ON UNCERTAINTY IN GEOGRAPHIC DATABASES DRIVING REGIONAL BIOGEOCHEMICAL MODELS*

(EXTENDED ABSTRACT)

Ferenc Csillag

Geography Department & Institute for Land Information Management
University of Toronto, Erindale College
Mississauga, ONT, L5L 1C6, Canada
<fcs@eratos.erin.utoronto.ca>

ABSTRACT

Large regional geographic databases are becoming crucially important in driving regional environmental models. When these databases contain information from various sources about different geographic phenomena, defining a common reference of "environmental units" may be difficult. A general framework is proposed to evaluate regular and irregular landscape partitioning strategies, as well as numerous spatial predictors to provide input data for complex models. The strategy is based on information available on spatial (and temporal) variability. It is concluded that it is advantageous to adjust the "effective scale" of representation to local variability.

INTRODUCTION

There is increasing demand to link environmental simulation models and geographic information systems (GIS) to predict the status of various environmental phenomena. Typically, these efforts extend a *site-based (process) model* over a region (Figure 1), which may be several orders of magnitude larger than "calibrated sites". Therefore, the extrapolation requires some *spatial model*, which accounts for the spatial variability of the landscape.

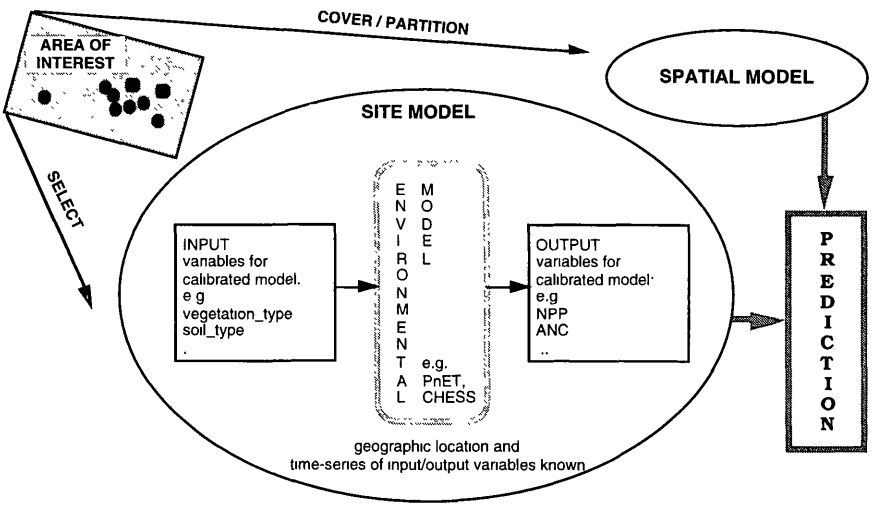

Figure 1. Linking "calibrated" site-based models with GIS for prediction

* This research has been supported by NSERC and the IBM Environmental Research Program.

Considerations must be given to the nature of spatial models, for example, when net primary productivity of forested ecosystems (Aber et al. 1993), acid neutralizing capacity of aquatic ecosystems (Driscoll and Van Dreason 1993), or grassland soil nutrient availability (Parton et al. 1988) is to be predicted at *regional scales*.

This paper focuses on possible choices of the spatial model for regional application of environmental process models as a function of information available about the modeled landscape. A general framework is presented for partitioning the landscape into "soft objects" (mapping or environmental units), which can be derived within a GIS by analysis of local variability. These units facilitate both loose and close coupling of GIS and the model (across data structures), and serve as common ground for sensitivity analysis of the predictions.

LINKING SITE-LEVEL MODELS AND REGIONAL MODELS

Extending site-level models to large regions frequently relies on the assumption that the region is a collection of comparable sites; thus they imply some form of *partitioning* of the region. This regional partitioning may require different strategies depending on which landscape, or environmental characteristics are predicted. The geographic context, which is quite different for climate variables, lake chemistry, or vegetation composition, is usually described within a GIS, and it will constrain the *interpolation* within the partitions (Figure 2).

