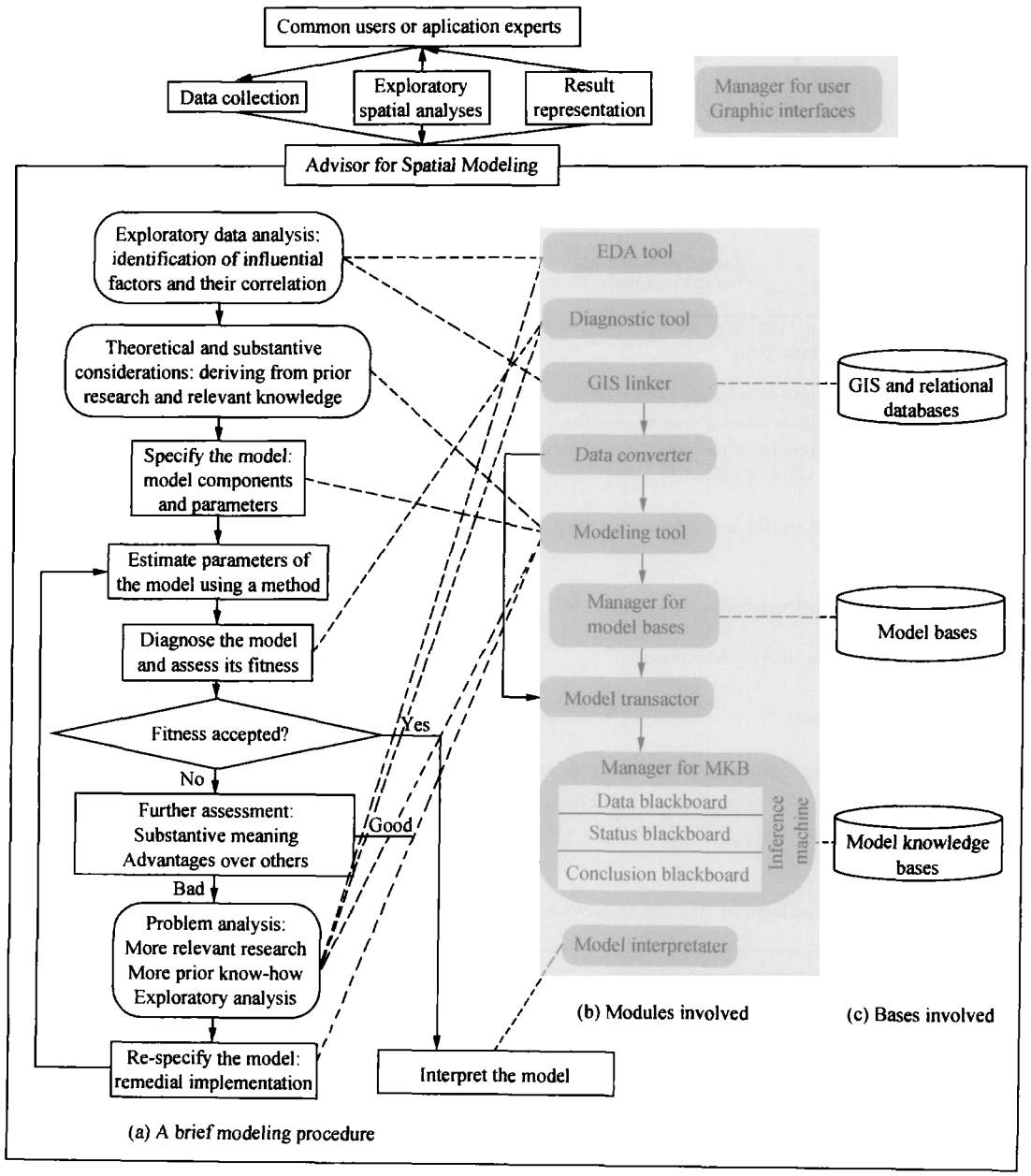


Fig. 7. The modeling procedure and modules involved.

model's type, components and coefficients of interest according to prior theories, relevant considerations and the output from the EDA step (1).

For instance, modeling the ward-level crime rate in a city needs to consider different predicators, effects from different level spatial contexts (i.e. spatial structure), correlation of the response variable from neighbors and spatial unstructured errors. A multilevel model is like this:

$$Y(i,j) = \mu(i,j) + \rho \sum_{k \in N(i,j)} w(i,j,k) Y(i,j) + e(i,j),$$

$$\mu(i,j) = \eta_0 + \mu_1 X(i,j) + \delta_1 V(j) + \delta_2 V(j) X(i,j),$$

where $Y(i,j)$ is the response in (level 1) area i which is a member of (level 2) area grouping j; $N(\cdot)$ denotes the neighbor units and $w(i,j,k)$ is the element of the connectivity's weight matrix dealing with the spatial scale of the neighborhood effect; $X(i,j)$ denotes the effects of predicators at different spatial levels (e. g. neighborhood, ward, city to province).

