SPATIO-TEMPORAL INTERPOLATION OF CLASS VARIABLES BY INTEGRATING OBSERVATIONAL DATA AND A BEHAVIORAL MODEL WITH GENETIC ALGORITHM(GA)

Shaobo HUANG* and Ryosuke SHIBASAKI**
Center for Environmental Remote Sensing, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba 263, Japan
shaobo@rsirc.cr.chiba-u.ac.jp
Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi, Minato-ku Tokyo 106, Japan
shiba@shunji.iis.u-tokyo.ac.jp

KEY WORDS: Interpolation, Integration, Behavioral model, Genetic Algorithm(GA)

ABSTRACT

Spatio-temporal interpolation to generate voxel-field data in space-time domain from observational data is indispensable to many spatio-temporal analysis and visualization of dynamic spatial objects. However only very primitive interpolation methods such as nearest neighbor interpolation based on voronoi diagram are proposed for nominal or "class variable" data such as land use data. In interpolating nominal data with these primitive methods, we cannot make use of knowledge on spatial or temporal patterns or behaviors of the object. The authors proposed a spatio-temporal interpolation scheme for generating a voxel-field of nominal data under the framework of optimization of likelihood which is computed from the fitness to both observational data and expected patterns/ behaviors described by a behavioral model or rules specific to the object. Any model which provide likelihood or probability to a given spatio-temporal pattern can be used in this framework. For the optimization of likelihood, a genetic-algorithm (GA) was combined with Hill-climbing (HC) method to increase the efficiency and reliability of optimization. Through some experiments, it is demonstrated that GA/HC based interpolation method can generate voxel- fields which fits both to observational data and to knowledge on it's behavior and that the reliability of interpolation can be compared quantitatively in terms of the maximal likelihood.

1. BACKGROUND & OBJECTIVE OF THE STUDY

Temporal or dynamic analysis of spatial data are needed in various fields such as environmental systems analysis. One of the most fundamental problems which users are facing is the difficulties in generating spatio-temporal field(3D or 4D voxel field) of quality data for analysis through an interpolation or integration of observational data. It is because observational data from multi-sources sometimes have only sparse or biased distribution, different forms (point, edge, polygon and solid in a spatio-temporal space), different resolution and accuracy/reliability.

In several fields, to improve reliability of spatio-temporal interpolation/ extrapolation in generating quality data, models and/or equations describing a mechanism and structure underlying a spatial or behavioral pattern is integrated with observational data. By integrating observational data and models describing underlying mechanisms and structures of object-phenomenon with a GIS, we can provide a GIS-based environment which allow dynamic update of spatio-temporal field of data whenever a new observational data and an improvement of models are given.

Integration methods for data and models have been mainly developed for continuous variables such as temperature and precipitation in meteorology and oceanography. They are known as 4DDA (Four Dimensional Data Assimilation). For nominal or class variables such as land use types, there are only very primitive interpolation methods such as nearest neighbor interpolation and so forth. In this paper, the authors propose a integration methods of models and class data from multi-sources under the framework of optimization of likelihood of spatio-temporal events. For optimizing the likelihood, genetic algorithm (GA) is combined with classical "Hill climbing" method. Experimental results demonstrate that GA with HC can be successfully applied to the integration.

2. GENETIC ALGORITHMS(GA) & NOMINAL VARIABLE INTERPOLATION

2.1 Introduction of Genetic Algorithm (GA)

Genetic algorithms are developed by John Holland and his colleagues as an approach to optimization which requires efficient and effective search in natural and artificial systems. They are search algorithms based on the mechanism of natural selection and evolution of natural genetics. They combine survival of the fittest among string structures with a structured but randomized gene exchange to form a search algorithms with some of innovative flair of human search (D.E.Goldberg, 1989).

Genetic algorithms are computationally simple and powerful in their search without restrictive assumptions about search spaces. In a simple genetic algorithm, five basic aspects should be considered; the representation or coding of problem, the initialization of population, the definition of evaluation function, the definition of genetic operators, and the determination of parameters.

