Lecture Notes: “Land Evaluation”

by

David G. Rossiter

Cornell University
College of Agriculture & Life Sciences
Department of Soil, Crop, & Atmospheric Sciences

August 1994

Part 5 : Uncertainty

Disclaimer: These notes were developed for the Cornell University course Soil, Crop & Atmospheric Sciences 494 ‘Special Topics in Soil, Crop & Atmospheric Sciences: Land evaluation, with emphasis on computer applications’, Spring Semester 1994, and were subsequently expanded and formatted for publication. They are not to be considered as a definitive text on land evaluation.

Contents for Part 5 : “Uncertainty”

The world is full of *uncertainty*, much of which has a direct effect on the kinds of predictions we want to make in land evaluation. *No* serious land evaluation should be without some *estimate of uncertainty* in its results, even if in descriptive form. In this unit, we discuss various aspects of uncertainty, including the concepts of data and rule uncertainty, spatial variability of land characteristics, and 'fuzzy' logic. The emphasis is on how to describe and evaluate uncertainty, and how to determine and express the uncertainty, especially over space, of our predictions in land evaluation.

1. Uncertainty

‘Uncertainty’ refers to our imperfect and inexact knowledge of the world. We can distinguish two classes of uncertainty: *data* and *rule* (Eastman *et al.*, 1993). *Data uncertainty* has to do with our observations of nature or society: we are unsure of what exactly we observe or measure. *Rule uncertainty* has to do with how we reason with these observations: we are unsure of the conclusions we can draw from (even perfect) data.

No serious land evaluation should be without some *estimate of uncertainty* in its results, even if in descriptive form. This can be from a *sensitivity analysis* (as we discussed in dynamic simulation modeling and economic optimization), from *confidence limits* (as we discussed in statistical modeling) or from direct estimates of errors in the data as *propagated* through the model (as we will discuss in this lecture).

Note: the use of the term *error* to refer to uncertainty is technically correct but has a negative connotation. It should not be confused with ‘error’ meaning ‘blunder’ or ‘mistake’.

1.1 Data uncertainty: measurement and sampling errors

The true value of a parameter or datum is unknown (unknowable?). Two source of uncertainty are *measurement* and *sampling*. A good introduction to data uncertainty in the context of GIS is (Goodchild & Gopal, 1989).

1.1.1 Measurement uncertainty

Measurement ‘error’: there is always some uncertainty in any measurement, because of limited precision of the measuring device. This can usually be determined from the characteristics of the device and by repeated sampling and statistical characterization. These errors are usually (correctly) considered to be independent, normally distributed, and more-or-less exactly characterizable. In a land evaluation context, they are rarely significant, when compared with sampling errors (next section).

Example of pure measurement error: diameter at breast height (DBH) of a single tree, with a tape measure (for the diameter) and ruler (for the standard breast height).

1.1.2 Sampling uncertainty

Almost always we are only be able to measure a small part of the object of interest; this causes *sampling* error. These errors are more difficult to characterize and correct for than measurement errors. We must make (sometimes unjustified or un-testable) assumptions about our sampling strategy. The type and magnitude of these errors can be determined by repeated sub-sampling or by more exhaustive sampling.

Example of sampling error: we can only measure the DBH of a small proportion of the trees in the forest.

For both types of data errors, there is extensive statistical theory. In the usual case in natural resources or economic survey, we obtain an *expected value* (usually given by the mean) and a *variance* of a normal or Student's distribution, which can be used to express the uncertainty of the expected value. If the distribution of errors is known to have a non-normal form, we can estimate the parameters of a different distribution.

1.2 Representing the data values of a map unit

In land evaluation we are always describing and evaluating *areas* not *points*, so that the sampling problem is unavoidable. How should be characterize the data values for a map unit, and combine them with other (uncertain) values?

This discussion is most relevant to polygonal map units of variable size and shape (including polygons of a vector GIS), but can also be applied to cells in a grid GIS. It has great practical importance when applying statistical or dynamic simulation models to land areas: these require a set of specific numeric inputs, which must be determined from the available data.

1.2.1 Representing the map unit: (1) a single 'representative' value

Basic idea: Describe the datum by a 'representative' or 'typical' value.