After the EDA step that uses the ES shell's induction tool and analysis of spatial correlation, the ef-

fects of different spatial scales are removed, and just contributing predictors and spatial autocorrelation of the response are kept in the model. The simplified matrix model is represented as: $Y = (I - \rho W)^{-1} X\beta + u$.

(3) Estimation of parameters

This step selects suitable algorithms, say Maximum Likelihood (ML) to estimate the values of the parameters of interests.

(4) Model's diagnosis

This step examines the effects of the model in terms of its fit acceptability, substantive meaning and advantages over other models. There are some traditional step-by-step diagnostic tools, e.g. cross validation and error comparison for fitness. Our tool is based on the knowledge base system and its artificial inference improves the automatic effects of the diagnosis by the fault-detection rules in combination with traditional methods (Fig. 4). For example, a rule examining the conflict between the parameter's estimate and the true value's constraints can easily find the fault of the model or abnormality of the predicator's contribution to the response.

(5) Problem analysis and model's re-specification

When the model cannot satisfy the criteria or a new theory makes the model up dated, re-specification of the model is needed. The above steps from (1) to (4) are repeated with new information or assumptions considered critically in the updated model so as to reflect the real or new circumstance. Consider the example in step (2). If we find the considerable effects of wider spatial scales and the trivial effect of spatial autocorrelation of the response, the model needs to remove its autocorrelation component $\left(\rho \sum_{k \in N(i,j)} w(i,j,k) Y(i,j)\right)$ and add the components ($\mu(i,j)$) from the effect of bigger spatial scales. The new model is like this: $Y(i,j) = \mu(i,j) + e(i,j)$.

(6) Interpretation of the model

Due to the fact that modeling of spatial data is often used to simulate and explain a social, economical or physical phenomenon, there are strong background, spatial context, mechanisms and process behind the model and interpretation of spatial modeling is ad hoc important in terms of the model's acceptability. The explanation facility of the ES shell and its inherent functionality handing the unstructured knowledge provide the natural mechanism for the interpretation. There are three types of interpretations (see Section 2.2). The first one is the explanation of inference naturally provided by the ES shell; the second one is interpretation of the diagram that is important for EDA and representation of the result and the rule base of pattern recognition of the diagram is a facility for this functionality; the third one is interpretation of model's parameters of interest, that is substantive meaning of the parameter's values for the problem and this is supported by prior knowledge, e.g. the weight for every predicator expresses the contribution degree of the predicator to the response.

4 Discussion with an application case

This section discusses an application of our system in spatial sampling decision.

Our study's area is Shandong Province, China. Our direct goal is to estimate the ratio of the pure cultivatable land's area to the total land's one in a region and our study purpose is to design an optimal sampling scheme for later estimation of the ratio in order to save the survey's cost.

An amount of knowledge about spatial sampling[19—22] has been referred, categorized and organized as objects (for sampling models), facts (for factual stuff such as criteria), and rules (for knowledge about modeling and conditional judgment) into the knowledge bases. With the help of the system, the modeling of spatial sampling decision is carried as the following.

(1) EDA and information collection

This step is to find influential factors, explore the autocorrelation of the samples and estimate some prior values, i.e. the prior mean and variance of the samples.

Fig. 8 is the dialogue box for collecting information from users. According to the type of the duty and the MC value of the samples' autocorrelation, the output from this step indicates that there is no direct influential factors and that spatial sampling methods are preferred due to MC>1.

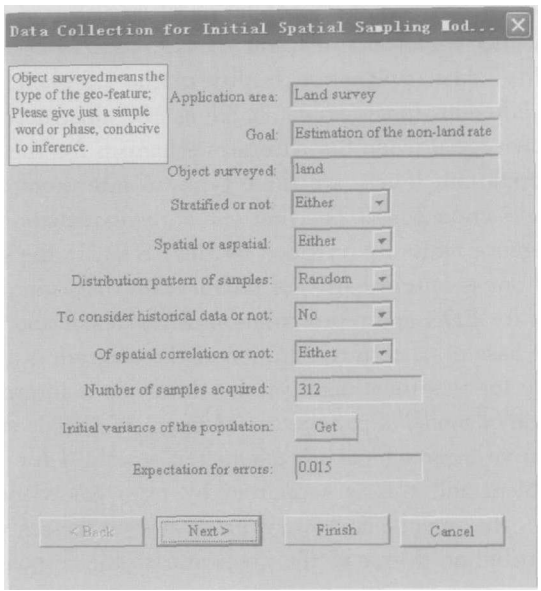


Fig. 8. The dialogue box for collecting information from users.

