York University
Faculty of Environmental Studies

ENVS 7189

Advanced Geographic Information System

Modeling with GIS

Introduction

Several points must be made before discussing GIS modeling. First, GIS models can be vector-based or raster-based. The choice between these two types of models is determined by the nature of the model, data sources, and the computing algorithm. Generally speaking, a raster-based model is preferred if the modeling involves intense and complex computation; for this reason, process models and, to a lesser degree, regression models are usually raster-based. A raster-based model is also preferred, if raster data such as satellite images and DEMs (digital elevation models) constitute a major portion of the input data. Second, it is very common for GIS researchers to use vector data in constructing a raster-based model and vice versa. This is because existing public data are in both formats. Third, algorithms for conversion between vector and raster data are easily available in most GIS packages and the decision as to which data format to use in analysis is no longer restricted to the format of the original data. Because of the above reasons, the topic of GIS modeling is important to both vector and raster data analysis.

Depending on the degree of complexity, GIS modeling may take place in a GIS or require the linking of a GIS to other computer programs. Generally speaking, **binary models** and **index models** belong to the former, and **regression models** and **process models** the latter. A statistical analysis package can better accommodate regression analysis, which is the core of regression models, than a GIS. In the case of process models, a GIS is often a tool for database management, data visualization, and spatial analysis while simulation models embedded in other computer programs work with complex and dynamic analysis. A GIS can be linked to other computer programs using the following three strategies. A loose coupling involves data transfer from one to another, a tight coupling provides a common user interface to the GIS and other computer programs, and an embedded system bundles the GIS and other computer programs with shared memory and a common interface. Most GIS applications described in the literature are based on the first two strategies.

Binary Models

A binary model uses logical expressions to select map features from a composite map. The output of a binary model is in binary format: 1 (True) for map features that satisfy the logical expressions and 0 (False) for map features that do not. A common application of binary models is siting analysis, with each logical expression corresponding to a siting criterion. Suppose a county government wants to select potential industrial sites based on the following criteria:

- At least five acres in size
- Commercial zones
- Vacant or for sale
- Not subject to flooding

- Not more than one mile from a heavy duty road
- Less than 10% slope

To proceed with the task, the county government can gather all digital maps relevant to the criteria and create a one-mile buffer zone map by buffering heavy-duty roads. A series of map overlay operations are followed to combine the road buffer zone map and other maps. A query of the composite map involving all the selection criteria can reveal which parcels are potential industrial sites.

Another example of siting analysis is the Conservation Reserve Program (CRP) administered by the Farm Service Agency (FSA) of the US Department of Agriculture. Land eligible to be placed in the CRP includes cropland that is planted to an agricultural commodity two or the five most recent crop years. Additionally, cropland must meet the following requirements:

- Have an Erosion Index of 8 or higher or be considered highly erodible land
- Be considered a cropped wetland
- Be devoted to any of a number of highly beneficial environmental practices, such as filter strips, riparian buffers, grass waterways, shelter belts, wellhead protection areas, and other similar practices
- Be subject to scour erosion
- Be located in a national or state CRP conservation priority area, or be cropland associated with or surrounding non-cropped wetlands

Government agencies, such as the Natural Resources Conservation Service, support the CRP by provide information to farmers, who may be eligible for the program. A binary model based on the program requirements would be an efficient way to compile a list of farmers to contact.

Binary models can also be used for change detection. The most common method for change detection is to overlay maps representing the spatial distribution of a variable of interest at two different points in time and to work with the attribute data of the composite map. For example, a national forest is interested in change of vegetation covers within a ranger district from 1980 to 1990. This task can begin with compiling two vegetation cover maps, one for 1980 and the other for 1990. The next step is to overlay the two maps so that the boundaries and attributes of vegetation covers are combined to form the output. Queries of attribute data in the composite map can reveal which polygons have undergone changes and which polygons have not.

Sometimes the role of a binary model is to help the building of a more refined model. This often happens at the beginning stage of a research project when data needed for the analysis are incomplete and the relationship between spatial features is still not clear. The *candystick* is a rare plant species that has been recorded in central Idaho and western Montana. The *candystick*'s favorable habitat cannot be defined on the basis of a limited number of *candystick* observations, although it is believed that the species occurs primarily in high-elevation lodgepole pine forests and on gentle to moderate slopes. A binary model based on the preliminary findings of the *candystick*'s habitat conditions can narrow the areas for field survey, which can then lead to more observations of the *candystick* and a better defined habitat model for the species.