Example: the attributes of a soil map unit is synthesized from many actual observations, from the surveyor's judgment, or from a single real profile which is considered to be 'typical'.

Advantage: in the absence of sufficient data to characterize the actual variability, a well-chosen single value usually gives a reasonable estimate of the *expected value*.

Disadvantage: no information about variability within the map unit.

Disadvantage: any information on how good is the estimate is lost when only one data point is used.

1.2.2 Representing the map unit: (2) a range of values (classes)

Basic idea: Instead of *one* representative value, describe the datum by *two* numbers, presumably spanning 'most' of the variability. We make no statement about most-probable values within the range.

Example: Slope class 'A' is defined as 0-3% slopes, Depth class 'moderately deep' is defined as 50-100cm depth.

This approach is usually employed in applications of the FAO method, including most ALES models.

Advantage: explicit statement of the 'entire' or at least the relevant portion of the range of the variable, useful for sensitivity analysis

Disadvantage: no single expected value, no information on distribution of values within the class

1.2.3 Representing the map unit: (3) a statistical distribution

Basic idea: Describe the datum by several parameters which depends on the *distribution* which the variable is assumed to have. We must also say which distribution. See (Morgan & Henrion, 1990) Chapter 5 for some common distributions. (Law & Kelton, 1991) is more comprehensive.

Advantage: completely describes the data values and their probability of being encountered.

Disadvantage: rarely is this theoretically and observationally justified.

Established by sampling from the map unit. (Forbes, Rossiter & Van Wambeke, 1982) Ch. 4 provide examples in the context of judging the adequacy of soil surveys.

Example: normal distribution with two parameters: mean and variance, can report as ± 1 (etc.) sample standard deviations from the sample mean, or ±z scores to achieve a given probability level.

Example: exponential distribution with one parameter: decay constant.

Example: uniform distribution with two parameters: the extremes.

Example: triangular distribution with three parameters: the extremes and the most probable value.

1.2.4 Representing the map unit: (4) a non-parametric distribution

Basic idea: Describe the datum by observed frequencies from some sample, without assuming any parametric distribution

Established by sampling from the map unit, but without assuming any underlying distribution of the data.

Advantage: completely describes the data values and their probability of being encountered

Advantage: no assumptions about the distribution, there may in fact be no single distribution in the map unit (e.g., if it is heterogeneous so that there really isn't a single 'population' to be parameterized)

Disadvantage: less statistical power and harder to combine with other values than a variable with known distribution.

Example: sample interquartile range, middle 8 deciles etc.

Example: fences of Box plot (-1.5 to +1.5 sample inter-quartile range from the sample median)

1.3 Data uncertainty: correlated variables

If several variables are *correlated* (i.e., not independent), it is not sufficient to describe their univariate distribution. Instead, we must determine their *multivariate* distribution, which in general involves the computation of a *variance-covariance matrix*. This is much more difficult to establish than univariate distributions.

The covariance structure is important because if we compute a function of several variables, the uncertainty of the result depends not only on the individual variances of the variables, but also on their covariances. We will see why in the section 'Error Propagation', just below.

1.4 Rule uncertainty

Even if all data values were known without error, the combination of variables to a result may be 'uncertain' in various senses:

(1) The true form of a *function* is unknown (e.g., logarithmic vs. polynomial yield response);

(2) In *expert judgment*, the result is uncertain (all facts being 'perfectly' provided, the expert still can't give an unambiguous answer).

1.5 Error propagation

If data values are described by probability distributions, and the combination of these into e.g. land suitability is by a *continuous function*, the classical theory of *error propagation* as developed by Gauss can be applied to determine the error (uncertainty) in the result. Thus we can give precise *confidence limits* in the results of a land evaluation, in the case where land characteristics are combined by continuous functions.

An example is a land evaluation based on predicted yields of an indicator crop, this yield being predicted by a multiple regression equation from a set of land characteristic values, where each land characteristic has a probability distribution.