2.2 Optimization Scheme for Nominal Variable Interpolation

Most natural properties seems to vary continuously. Spatial continuity and temporal continuity are intuitive assumptions which provides rationale for interpolating observational data (M.A.Olover, 1990). However, knowledge and rules governing spatio-temporal patterns and behavior of geographic objects (e.g. environmental systems) are now being rapidly accumulated and represented by many simulation models. They can provide more robust and quantitative basis for interpolating observational data, though many of the models still may not be accurate and reliable enough. On the other hand, it can be said that not very reliable results estimated from model simulation can be improved by combining observational data. Actually, integration of observational data and models (GCM etc.) are conducted in meteorology as daily routine. There are not no such attempt to extend the idea of integration to more generic geographic objects.

It is reasonable to assume that spatio-temporal events or "voxel-field" of nominal variables which are estimated should maximize likelihood under given observational data and behavioral models, if we suppose that observation is a probabilistic event and behavioral models are structured and probabilistic *a priori* knowledge on behavior of the object phenomenon. Observational data and behavioral models/ rules can be integrated in the process of maximizing likelihood of spatio-temporal events,

By the way, spatial-temporal data can be divided into two types: continuous variables and nominal variables. Although relatively more interpolation methods have been developed for continuous variables even from multi-source data (e.g.R.Shibasaki et al.(1993) etc.), few interpolation methods have been proposed for nominal data.

In this article we propose a spatio-temporal interpolation of nominal variables which allows integration of observational data with behavioral/structural models/ rules. Since searching for the most likely spatio-temporal voxel-field of nominal data is typical combinatorial optimization problem, we introduce the genetic algorithm as a optimization scheme for class variable data to get optimized interpolated time-slice data. The likelihood is computed based on the fitness of interpolation results both to observational data and to behavioral/structural models/ rules.

3.APPLICATION OF GA FOR INTEGRATING BEHAVIORAL MODELS AND OBSERVATIONAL DATA TO CLASS VARIABLE INTERPOLATION

3.1 3D Representation of an Individual (coding)

In the following sections, three dimensional array is defined to represent the individual(see **Figure.1**). While, the horizontal surface is used to represent 2D space and vertical dimension is used to represent temporal dimension.

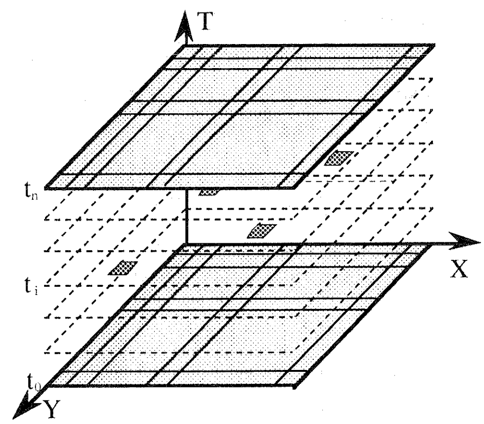

Figure.1 Representation of Individual

3.2 Initialization of Population

An initial population for a genetic algorithm is usually chosen at random; one random trial is made to produce each individual. All members of initial population are chosen automatically by same procedure so that the expected value of each member of initial population is same. In addition we use cubes of 1*1*1, 2*2*2 and 3*3*3 pixels as the initial unit for the initialization of population to increase the efficiency of procedure.

3.3 Definition and Computation of Individual's Fitness

3.3.1 Spatio-temporal Behavioral Models/Rules of Class Variable Data

Any types of behavioral/structural models/rules can be used for the GA-based interpolation if they can determine the probability of every possible behavior/transition of nominal or "class" variables. For nominal variable data, possible changes of a class at one pixel is basically defined by the probability of the changes from one class to another.

One of the simplest example is a Markov chain, where transitional probability is determined only by the previous class. In addition the probability also can be affected by combination of classes in the neighborhood. In this study, we assume a model at which transitional probability is determined by the combination of classes in the neighborhood. And landcover data with five classes is used as test data.