Index Models

An index model evaluates attributes of a composite map, calculates index values from the attributes, and outputs a rank map based on the index values. Evaluation of attributes can take place at two different levels. The first is to rank the relative importance of each attribute with respect to other attributes and assign a weight to the attribute. The second evaluation is with the values of each attribute. Attribute values are usually scored using a 1 to 9 scale, 1 to 5 scale, etc. After evaluation of

attributes is made, an index model can be expressed as a linear equation, with an index variable on the left and the attributes and their weights on the right, and the index value can be calculated by summing the weighed numeric scores from each attribute. Both weights and ratings in an index model are ordinal data; this is why the output of an index model is a rank map.

There are variations to the above method for creating an index model, especially if the model is raster-based. One can use, for example, the lowest score, the highest score, or the most frequent score among the attributes in calculating the index value. These variations can be easily incorporated through local functions in raster-based GIS.

Index models are commonly used in suitability analysis and vulnerability analysis. Numerous examples are available in the literature; here we look at four examples. The first example is a study of prioritizing lands for conservation protection in Sterling Forest on the New York-New Jersey border. The study selected the following five parameters as input for the assessment:

- Development limitations due to soil conditions/steep slopes/flooding
- Non-point source pollution potential due to proximity to water/wetlands
- Habitat fragmentation potential due to distance from existing roads and development
- Sensitive wildlife habitat areas
- Aesthetic impact (visibility from the Appalachian and Sterling Ridge Trails)

The environmental cost/development constraints for each parameter were ranked from 1 to 5, with 1 being very slight and 5 being very severe. A grid was made for each parameter and the input grids were overlaid. The maximum value of any of the five input parameters was assigned to each cell in the output grid. Lands suitable for conservation protection were those cells with low values in the output grid.

The second example is the DRASTIC model developed by the Environmental Protection Agency (EPA) for evaluating ground water pollution potential. The acronym DRASTIC stands for the five parameters used in the model: Depth to water, net Recharge, Aquifer media, Soil media, Topography, Impact of the vadose zone, and hydraulic Conductivity. The use of DRASTIC involves rating each parameter, multiplying the rating by a weight, and summing the total score by

Total = summation of W_iP_i, i = 1 to 7

where P_i is the input parameter i and W_i is the weight applied to P_i.

The third example is a habitat suitability index (HSI) model. HSI models evaluate habitat quality by using attributes considered to be important to the wildlife species. The process of developing a HSI model involves interpreting habitat variables based on the life requisites of a wildlife species and converting these habitat variables into spatial variables that can be rated and used in a HSI model. Kliskey et al. (1999), for example, used the following equation to calculate habitat suitability for pine marten:

$HSI = sqrt \{[(3SR_{BSZ} + SR_{SC} + SR_{DS})/6] [(SR_{CC} + SR_{SS})/2]\}$

where SR_{BSZ}, SR_{SC}, SR_{DS}, SR_{CC}, and SR_{SS} are the ratings for biogeoclimatic zone, site class, dominant species, canopy closure, and seral stage, respectively. The model was scaled so that the HSI values ranged between 0 for unsuitable habitat to 1 for optimal habitat.

The last example is a model of human vulnerability to chemical accident. Hepner and Finco (1999) built their model by including the following demographic parameters:

- Total population
- Number of people younger than 18 years of age and older than 65 years
- Economic status as measured by household income
- Proximity to sensitive institutions such as schools, hospitals, and health clinics

Each of the parameters was rated from 0 to 10, with 10 being most vulnerable. The total population and sensitive population parameters were positively related to vulnerability, whereas the economic status parameter was inversely related to vulnerability. The assumption was that the economically disadvantaged had less access to information about chemical accidents and were therefore more vulnerable. Another assumption was 100 meters as the zone of influence for sensitive institutions. Finally, the four parameters were combined into a single measure of vulnerability.

The usefulness of an index model depends on the selection of parameters and the interpretation of rates and weights. To minimize subjective biases, rates and weights should be determined by consensus of a panel of experts. Additionally, it may be helpful to assess the sensitivity of the vulnerability measure to the weighting scheme by using different weight combinations in a sensitivity analysis.