Further, according to the type of the geo-feature under study (land) and prior knowledge[1,20], the output suggests two algorithms for estimating the spatial correlation, Wang's $E_z[r_h(a-a')]$ and the variogram $\gamma(h)$[17].

(2) Specification of models

According to the step (1)'s output, heterogeneity of the region's land, and the long-range spatial variation (Fig. 9), the step's output indicates the stratified and random sampling frames and three specified spatial sampling models: Wang's sampling model of discrete two-stratified geo-features (SMDTSG)[20], stratified spatial sampling (SSS)[21], and the Kriging model. Also the change-curve-of-error method is also recommended[21].

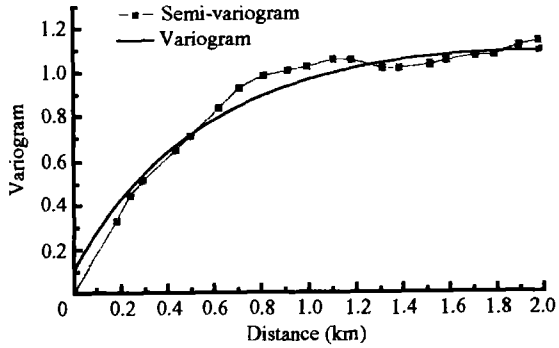


Fig. 9. Semi-variogram and its fit curve. The unit of the sample is the percentage of the cultivable farm's area in the total land's area. About varigram's calculation, please refer to Ref. [5].

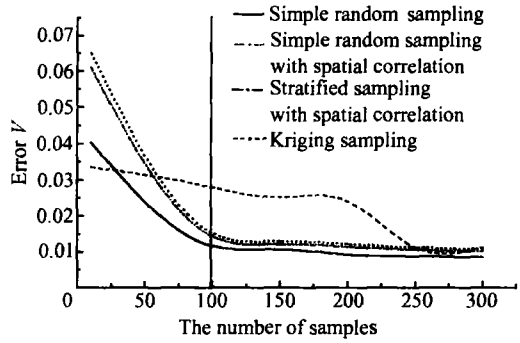
(3) Fitness of the models

There are two kinds of fitness. The first is of the variogram and the other is of the change curve of the error with the number of samples as an independent variable.

For the semivariogram, based on the plot pattern, the spherical model has been chosen to fit the data and the parameters' estimates are: $C_0 \approx 0.1$, $C_1 \approx 1.1$; $a_{max} \approx 0.3$ (Fig. 9).

For the error curve, the number of samples needs to change from small to big to get the curve's trend. To decrease the effect of uneven sampling and too-clustering of samples, sampling has been carried out many times at every fixed number of samples to get as objectively the curve's trend as possible. These opinions are output when users inquire the system for directives of methods. According to the error scatterplot's pattern, the exponential fit model has been recommended to fit these points by the ES shell and Fig. 10(a) gives the error's fit curves for the intended models (SMDTMG, SSS and Kriging) and the model for comparison (simple random).

(a)



(b)

Rule inference	Knowledge source
LHS Geo-feature: land and MC>1 => RHS: algorithms for spatial correlation	Prior research and theory of spatial variation
LHS regional heterogeneity and long-range variation =>RHS: stratified and random sampling frame	Theories of sampling and spatial variogram
LHS geo-feature: land, heterogeneity, and obvious spatial variation=>RHS 3 sampling models suggested	Prior research on spatial sampling and experience
LHS purpose: the trend curve of the error =>RHS multiple sampling at each number of samples	Experience knowledge. for even sampling
LHS recognition of the the plot's pattern => RHS fit model recommended	Analysis of the plot's pattern, experience of fit
LHS the downward trend of the error curve =>RHS substantive meaning of the model	The universal pattern of the error's change

Fig. 10. The interpretation to the plot's output and inference cases in the sampling application. (a) The interpretation of the comparison plot of the error; (b) several typical inference rules and their sources in the sampling case.

(4) Diagnosis of the models

These models selected have been examined using the diagnosis tool. In terms of fit effects, plot interpretation of several re-sampling calculations and visual examination gave evidence to support the fitness. As for substantive meaning, the variogram's fit curve (Fig. 9) shows the long-range spatial variation's type and the error fit curves (Fig. 10(a)) basically reflect a general downward trend. Finally, compared with other models, Wang's model proved to be the best with its least error at the same number of samples, $N(10<N)$.

(5) Re-specification of the models

Re-specification has been incorporated in the comparison of multiple models.

(6) Interpretation of the model

The ES shell provides the explanation facility for the inference procedure. As for the interpretation of the diagram, Fig. 9 is interpreted as a long-range structure of spatial variation and hence heterogeneity of the region and Fig. 10(a) indicates the smallest error of the SMDTMG and hence its optimal feature. Regarding substantive meanings of the model, its sampling structure and equation are suitable for sampling of the similar geo-feature. The values produced from the model gives valuable clues for surveyors to design the sampling scheme for the land survey in a large region (sampling frame, error's demand and the number of samples).