In an example model which we used in an experiment, spatial and temporal relations affect the transitional probability in three ways as shown in the **Figure.2**. The first one can be called "spatial continuity", based on the assumption that the same class data tends to continue in spatial dimension. Second one is called temporal continuity. This is an extension of spatial continuity to temporal domain. The third aspect is expansion-contraction relations based on the assumption that some class data has high possibility to expand its area at next time-slice, while others tend to contract. The temporal change in the pixel with un-contractible class type will be determined by the pixels class itself. And the temporal landuse change in the pixel with contractible type will be determined by class of the pixel and classes of its expansible neighborhood.

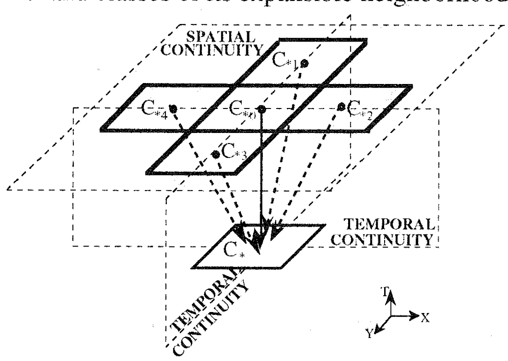

Figure.2 3D Spatial-Temporal Relation of Pixel-based Class Variable Data

3.3.2 Definition and Computation of Fitness of an Individual

Fitness of an individual is defined by the combination of behavioral fitness and observational fitness. Behavioral fitness is defined as combined probability of change events of nominal variables under the condition that these changes follow a given probabilistic behavioral model or rule. Observational fitness can be defined as combined probability that the observational nominal values occur under probabilistic functions of observational errors/uncertainties. Observational probability can be determined by accuracy, resolution and frequency of observation. By multiplying behavioral fitness and observational fitness, overall fitness can be computed. Behavioral/structural models and observational data can be integrated by optimizing the overall fitness.

1)Behavioral Fitness: As showing in **Figure.2**, let $P_V(C_{P,T}, C_{P,T+1})$ as the probability of changes from landuse class C_{*0} to landuse class C_* considering the expansion-contraction effect of its neighbourhood, and $P_{SC}(C_{*0} \sim C_{*4})$ as the probability of spatial continuity. If we assume that $P_{TC}(C_{*0}, C_*)$ and $P_{SC}(C_{*0} \sim C_{*4})$ are independent, we can compute behavioral fitness of each individual according to following formula:

$$\underset{\text{(behavioral fitness)}}{\text{FITNESS}} = \prod_{P=1}^{N_P} \{ \prod_{T=1}^{N_T} P_V(C_{P,T}, C_{P,T+1}) \}$$

$$= \prod_{P=1}^{N_P} \{ \prod_{T=1}^{N_T} P_{TC}(C_{*0}, C_*) P_{SC}(C_{*0}, C_{*1}, C_{*2}, C_{*3}, C_{*4}) \}$$

where N_P : is the pixel number in 2D space,
N_T : is the temporal slice number,
$C_{P,T}$: is the landuse class of the cell on the Pth pixel at the T time slice;

For the class change probability with spatial continuity, $P_{SC}(C_{*0} \sim C_{*4})$, we set values according to following five neighboring pixel's statues along the spatial dimension, which form a set of behavioral rules: 1)If or not classes in all neighbouring pixels are equal; 2) If or not classes in 4 neighbouring pixels are equal; 3) If or not classes in 3 neighbouring pixels are equal; 4) If or not classes in 2 neighbouring pixels are equal; 5) If all classes in 5 pixels are unequal.

To calculate the probability of class changes under the temporal continuity/expansion-contradiction effect, $P_{TC}(C_{*0}, C_*)$, three possible changing patterns of landuse classes in spatial-temporal distribution are picked up and listed in the **Figure.3**. They will be determined based on the probability integrating class changes in Markov chain, $P_M(C_{*0}, C_*)$, and expansion speed of class-types into neighboring pixels. Their behavioral rules can be reckoned from tables similar to **Table.1**.