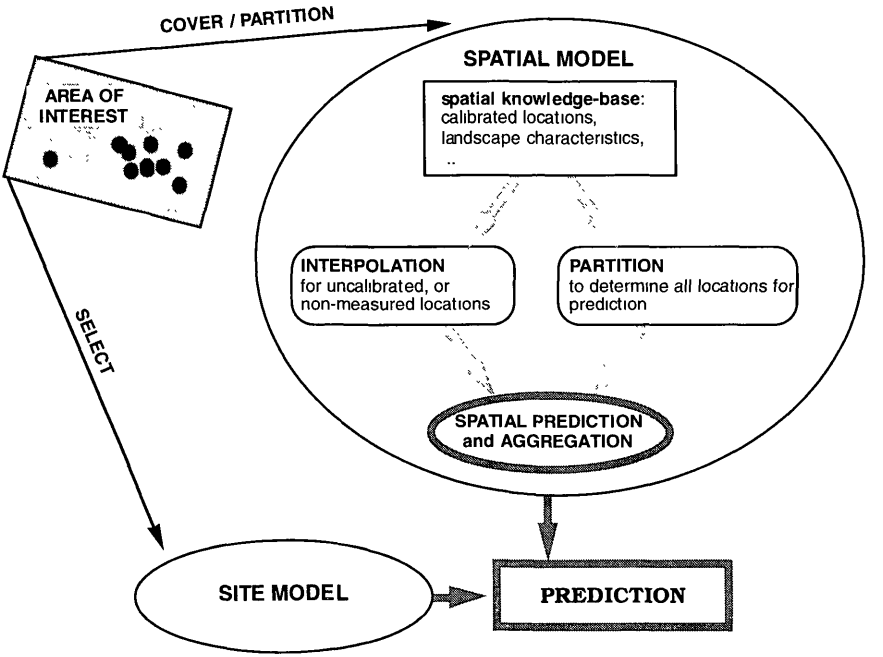

Figure 2. Expansion of the spatial model

Partitioning (which in itself may be scale-dependent) these data sets to units, that correspond to input and control requirements of the underlying processes and their models, is a function of spatial (and temporal) variability of the phenomena to be characterized, and the amount and nature of information available. In a general framework a variety of strategies can be identified to derive "mapping units" according to preference given to statistical, geometric and/or biophysical control over partitioning -- such as in geostatistical (Webster and Oliver 1990), tessellation-based (Gold and Cormack 1987), or watershed-oriented (Band et al. 1991) models, respectively.

These strategies can also be compared and classified from a data structure perspective; primarily considering whether they employ regular or irregular spatial units. For example, a grid of biomass (e.g., as derived from satellite images) does not use any "environmental" control in determining the spatial units (i.e., all sites/locations are "equal") and it is completely regular; as opposed to a biophysical and statistical knowledge-based watershed delineation (e.g., derived from a DEM) which is rather irregular; or a set of Voronoi-polygons constructed based on a set of observation sites (e.g., weather stations, or lakes) is irregular and based on geometric distribution.

Once spatial units are defined for the database, the next step is to supply data for the entire region of interest, which usually requires some form of interpolation. The more information is available about spatial structure of the phenomenon, the more reliable spatial prediction can be. The reliability of describing spatial structure, however, partly depends on the arrangement and size (distribution) of spatial units. Furthermore, efficiency and feasibility considerations may conflict with statistical optimality.

The combination of these functionality requirements for linking and interfacing GIS and environmental models results in a fairly complex system (Figure 3). There are two primary objectives for the system currently under development: (1) to keep the complete functionality of both GIS and environmental model, and (2) to provide guidelines for (or, potentially automate) the choice of the spatial model according to uncertainty analysis of the prediction.