References: A good introduction to error analysis in the context of measurements in physical sciences is: (Taylor, 1982). (Bevington & Robinson, 1992) is a more advanced presentation, with computer programs. Chapter 8 in (Morgan & Henrion, 1990) (written by Mitchell Small) is a more theoretical but well-illustrated introduction. (Burrough, 1986) Ch. 6 is a truly humbling recital of all the sources of error and how they can propagate in a GIS; we adapt some of this discussion later in this lecture. (Eastman, 1993) has a section on error propagation in the context of IDRISI. Heuvelink and collaborators (1993, 1993, 1989) have developed a theoretical and practical approach to this in the context of GIS; (Veregin, 1989) deals with error propagation in map overlay, in the context of uncertainty in spatial databases (Goodchild & Gopal, 1989).

GIS Software: IDRISI V4.1 module MCE (Clark University); ADAM (University of Utrecht)

1.5.1 Basic relation of error propagation

(For a detailed exposition, see (Taylor, 1982))

Suppose that various quantities (the so-called *independent variables*) $x_1, x_2, \cdots x_n$ are measured with 'small' uncertainty δx_i and that the uncertainties are *independent* of each other and *random*. These quantities are then used to calculate a result (so-called *dependent variable*) *z* by some function $z = f(x_1, x_2, \cdots, x_n)$. If the variables are *uncorrelated*, the uncertainty δz of the result is:

$$\delta z = \sqrt{(\frac{\partial z}{\partial x_1}\delta x_1)^2 + (\frac{\partial z}{\partial x_2}\delta x_2)^2 + \cdots + (\frac{\partial z}{\partial x_n}\delta x_n)^2} \qquad (1)$$

In words, the partial derivatives with respect to each variable are multiplied by the uncertainties, and the final uncertainty is the *sum by quadrature* of all these. This formula can be applied to all totally-differentiable functions. The disadvantage is that we must determine all the partial derivatives, although symbolic mathematics programs help derive these. Also, in practice, the partial derivatives are only *approximated*, usually by Taylor series expansions.

Warning! The requirements of independent and random errors are very important to this analysis.

We can avoid having to take the sum by quadrature, and assuming independent and random errors, if all we want is an *upper bound* on the error:

$$\delta z \le \left|\frac{\partial z}{\partial x_1}\right|\delta x_1 + \left|\frac{\partial z}{\partial x_2}\right|\delta x_2 + \cdots + \left|\frac{\partial z}{\partial x_n}\right|\delta x_n \qquad (2)$$

This is the *ordinary sum* of the error magnitudes, which is in general quite a bit larger than the sum by quadrature, so if we can assume independent and random errors, we should do so.

1.5.2 The basic relation for correlated variables

If some variables are correlated, the error may be more (positive correlation, i.e., the errors tend to reinforce each other) or less (negative correlation, i.e., the errors tend to cancel each other out) than is calculated by equation (1), because we must also compute the covariance and the partial derivatives in two dimensions:

$$\delta z = \sqrt{\sum_i \left(\frac{\partial z}{\partial x_i}\delta x_i\right)^2 + \sum_i \sum_j \frac{\partial x}{\partial x_i}\frac{\partial z}{\partial x_j}\delta x_i \delta x_j} \qquad (3)$$

In most practical problems, within a small neighborhood the covariances are small compared with the variances, and so are generally ignored. Note this formulation *still* assumes independent and random *errors*, even though the *variables* are correlated.

1.5.3 Error propagation for simple functions (uncorrelated variables)

In practice we often combine variables by simple functions such as addition, subtraction, multiplication and division. The basic relation (2) for uncorrelated variables simplifies for these functions.

Addition and subtraction: $z = f(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{k} x_i - \sum_{i=k+1}^{n} x_i$

In this case, all the partial derivatives equal 1, so that the error by sum of quadrature is:

$$\delta z = \sqrt{(\delta x_1)^2 + (\delta x_2)^2 + \cdots + (\delta x_n)^2} \qquad (4)$$

Or, the upper bound on the error (without assuming independent and random errors) is:

$$\delta z \le \delta x_1 + \delta x_2 + \cdots + \delta x_n \qquad (5)$$

In words, the *absolute* error of the results is the sum of the *absolute* errors of the individual variables. We can easily get a feel for the beneficial effect of assuming independent and random errors; suppose we have ten variables each with uncertainty $\delta x = 1$; formula (4) gives a final error = √10 ≈ 3.16, whereas formula (5) gives a final error ≤ 10.