2) Observational Fitness: Observational fitness can be computed with the following formula. Observational probability can be determined mainly by the accuracy of

Invasion C_{*i1}	Invasion C_{*i2}	Value of C_*: C_{*i1} $P_M(C_{*0}, C_{*i1})$	C_{*i2} $P_M(C_{*0}, C_{*i2})$	$C_{*i1} = C_{*i2}$ $P_M(C_{*0}, C_{*i1})$	Others $P_M(C_{*0}, C_*)$
Yes ($\alpha_{C_{*i1}}$)	Yes ($\alpha_{C_{*i2}}$)	0	0	Invasion	0
Yes ($\alpha_{C_{*i1}}$)	No ($1-\alpha_{C_{*i2}}$)	Invasion	0	Invasion	0
No ($1-\alpha_{C_{*i1}}$)	Yes ($\alpha_{C_{*i2}}$)	0	Invasion	Invasion	0
No ($1-\alpha_{C_{*i1}}$)	No ($1-\alpha_{C_{*i2}}$)	Markov Chain	Markov Chain	Markov Chain	Markov Chain

Notice: 1> Supposed C_{*i1} and C_{*i2} are expansible (i1, i2 = 1 ~ 4) ;
2> $\alpha_{C_{*i1}}$ and $\alpha_{C_{*i2}}$ are defined as expansion speed of C_{*i1} and C_{*i2} ;

Table.1 Behaviors of Class Changing in Case3

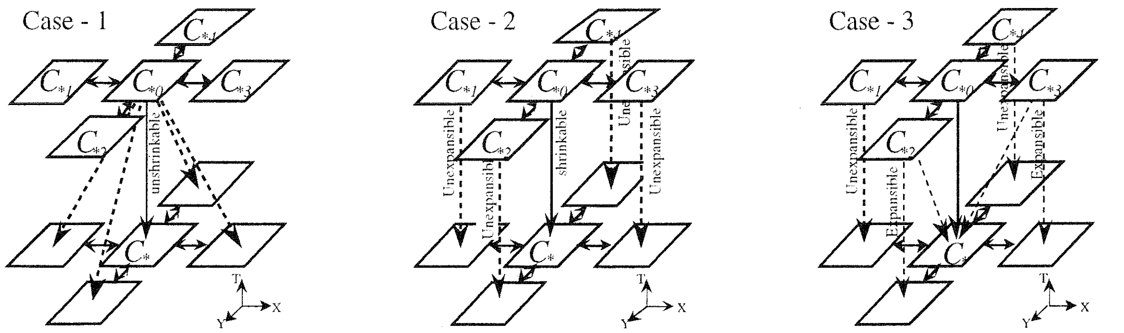

Figure.3 Possible changing patterns of landuse classes in temporal and spatial distribution

observation. Observational location and time/frequency can be represented by locating observational pixel in the two dimensional string. Spatio-temporal resolutions can be represented by setting an aggregation formula over the range of observation.

3) Total Fitness: Total fitness is computed from behavioral fitness and observational fitness by the following formula. Total fitness = Behavioral fitness * Observational fitness or ln(Total fitness) = ln(Behavioral fitness)+ln (Observational fitness)

$$\text{Observational Fitness} = \prod_{n=1}^{No} P_{Obs.}(C_{P,T}\ .C_{P,T,Obs.})$$

$P_{Obs.}(C_{P,T}\ .C_{P,T,Obs.})$: Probability that observational value $C_{P,T,Obs}$ is given when actual value is $C_{P,T}$

3.4 Definition of Operators

3.4.1 Reproduction

Reproduction is a process in which individual strings are copied according to their objective function values or the fitness values. Copying strings according to their fitness values means that strings with a higher value have a higher probability of contributing one or more offspring in the next generation (**Figure.4**). This operators, realizing an artificial version of natural selection, a Darwinian survival of the fittest among string creatures.