DATABASE AND INTERFACE COMPONENTS

One of the key challenges facing the linkage between GIS and site-based models lies in considering the sensitivity of the model to errors in the parametrization derived from the geographic database. This problem is often avoided in site-based calibration, which is sometimes called "validation", because the uncertainty (e.g. variance) in input parameters is set to zero. In regional studies, however, the emphasis is on the "collection of sites", therefore, it is imperative to consider the relationship between the uncertainty of location and attributes (Csillag 1991). For example, an effort to predict the acid/base chemistry in the lakes of the Adirondack Mountains, NY, requires, among others, input of precipitation, min/max temperature, slope and aspect (for solar radiation), vegetation cover and soil characteristics information to a biogeochemical model (PnET, Aber and Federer 1992). Whenever a set of input parameters is available, the model is "ready to run", and will

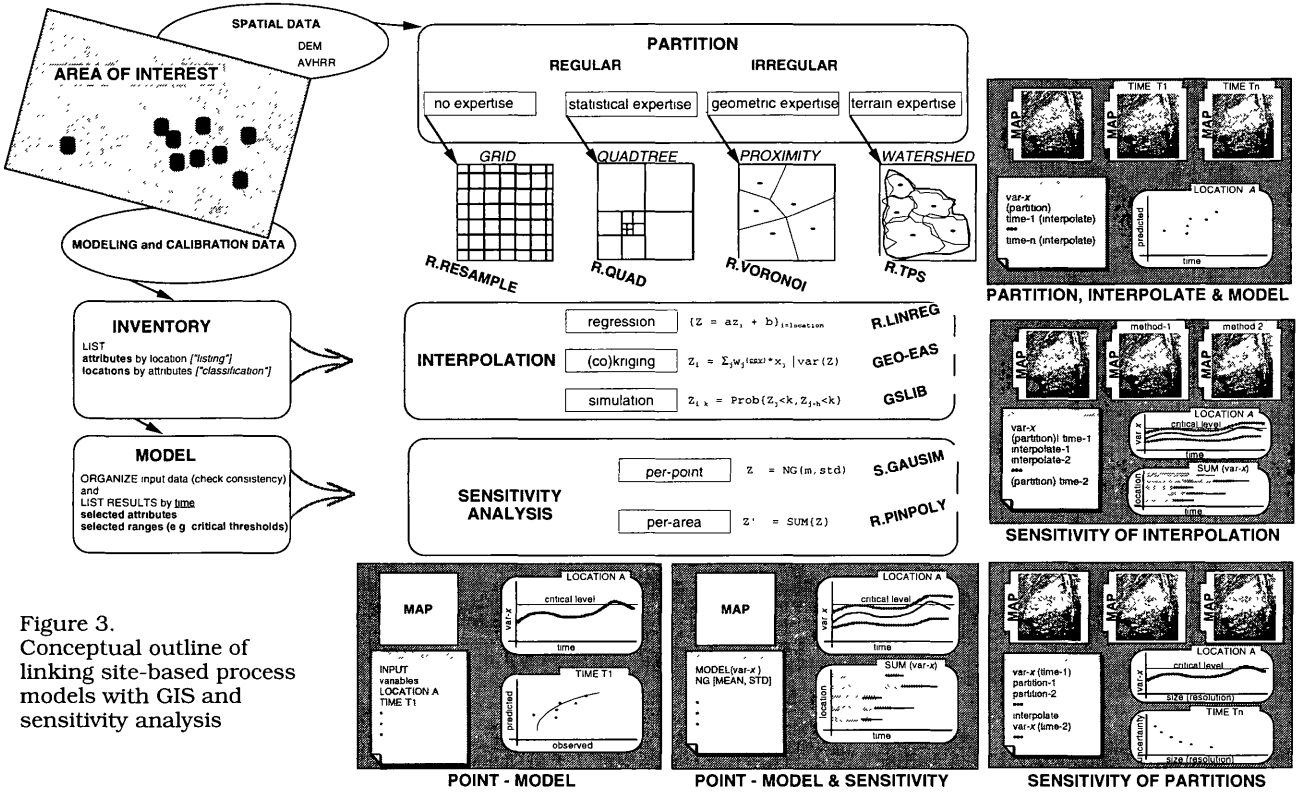

Figure 3.
Conceptual outline of linking site-based process models with GIS and sensitivity analysis

result in a time-series of a variable (e.g., the amount of dissolved N in drainage water).