Regression Models

A regression model involves a dependent variable and a number of independent variables. The output of a regression model is an equation, which can be used for prediction or estimation. Like an index model, a regression model can use map overlay operations to combine all the independent variables needed for the analysis. For example, in developing their habitat suitability model for red squirrel, Pereira and Itami (1991) started a database with 14 habitat variables. A logistic multiple regression analysis selected six statistically significant variables, including elevation, slope, aspect, and three categories of canopy. They were then able to calculate the probability (p) of squirrel presence for each grid cell in their study area by the equations

$Y = 0.002 \times \text{elevation} - 0.228 \times \text{slope} + 0.685 \times \text{canopy1} + 0.443 \times \text{canopy2} + 0.481 \times \text{canopy3} + 0.009 \text{ aspectE-W}$

$p = 1/(1 + \exp(-Y))$

Mladenoff et al. (1995) also built a logistic regression model to estimate the amount and spatial distribution of favorable gray wolf habitat. The study used vector data and required a fair amount of data processing. It began by creating wolf pack areas and nonpack areas in Wisconsin in polygon coverages. Wolf pack areas were home ranges derived from telemetry data of wolf location points. Nonpack areas were randomly located in the study area at least 10 km from known pack territories. Both pack and nonpack areas were then overlaid with the landscape coverages of human population density, prey density, road density, land cover, and land ownership. Using the composite coverages, stepwise logistic regression analysis converged on the following model:

$\text{Logit}(p) = -6.5988 + 14.6189 R$

where p is the probability of occurrence of a wolf pack, and R is road density. Probability values for occurrence of wolf presence can be calculated by

$p = 1 / [1 + e^{\text{logit}(p)}]$

where e is the natural exponent.

The logistic regression model, which was based on data from Wisconsin, was then applied to a three-state region (Wisconsin, Minnesota, and Michigan) to map the amount and distribution of favourable wolf habitat. The same model was used in a subsequent study (Mladenoff and Sickley 1998) to predict habitat suitable for wolves in the Northeast from New York to Maine. The subsequent study was based on raster data with a cell resolution of 5- x 5-km.

Process Models

Process models integrate existing knowledge into a set of relationships and equations for the purpose of quantifying physical processes. The output of a process model is an equation or, more likely, a set of equations, which can be used for prediction. Some process models use generalized equations

similar to those of index models, while others use sophisticated equations to describe the interaction among large amounts of environmental data. Process models are typically raster-based, and often use simulation models outside the GIS to work with complex and dynamic analysis.

The Universal Soil Loss Equation (USLE) is a well-known example of a generalized process model (Wischmeier and Smith 1978). USLE uses the product of six factors to predict soil losses for agricultural land:

$$A = R K L S C P$$

where A is the average soil loss in tons, R is the rainfall intensity, K is the erodibility of the soil, L is the slope length, S is the slope gradient, C is the cultivation factor, and P is the supporting practice factor. Of the six factors, the R, K, C, and P factors can usually be derived from available data on precipitation, soils, and land use. The L and S factors, which used to be estimated from field measurements, pose the major challenge to USLE users when the topography is complex and irregular. One approach is to combine L and S into a single topographic factor, which can then be calculated from the upslope contributing area of each cell and the slope of the cell.

The AGNPS (Agricultural Nonpoint Source) model analyzes nonpoint-source pollution and provides estimates of runoff water quality from agricultural watersheds (Young et al. 1987). AGNPS is event-based and operates on a cell basis. Using various input data, the model simulates runoff, sediment, and nutrient transport. For example, AGNPS uses a modified form of USLE to estimate upland erosion for single storms (Young et al. 1989):

$$SL = (EI) K LS C P (SSF)$$

where SL is the soil loss, EI is the product of the storm total kinetic energy and maximum 30-minute intensity, K is the soil erodibility, LS is the topographic factor, C is the cultivation factor, P is the supporting practice factor, and SSF is a factor to adjust for slope shape within the cell. Detached sediment calculated from the equation is routed through the cells according to yet another equation based on the characteristics of the watershed.

The SWAT (Soil and Water Assessment Tool) model predicts the impact of land management practices on water, sediment, and agricultural chemical yields in large complex watersheds (Srinivasan and Arnold 1994). It is a process-based continuous simulation model. Inputs to SWAT include land management practices such as crop rotation, irrigation, fertilizer use, and pesticide application rates, as well as the physical characteristics of the basin and subbasins such as precipitation, temperature, soils, vegetation, and topography. The creation of the input data files requires substantial knowledge at the subbasin level. Model outputs include simulated values of surface water flow, groundwater flow, crop growth, sediment, and chemical yields.

USLE, AGNPS, and SWAT are three examples of process models. The literature offers other process models on such topics as nonpoint source pollutants in the vadose zone (Corwin et al. 1997), landslides (Montgomery et al. 1998), and groundwater contamination (Loague and Corwin 1998).

Further Readings: Burrough and McDonnell, pp. 171-176; DeMers Chapter 13