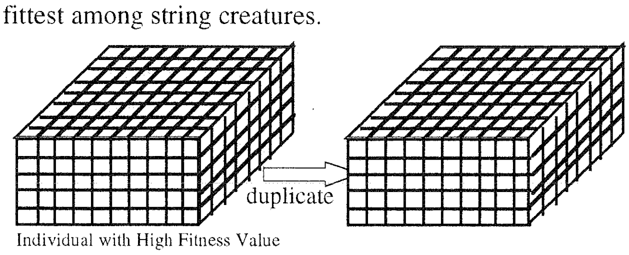

Figure.4 Reproduction of Individual

There are several proposals for selecting survival individuals. The most basic scheme is called the roulette wheel scheme, the deterministic sampling and the elitist scheme. In order to efficiently find the best solution in a search space, in our search, we proposed the selection scheme based on the combination of the deterministic sampling and the elitist scheme. The selected survival possibility in next generation of each individual is calculated as in the deterministic sampling. And the best individual is kept into the next generation as in the elite scheme.

3.4.2 Crossover

Crossover operator first randomly mates newly reproduced individuals in the mating pool. Then it randomly locates a window with random size for a pair of individuals. Finally, the contents of individuals within the window are swapped to create new individuals (**Figure.5**).

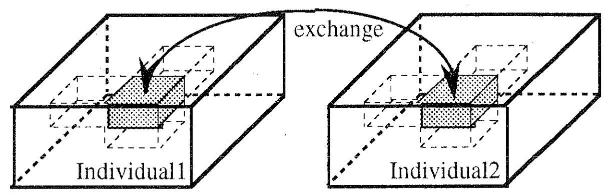

Figure.5 Crossover of Individuals

3.4.3 Mutation

Mutation operator plays a secondary role in the simple GA. It occasionally alters the value in a individual position (**Figure.6**).

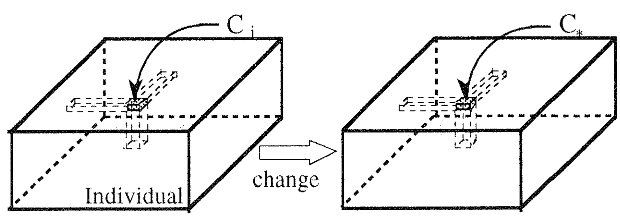

Figure.6 Mutation of One Individual

4. IMPROVEMENT OF THE SEARCH IN GAs

4.1 Hill-Climbing method to improve the efficiency of genetic algorithm

Searching a complex space of problem resolutions often involves a tradeoff between two apparently conflicting objectives: exploiting the best solutions currently available and robustly exploring the space (Lashon Booker, GA&SA). Generic algorithms have been toured as a class general purpose search strategies that strike a reasonable balance between exploration and exploitation. The power of these algorithms is derived from a very simple heuristic assumption: that the best solutions will be found in regions of the search space containing relatively high proportions of good solutions. The problem is that, if the complex space of problem resolutions become larger and larger, the population size and the generation size have to be increased at same time. Therefore, the efficiency of GA is one of obstacles to apply GA in reality.

Hill climbing is a good example of a search strategy that exploits the best among known possibilities for finding an improved solution. Although Hill-Climbing strategies is easy to trap in one of local maxima more far away from the optimal solution, it is a very good search strategy that exploits the best among known possibilities for finding an improved solution. So in our research we try to combine the Hill-Climbing strategy with GA.

4.2 Maintenance of Population Diversity

Estimates based on finite samples in GA inevitably have a sample error associated with them. Repeated iterations of algorithm compound the sample error and lead to search trajectories much different from those theoretically predicted. The most serious phenomenon is the premature convergence. The premature convergence is caused by early emergence of an individual that is better than the others in the population, although far form optimal. Copies of this structure may quickly dominate the population. Search continues then but is concentrated in the vicinity of this structure and may miss much better solutions elsewhere in the search space.