Multiplication and division: $z = f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{k} x_i \div \prod_{i=k+1}^{n} x_i$

In this case, all the partial derivatives equal z^2/x, so that the error by sum of quadrature is:

$$\frac{\delta z}{|z|} = \sqrt{(\frac{\delta x_1}{x_1})^2 + (\frac{\delta x_2}{x_2})^2 + \cdots + (\frac{\delta x_n}{x_n})^2} \qquad (6)$$

Or, the upper bound on the error (without assuming independent and random errors) is:

$$\frac{\delta z}{|z|} \le \frac{\delta x_1}{|x_1|} + \frac{\delta x_2}{|x_2|} + \cdots + \frac{\delta x_n}{|x_n|} \qquad (7)$$

In words, the *relative* error of the result is the sum of the *relative* errors of the independent variables. Again, we can see the benefit of independent and random errors: suppose we have ten variables each with relative uncertainty $|\delta x/x| = 0.1$ (i.e., 10% relative error); formula (6) gives a final relative error = √0.1 ≈ 0.316 (i.e., 31.6%), whereas formula (7) gives a final relative error ≤ 1 (i.e., 100%).

1.5.4 How do we measure the uncertainty to be propagated?

The 'small uncertainty' δx etc. in the preceding formulas can be estimated in various ways. A very common method is to use the *sample standard deviation*, also referred to as the *root-mean square* or *RMS error*:

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x}$ is the sample mean and n is the sample size, as usual. Then we can rewrite (4) and (6) as:

$$s_z = \sqrt{{s_{x1}}^2 + {s_{x2}}^2 + \cdots + {s_{xn}}^2} \qquad (4')$$

$$s_z = \sqrt{{s_{x1}}^2 \cdot (x_2 \cdot x_3 \cdots x_n)^2 + \cdots + {s_{xn}}^2 \cdot (x_1 \cdot x_2 \cdots x_{n-1})^2} \qquad (6')$$

this latter derived from the general relation (1) because of the fact that:

$$\frac{\partial \prod_{i=1}^{n} x_i}{\partial x_k} = \prod_{i=1}^{k-1} x_i \prod_{i=k+1}^{n} x_i$$

In the simple *two-variable* case (e.g., a single IDRISI OVERLAY), (4') and (6') reduce to:

$$s_z = \sqrt{s_x{}^2 + s_y{}^2}, \quad s_z = \sqrt{s_x{}^2 y^2 + s_y{}^2 x^2} \qquad (4'', 6'')$$

where the two input variables are x and y, and the output variable is z. Some formulas are even simpler. For example, multiplication by a constant k with no error (e.g., π) results in the error $|k| \cdot s_x$, and addition of a constant k with no error does not affect the error.

1.5.5 Error propagation for simple functions (correlated variables)

Recall that equation (3) must be used instead of equation (1) when the variables to be combined are correlated. The term $\delta x_i \delta x_j$ of this equation corresponds to the *covariance* σ_{xy} between the two variables. This is what can make the error larger or smaller than for the uncorrelated case. In practice, σ_{xy} is estimated from the sample with the *sample covariance* s_{xy} . The derivations based on equation (1) must be modified by also computing the *partial cross-derivatives* which must be established for each case, i.e., we can't just compute a single relation for addition or multiplication, it depends on how the particular variables being added or multiplied co-vary.

1.5.6 Application to GIS

References: IDRISI: (Eastman, 1993); ADAM: (Heuvelink, 1993, Heuvelink & Burrough, 1993, Heuvelink, Burrough & Stein, 1989); general reference on map overlay: (Veregin, 1989); multiple sources of error as well as error propagation: (Burrough, 1986) Ch. 6.

The basic idea is to compute errors at the same time that we compute results when combining maps. The IDRISI command OVERLAY allows us to add, subtract, multiply, or divide maps; the error propagation can be determined by the formulas of the previous section.

The uncertainty for a map can be *uniform* for the entire map or, more realistically, it may be different for each cell. So in general we have *two* maps for each coverage: (1) the *expected value* and (2) its *uncertainty*. All the uncertainties must be expressed in the same relative terms, e.g. one standard deviation for normal variates.

Software: IDRISI modules MCE and SURFACE will automatically propagate RMS errors; ADAM (Heuvelink, 1993) (University of Utrecht) can handle more complicated functions.