Whenever collections of sites are used as input to the model, it is worthwhile and necessary to assess the sensitivity of the model to the errors in the input parameters. Since most of the process models are non-linear, their sensitivity to variation in input values is generally assessed by Monte-Carlo simulation, and is reported as "confidence intervals" around calibrated values. In spatial context then, if we can determine, or even limit, the (residual) variance of our partitions, this information can be directly related to the sensitivity of the model output. For example, if we have two (or more sites) close to each other with similar characteristics, we may be better off aggregating them into one "soft unit"; it would result in significant (50% or more) savings in computing requirements while the uncertainty in prediction may be kept below a required threshold. Since the "calibrated sites" are usually very small compared to the regions in question (e.g., a 4 ha lake watershed compared to the 3.5 million ha Adirondack ecological zone), this strategy potentially offers major compensation of attribute versus spatial accuracy. Consequently, during partition and interpolation one should control the tradeoff between the level of spatial detail (resolution) and the accuracy of prediction.

TOWARD USING MULTI-PARTITIONS IN SPATIAL ESTIMATION

As outlined on Figure 3, there are many feasible approaches toward partitioning the geographic landscape and interpolating environmental state-variables across partitions. Depending on the nature of data (e.g., a DEM versus a collection of lakes), and the level of expertise (e.g., is a detailed DEM available, or not), the partitioning strategies are classified into four groups. Since data structures in GIS and environmentally meaningful "units" do not necessarily coincide, the strategies are also grouped from a data structure perspective. Regular partitions (Figure 4) include *grids*, which do not utilize any expertise, and *quadtrees*, which can be constructed by statistical constraints on within-leaf variance (Csillag et al. 1994). Irregular partitions include *Voronoi-polygons*, which rely on geometric expertise, and *watersheds*, which are based on terrain expertise.

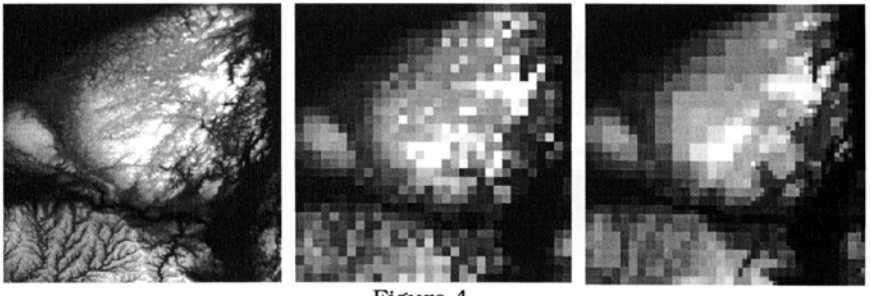

Figure 4.
Original (100m) DEM (left), regular grid aggregation to 1460 units (middle)and quadtree-tessellation with 1460 leaves (right) of the Adirondack Mts. (Note: the total variance of aggregates is reduced by using the quadtree instead of grid-based aggregation.)

Once the GIS is capable of controlling the within-unit variance, it can provide guidance to the user, who sets the accuracy thresholds for the model prediction. At the same time, it is not required that all input variables be partitioned the same way; the "soft objects" can be carried over and can be further used in interpolation within the partitions.

The combination of interpolation with (optionally limited-variance) partitions facilitates further control of uncertainty (Dungan et al. 1994, Mason et al. 1994). Since interpolation is always carried out using as much information about spatial variability as possible, for each partition the partition-mean will be more robust, and the lack of information will not spread from one partition to another. Furthermore, during interpolation the uncertainty can also be determined for each partition. The combined uncertainty associated with partitioning and interpolation can be reassessed before running the model. This is particularly important when information on one variable (e.g. elevation) is used to (co-)estimate another (e.g. acid deposition). Without partitioning the predictor variable simple estimators (e.g., non-spatial regression) lead to enormous residual variance; however, partitions can dramatically reduce it (Figure 5).