To avoid the premature convergence, one has to maintain population diversity or to reduce the different of best fitness with others. Although, to reduce the reproduction number can not eliminate the premature convergence, it can be used as a simple way to reduce the rapid convergence. Therefore, in our research, we limited the duplicated number of individuals less than two. It means that if individual's expected duplicated number is larger than two, we will force it to equal two. To do so, the premature convergence speed can be reduced.

5. EXPERIMENTS

The test program of GA/HC for spatio-temporal interpolation of pixel based landuse data was coded with C language and was run on SPARC/station2. Several simple landuse class variable data had been used to test program. Small spatio-temporal datasets are used in the experiment to check the behavior of the GA/HC based interpolation under different conditions. The test data size of individual had been defined with 20 pixels * 6 time-slices for two dimensional case, while 11lines * 11 columns * 6 time slices for three dimensional case. The first and last time-slice in

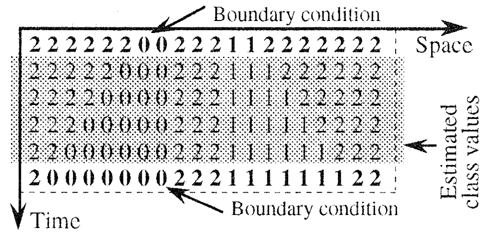

Figure.7 a) Result of GA/HC (2D case)

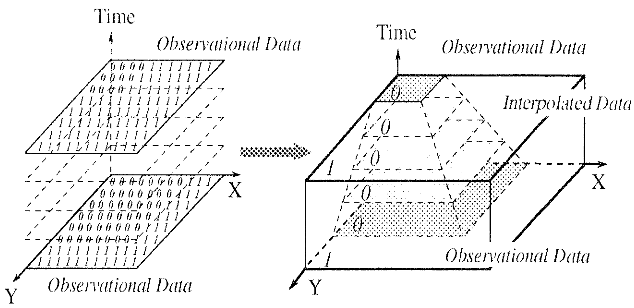

Figure.7 b) Result of GA/HC (3D case)

the individual are supposed to be sample (observational) data and all middle time-slices should be estimated by the interpolation. In these experiments, we set the generation size of GA/HC to 2000, which was large enough to get stable results. The probability of crossover operation was defined as 0.7, while the probability of mutation operation was relative small in natural population, so that we used 0.01 as the probability of mutation. **Figure.7** is two dimensional and three dimensional experiment results of GA/HC, in which the individual has largest fitness value. With the assumed model, smooth transition/expansion has the largest fitness or likelihood.

In Figure 8, observed location of class '0' in the first time slice and the last time slice are overlapping spatially. The Interpolation result naturally connect class '0' together, forming a band of class '0'. On the other hand in **Figure 9**, while class '0' is not overlapping, it is demonstrated that the most likely interpolation do not connect class '0' together, and that a case forming a band of class '0' apparently have lower likelihood, though it looks natural. In**Figure.10**, another observational data is given at the middle. In this case, the most likely spatio-temporal pattern of class changes has a band of class '0'. It is concluded that the interpolation method integrating observational data and behavioral models/rules can estimate the most likely voxel field under different conditions.

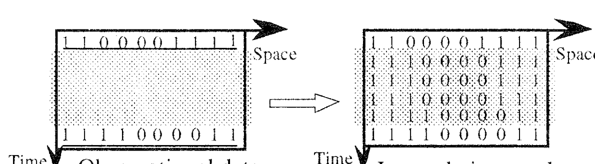

Figure. 8 Interpolation Result (Overlapping case)