Problem 1: propagation is strongly affected by correlations between *variables*. This is not accounted for in IDRISI; in ADAM the form and degree of the covariance must be known, and this is difficult to determine.

Problem 2: Strong positive covariance between *errors* of different variables can lead to exaggerated errors in the result (i.e., the result is less certain than it seems by applying the error propagation formulas), whereas strong negative covariance between errors can lead to cancellation of errors and a more certain result. The conservative position, if we can't assume independent and random errors, is to use equation (2) and its derivatives, i.e., ordinary sum instead of sum by quadrature. This leads to very loose error bounds.

Problem 3: some GIS operations do not have closed-form error propagation formulas. We would have to use (1) directly, and compute the numeric partial derivatives.

Problem 4: we have assumed *spatial independence*, i.e., only the values of land characteristics at each point are used to compute the result at that point. This ignores useful information from other nearby sampling points, if there is any *spatial dependence* in the land characteristic, as is often the case. We will consider this further in a subsequent lecture.

1.5.7 Example of error propagation: estimating soil loss with the USLE

(Burrough, 1986) p. 130-131 has a sobering example, which we present here with slightly different numerics. The exercise is to predict soil erosion in the Kisii district of Kenya using the 'Universal' Soil Loss Equation (USLE), a multiplicative index with the general form:

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

where A is the predicted annual soil loss in T ha^{-1} . Burrough explains how each independent variable is estimated, along with its uncertainty, from the best available local information. He obtains the following values and uncertainties:

Variable	Factor name	Estimated value	Estimated Standard Deviation	How estimated	Reliability
R	rainfall intensity	297 cm	72	FAO formula, from average annual precipitation, maximum-day in 2yrs precip, maximum 1-shower precip in 2yrs.	Low, due to sparse climate data for the area
K	soil erodability	0.1	0.05	measurements on soils outside the study area	Low; assume CV of 50% for relevant properties (e.g. silt and fine sand content) for soil map unit
L	slope length factor	2.13	0.045	regression from slope length	Good; 100±20m field size
S	slope gradient factor	1.169	0.122	parabolic regression from slope gradient	Good; 10±2% slope
C	cropping practices	0.63	0.15	from tables of similar crop cover in erosion experiments	Moderate; part of observed range of results
P	erosion control practices	0.5	0.1	from tables of similar practices in erosion experiments	Moderate; part of range of observed results

Using (5), i.e., assuming independent and random errors, the predicted soil loss is computed as:

$$A=23.0 \pm 14.8 \text{ T ha}^{-1} \text{ yr}^{-1}$$

which corresponds to lowering the soil surface by 7.8 ± 5.0 cm in 40 years (see calculation below). Since the uncertainty is a standard deviation, we can compute confidence limits. For example, 95% of the grid cells (in a grid GIS) having the combination of soils, rainfall, slope, and management practices specified here will have a lowering of the soil surface between 0cm (insignificant) and 16.1cm (substantial, probably will expose the subsoil to the surface) after 40 years. This is a serious practical difference for a planner.

(Calculation, supposing a bulk density of 1.175 g cm^{-3} for the surface soil: 1T ha^{-1} yr^{-1} × 10^{-4} ha^{1} m^{-2} × 10^{3} kg T^{-1} × 10^{3} g kg^{-1} × $(10^{-2})^{2}$ m^{2} cm^{-2} × $(1.175)^{-1}$ cm^{3} g^{-1} = 0.85... × 10^{-2} cm yr^{-1} = 0.34 cm 40yr^{--1}; multiply this factor by the 23 ± 14.8 T ha^{-1} yr^{-1} to obtain 7.82 ± 5.03 cm 40yr^{-1}, then round these to two significant figures: 7.8 ± 5.0 cm 40yr^{-1}. Note that we have not accounted for the variability in the bulk density for a soil map unit or the uncertainty in its measurement!).

(Calculation: 7.8 ± (5·1.65) = 7.8 ± 8.3; results less than 0 are unphysical, so we get the range 0-16.1cm. The factor '1.65' comes from the normal distribution; it is the value of Z above which only 5% of the deviates are expected to occur; since this is two-tailed the total deviates accounted for are 90%.)