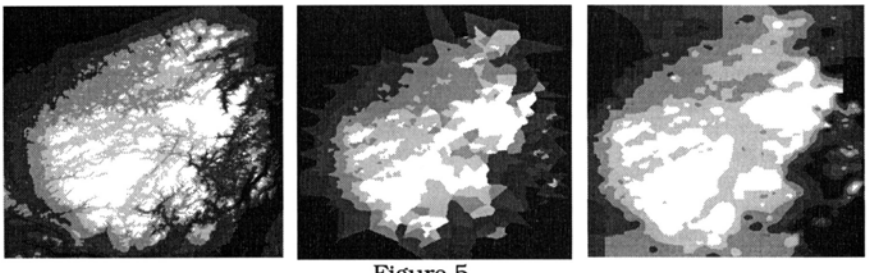

Figure 5.
Original (100m) DEM grid of the Adirondack Park (left); Voronoi-polygon mean elevations based on 1468 lakes (middle); kriged elevation based on 1468 lakes (right).

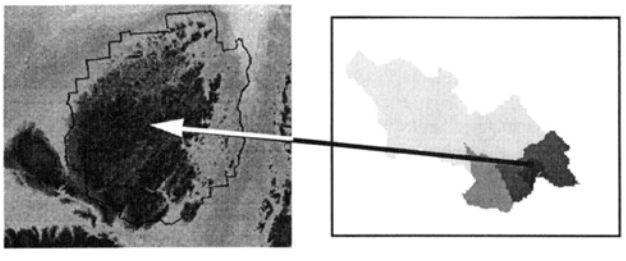

Figure 6.
PnET prediction of the amount of dissolved N on a regular (1 km) grid. (The gray-levels represent 0.4-1.4 mg/liter.) Arbutus Lake (one of the calibration sites) is marked with watersheds derived with various limits on internal heterogeneity.

CONCLUSION AND WORK IN PROGRESS

One of the major challenges in linking environmental models and GIS is to control the uncertainty related to input derived from the geographic database to drive the environmental (process) model. A general framework is proposed to combine the analytical capabilities of GIS with sensitivity analysis of a biogeochemical model (PnET) to control uncertainty in a simulation study, aiming to predict acidification in the northeast US (Figure 6). Most of the elements for interfacing GIS and environmental models by "soft objects" have been implemented in grass (see Figure 3). Current efforts are focused on automating the use of the analytical components.

REFERENCES

Aber, J.D., C. Driscoll, C.A. Federer, R. Lathorp, G. Lovett, J.M. Melilo, P.A. Steudler and J. Vogelman. 1993. A strategy for the regional analysis of the effects of physical and chemical climate change on biogeochemical processes in northeastern US forests. Ecological Modelling 67:37-47.

Aber, J.D. and C.A. Federer. 1992. A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary productivity in temperate and boreal ecosystems. Oecologia 92: 463-474.

Band, L.E., D.L. Peterson, S.W. Running, J. Coughlan, R. Lammers, J. Dungan and R. Nemani. 1991. Ecosystem processes at the watershed level: basis for distributed simulation. Ecological Modeling 56: 171-196.

Csillag, F. 1991. Resolution revisited. AutoCarto-10 (ASPRS-ACSM, Bethesda) p. 15-29.

Csillag, F., M. Kertész and Á. Kummert. 1994. Sampling and mapping of two-dimensional lattices by stepwise hierarchical tiling based on a local measure of heterogeneity. International Journal of GIS (in press)

Driscoll, C.T. and R. Van Dreason. 1993. Seasonal and long-term temporal patterns in the chemistry of Adirondack lakes. Water, Air and Soil Pollution 67: 319-344.

Dungan, J. L., D.L. Peterson and P.J. Curran. 1994. Alternative approaches to mapping of vegetation amount. In: Michener, W.K., J.W. Brunt and S.G. Stafford (Eds.) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis, London (in press)

Gold, C. and S. Cormack. 1987. Spatially ordered networks and topographic reconstructions. International Journal of GIS 1:137-148.

Mason, D.C., M. O'Conaill and I. McKendrick. 1994. Variable resolution block-kriging using a hierarchical spatial data structure. International Journal of GIS 8:429-450.

Parton, W.J., J.W.B. Stewart, C.V. Cole. 1988. Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry 5: 109-131.