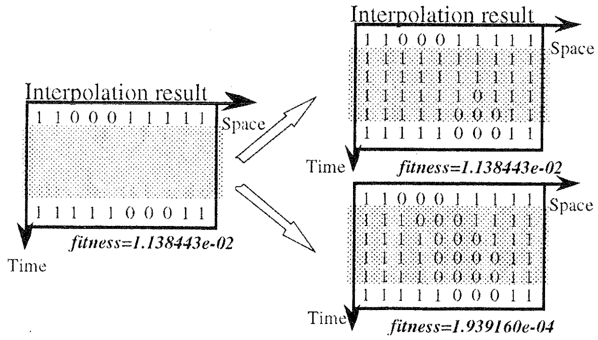

Figure.9 Interpolation Results in Non-overlapping Case

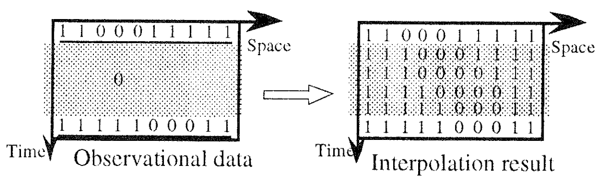

Figure.10 Interpolation Result with Additional Observational Data

6. CONCLUSION AND FUTURE PROSPECTS

In this study, a spatio-temporal interpolation scheme is proposed for raster nominal data which can integrate observational data with behavioral/structural models/rules under the framework of maximizing likelihood of spatio-temporal events. Genetic-Algorithm/Hill-Climbing can be successfully applied to the combinatorial optimization of nominal voxel-field data. Conclusions from the experiments can be summarized as follows:

1) GA/HC can be very rigorous because it can generate the most likely spatio-temporal distribution of class variables under observational data and a behavioral model;

2) Hill-Climbing method can be effective method to greatly improve the efficiency of GA;

Although the GA/HC authors proposed can be a good scheme for spatio-temporal interpolation, it is just a first attempt to apply GA in the field of spatio-temporal interpolation. We will apply GA/HC for larger size of class variable data, to reduce the speed of premature convergence and get higher efficiency of GA/HC.

[1] Bramlette, M.F. (1991): **Initialization, Mutation and Selection Methods in Genetic Algorithms for Function Optimization**, Proc. of the 4th. Conf. on GA, R.K.Belew and L.B.Booker (Editors), July, 1991, pp.100-107.

[2] Booker, L. (1987): **Improving Search in Genetic Algorithms,** Genetic Algorithms and Simulated Annealing, edited by Lawrence Davis, Morgan Kaufmann Publishers, Inc. Los Altos, California, 1987, pp.61-73.

[3] Burrough, P.A. (1986): **Methods of Spatial Interpolation**, Principle of GIS for Land Resource Assessment, Monographic on Soil and Research Survey, No.12, 1989, pp.147-166.

[4] Davis, L. (1987) : **Genetic Algorithms and Simulated Annealing**, Pitman Publishing, 128 Long Acre, London WC2E 9AN, 1987.

[5] Eshelman, L.J. and J.D.Schaffer (1991): **Preventing Premature Convergence in Genetic Algorithms by Preventing Incest**, Proc. of the 4th. Conf. on GA, R.K.Belew and L.B.Booker (Editors), July, 1991, pp.115-122.

[6] Gold, C.M. (1989): **Surface interpolation, spatial adjacency and GIS**, Three Dimensional Applications in Geographic Information System, Edited by J.Raper, Taylor & Francis Ltd., 1989, pp.22-35.

[7] Goldberg, D.E. (1989): **Genetic Algorithms in Search, Optimization & and Machine Learning,** Addison-Wesley Publishing Company, Inc., 1989.

[8] Olover, M.A. and R.Webster(1990): **Kriging: a method of interpolation for geographical information systems**, Int.J.Geographical Information Systems, 1990, Vol.4, No.3, pp.313-332.

[9] Shibasaki,R., T.Ito and Y.Honda (1993): **Integration of Remote Sensing and Ground Observation Data for Developing Global GIS**, Proc. of SEIKEN Symposium, Vol.12, Aug. 1993, pp.263-277.

[10] Shibasaki,R., M.Takaki and Y.Honda (1994): **Spatio-temporal Interpolation Using Multi-Source Data for Global Dataset Development**, COMMISSION III, IGWG III/IV; 1994.

[11] Huang, S.B. and R. Shibasaki (1995): **Development of Generic Algorithm / Hill-climbing Method for Spatio-temporal Interpolation**, Sixth symposium on functional graphics and geographic information systems ,IIS, Univ. of Tokyo, 1995.