Lessons from this example: (1) multiplicative indices are to be avoided if possible; (2) the number of factors in a predictive equation should be as small as possible; (3) identify the largest errors and try to control them. In this case since we are multiplying, the *relative* errors are what we compare; the *R* and *K* values were the least certain. We could improve *R* with better rainfall records, although even with perfect records there is great uncertainty in future rainfall. We could improve *K* with more homogeneous map units (narrower range of relevant soil properties), however, some map units are inherently heterogeneous at any scale.

Questions: are there any physical or logistical (sampling) reasons to suppose that the errors for any of these six variables are positively or negatively correlated? How about the variables themselves? How would these correlations affect the uncertainty of the result?

1.6 Monte Carlo simulation

In many situations, we can not fully analyze errors. Reasons: (1) the functional form is unknown; (2) the function's parameters are not known; (3) the function has no total differential or is too difficult to differentiate; (4) there is known to be covariance but we want a tighter error bound than given by the ordinary sum of errors (i.e., this may be too pessimistic).

In these cases we can determine the amount and distribution of errors by *simulation* (yet another use of the word). The basic idea is to *set up the model* (function), then *randomly vary the input variables* according to their probability distribution (note: if there are known joint distributions we can sample from these, thereby accounting for covariance), *compute* the model and *record the result*. If we do this enough times we get a frequency distribution of the result, from which we can compute its expected value and various statistics that quantify its error.

This is a very powerful and flexible technique. It is applicable to variables with and without known probability distributions (i.e., frequency distributions are acceptable). Without modern computers it would be impossible. Example software: @RISK. Tricky points: establishing the distributions of the input variables, designing a sampling strategy that will give reliable results in a reasonable number of simulations (typically on the order of 1,000), making sure we have examined enough simulations, quantifying the true distribution of the output errors.

In a grid GIS such as IDRISI, we can undertake a Monte Carlo simulation by introducing errors (calculated with the RANDOM module) into the data layers, then computing the model (e.g., a sequence of OVERLAY, SCALAR, and TRANSFORM), then examining the frequency distribution of the result.

1.6.1 Example of Monte Carlo simulation in IDRISI

Suppose we want to generate a slope map from a Digital Elevation Model (DEM) with a RMS error of 3 meters (this would be established by random sampling of the real landscape, or a very accurate map of it, vs. the DEM, and tabulating the errors). What is the RMS error of a slope map computed by IDRISI's SURFACE module?

It is not convenient to calculate the RMS error of the slope map (taken as a whole) directly, because the algorithm for computing slope depends on the neighborhood of each cell. However, we can simulate the RMS error:

(0) Generate an 'expected value' slope map from the 'expected value' DEM, with SURFACE.

Repeat steps (1) - (4) a large number of times (this can be automated with a batch file):

(1) Generate a map of the *simulation errors* using RANDOM with parameters 0m mean (i.e., no bias) and 3m RMS error. Each cell will have a different error, sampled from the normal distribution with parameters (0, 9).

(2) Add the simulated error to the expected-value DEM using OVERLAY (addition); this creates a modified DEM which *could* be the true DEM.

(3) Compute the slope map from the modified DEM with SURFACE

(4) Subtract the modified slope map from the expected-value slope map, with OVERLAY (subtraction). This is a single estimate of the error in the slope map; save it.

Now we analyze the slope-error maps as a group to establish the overall error:

(4) Compute the mean and standard deviation (RMS error) of all the slope-error maps, cell by cell. For example, the mean can be computed by repeated additions (OVERLAY) followed by a single division (by N, the number of simulations). This produces two maps: the expected error (bias) and RMS error, per-cell. These maps can be used to identify locations on the map where the uncertainty in the DEM had a large effect on the uncertainty in the slope; we could improve our efforts to make a good DEM in these regions.

(5) Compute the expected (mean) value of the by-cell means and RMS errors, using HISTO. The mean value of the means is the bias; if it is significantly different from zero, the SURFACE module has a systematic bias. The mean value of the RMS errors is the overall RMS error of the slope map.

Problem with this procedure: it assumes that there is no spatial correlation between errors in the DEM. In fact, we might expect positive correlation between the errors nearby cells, depending on how the DEM was constructed (e.g., if a contour line was mis-drawn in a certain area of the map, the elevations of all cells near the contour line will be similarly affected).