Fig. 10(b) gives several rules and their knowledge sources for the spatial sampling application. The previous research's conclusions, prior knowledge, theories of spatial correlation, information extracted from the diagram and declarative knowledge in models form the foundation for producing the modeling rules.

5 Summary

For complexity of modeling of geospatial data that involves theoretical and substantial considerations, this paper suggests a prototype auto-human support system for spatial analysis. With its functionality of artificial inference, the system, different from the previous softwares for spatial analysis, can code and materialize prior knowledge from researchers and appliers for EDA, specification, fitness, diagnosis and interpretation of the model in modeling, hence being able to use the special knowledge to automatically guide users to model. With a case of optimal spatial sampling, we have showed how to design the modeling and carry out the modeling.

Due to the system's open architecture, the suite of tools can be scalable, interactive and integrative, and be improved continually with updates of the knowledge base and modules. For the coming information-grid era, the suite of tools has several issues deserving endeavors in several directions. The first is strengthening of the self-learning ability by adding modules of machine learning and data mining that can efficiently extract spatial association and relationship from data. The second is expansion of the knowledge base that can improve representation and inference of knowledge. The semantic web having a good logic, scalability and interoperability of knowledge representation in support of powerful open-source software is a very promising technology for guiding modeling. The third is the parallel calculation of the system that can make the suite of tools be able to deal with magnanimous spatiotemporal data in the global web environment.

References

1 Cressie N. Statistics for Spatial Data. New York: Wiley & Sons, 1991, 1—325.
2 Anselin L. Spatial Economics: Methods and Models. Dordrecht: Kluwer Academic, 1988.
3 Haining R. Spatial Data Analysis: Theory and Practice. Cambridge: Cambridge University Press, 2003.
4 Webster R. and Oliver M. A. Geostatistics for Environmental Statistics. Chichester: John Wiley, 1994.
5 Goovaerts P. Geostatistics for Natural Resources Evaluation. London: Oxford University Press, 1997.
6 Insightful Corporation. S-PLUS for ArcView GIS, 2003. http: // www. insightful. com/products/splus.
7 Haining R. P., Wise S. and Ma J. Designing and implementing software for spatial statistical analysis—a GIS environment. Journal of Geographical Systems, 2000, 2: 257—286.
8 Wise S., Haining R. P. and Ma J. Providing spatial statistical data analysis functionality for the GIS user: The SAGE project. International Journal of Geographical Information Science, 2001, 15: 239—254.
9 Goodchild M., Haining R. and Wise S. Integrating GIS and spatial data analysis: problems and possibilities. International Journal of Geographical Information Systems, 1992, 6: 407—423.
10 Leung Y. Intelligent Spatial Decision Support System. Berlin-New York: Springer Verlag, 1997.
11 Li D., Wang S., Li D. et al. Theories and technologies of spatial data mining and knowledge discovery. Geomatics and Information Science of Wuhan University, 2002, 27(3): 221—233.
12 Reistma F. and Albrecht J. Modeling with the semantic web in the geosciences. IEEE Intelligence Systems, 2005, 86—88.
13 Li J. Model System of Regional Development (in Chinese). Beijing: Science Press, 1999.
14 Giarratano J. and Riley G. Expert Systems: Principles and Programming. Boston: PWS Publishing Company, 1998.

15 Li L., Wang J. and Wang C. Typhoon insurance pricing with spatial decision-making support tools. International Journal of Geographical Information Science, 2005, 19(3): 363—384.
16 Perdita S. and Rob P. Using UML: Concerning Object and Component Software Engineering (Chinese). Beijing: People's Posts and Telecommunications Press, 2003.
17 Zhang R. Theory and Applications of Spatial Variation (in Chinese). Beijing: Science Press, 2005.
18 Duda R., Gaschnig H. and Hart P. Model Design in the PROSPECTOR Consultant System for Mineral Exploration, Expert Systems in Micro-Electronic Age. Edinburgh: Edinburgh University Press, 1979.
19 Cochran W. G. Sampling Techniques. 3d ed. New York: John Wiley & Sons, 1977.
20 Wang J., Liu J., Zhuang D. et al. Spatial sampling design for monitoring the area of cultivated land. International Journal of Remote Sensing, 2002, 23(3): 263—284.
21 Li L., Wang J. and Liu J. Optimal decision-making model of spatial sampling for survey of China's land with remotely sensed data. Science in China (Series D), 2005, 48(6):
22 Feng S. and Shi X. Sampling Survey: Theory, Method and Practice (in Chinese). Shanghai: Shanghai S&T Press, 1996.