2. Fuzzy logic & continuous classification

Bibliography: A good general introduction to fuzzy logic is (Zimmerman, 1991), a bit more mathematical is (Klir & Folger, 1988), quite mathematical but with many interesting applications is (Kandel, 1986). (Burrough, 1989, 1992) give preliminary applications to soil survey and land evaluation.

2.1 Why use continuous classification in land evaluation?

When the *concept* being classified is not precisely defined (especially in the case of *linguistic* uncertainty), the techniques of fuzzy logic may be applicable. This allows us to compute and express land evaluation results when the land characteristics are not precisely measured, but instead are expressed by well-understood linguistic terms that can be quantified in a manner presently to be discussed. Fuzzy logic is an attempt to quantify ordinary expert discourse.

For example, it does not make sense to talk about the 'probability' that a particular year in the past was 'very wet', because we have the actual climate data for the year in question, and can determine exactly how wet was the year. What remains 'fuzzy' is what we mean by the term 'very wet'; so, instead of talking in terms of probability, we use the language of *possibility*, e.g., that the year in question can be considered 'very wet'.

Burrough uses the term *continuous classification* in preference to 'fuzzy sets'. As we will see, the techniques of fuzzy logic allow us to classify according to a continuous scale of membership. For example, a given year isn't either 'very wet' or 'not very wet'; instead it is *to some degree* 'very wet', this degree varying from none to completely.

(There is great controversy about whether fuzzy logic is simply a restatement of subjective probabilities. It seems to have practical application as it stands, so we will use its terminology.)

2.2 A fuzzy set and its membership function

A *fuzzy set A* over a universe of possible members *X* consists of members, a generic member being labeled as *x*, along with a *membership grade* for each member *x*, defined either by enumeration or by a function:

$$A = \{(x, \mu_A(x)) | x \in X\}$$

where the membership function $0 \le \mu_a(x) \le 1$. Intuitively, 1 = totally in the set, 0 = totally not in the set. Traditional *crisp* sets only allow values of 0 or 1, corresponding to *false/true, out/in, wrong/right* etc.

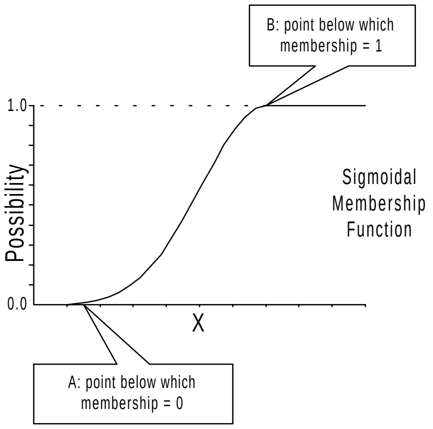
A good example is the concept of 'steep' slopes. We decide that 25% slopes or steeper are definitely 'steep' (fuzzy membership 1), and that 10% slopes or less are definitely not 'steep' (fuzzy membership 0). We must now decide how to qualify slopes between 10% and 25%; a typical (and arbitrary) choice is the *sigmoidal* membership function, which is defined by the relation $\mu = \cos^2\alpha$, where μ is the membership grade. We determine α from two points: the lower and upper limits of our concept:

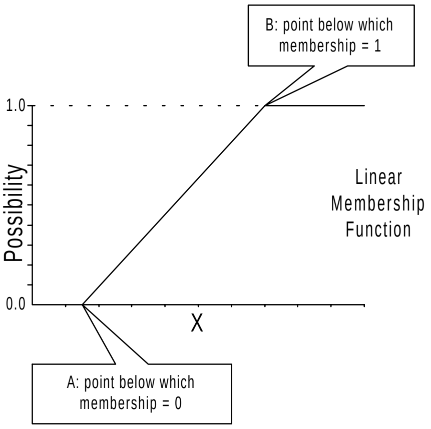
$$\alpha = \frac{\pi}{2} \cdot \frac{(x - l_1)}{(l_2 - l_1)}, x \le l_2$$

If $x \ge l_2$, then $\mu = 1$.

Note that we put the term 'steep' in quotes to show that we have a mental model of this term, but that it must be defined by a fuzzy function.

Membership functions are at the discretion of the analyst. The idea is to quantify the linguistic uncertainty. Here are two examples of membership functions for monotonic concepts, e.g. 'steep'.





The sigmoidal function is intuitively appealing, but the linear function may provide a reasonable approximation, and makes fewer assumptions about how our concept of membership changes in the range where the possibility is in the interval (0,1).

Other functional forms are used when the concept has a maximum membership at some value and less membership at both a higher and lower value. For example 'moderately deep' soils could be considered to have membership 1.0 between 60 and 80cm, tapering off to 0 at 40 and 100cm.

So using expert opinion and the wide variety of functional forms available, we can quantify any linguistic term.

2.3 Computing with fuzzy sets

The difficult point is in *combining* fuzzy sets in a way that is (1) mathematically consistent and (2) corresponds with the way we combine the linguistic concepts. For example, in rigid (Boolean) classification, we might say that a land area is 'very suitable' if it combines 'gentle' slopes and 'deep' soils. We would have to define these sets with crisp limits: e.g. 'deep' ≥ 1m, so 'not deep' < 1m; 'gentle' slopes ≤ 8%, so 'not gentle' slopes > 8%. Now in fuzzy classification, each site is not just 'deep' or 'not deep' etc., but some grade of 'deep', given by the membership function. How can we obtain a grade of 'suitable' from grades of 'deep' soil and 'gentle' slopes? The linguistic statement is 'A land area is very suitable for use X if it has both gentle slopes and deep soils'. There are many possible combinations with different desirable properties; see (Klir & Folger, 1988) and (Kandel, 1986) Ch. 6 for a theoretical treatment. In general the following combinations are used:

AND: the *minimum* of the two membership grades:

$$\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$$

OR: the *maximum* of the two membership grades:

$$\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$$

NOT: the *complement* of the membership grade on [0,1]:

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

CONcentration: the *square* of the membership grade. Corresponds to the linguistic modifier 'very'.

DILation: the *square root* of the membership grade. Corresponds to the linguistic modifier 'somewhat'.

INTensification: increases the contrast among members of a set: If a<0.5 → $2a^2$, otherwise $(1-2((1-a)^2))$, i.e. the middle (near 0.5) is eliminated, and membership grade is made more extreme. This is useful for eliminating middle ground.

NORMalization: ensure that at least some members of the set can be fully in it (i.e., membership value = 1). This operation is necessary *if* a derived fuzzy set should have at least one 'full member' (membership grade = 1), because the AND and other operations may leave the set without any full members.

(All these can be implemented by IDRISI operators such as TRANSFOR, SCALAR, and OVERLAY.)

The *big problem* with the practical application of fuzzy sets is that there is no single theoretical basis for the functional forms of the original variables nor for their logical combination, nor for their concentration, dilation and intensification with respect to our linguistic concepts. Normalization is a problem, and the final result may not be *convex* (i.e., may have peaks and valleys) even when it should not; other *ad hoc* techniques are used here.

3. Spatial variability

A fundamental fact is that land resources vary over space. In fact it is this pattern that makes land evaluation important. Furthermore, the variation is not random over space, but very often exhibits *spatial dependence*, i.e., knowing the value of a land characteristic at a certain point already provides information about non-sampled points nearby. In this lecture we discuss how to describe and deal with spatial variability in land evaluation.

References: (Oliver & Webster, 1991) is a nice introduction to spatial statistics for the soil scientist. (Isaaks & Srivastava, 1989) is an excellent practical introduction to the theory and practice of geostatistics and estimation. (Burrough, 1986) Ch. 8 discusses this from a GIS perspective. (Burrough, 1993) is an exhaustive review of the literature of soil spatial variability. (Vieira *et al.*, 1982) give a fascinating review of historical field experiments from the early 1900's, using geostatistical methods to expose the underlying spatial structure and draw new conclusions from old experiments. (Cressie, 1991) is a theoretical treatment which emphasizes the underlying similarity of a variety of spatial statistics.

3.1 Why natural resources vary in space

The causes of natural phenomena vary in space, therefore, so do the natural resources themselves. For example, considering the classic 'equation' of soil formation due to Jenny (see (Buol, Hole & McCracken, 1989)):

Soil = *f*(parent material, climate, organisms, topography, time)

we can imagine how these factors (other than time) vary over space: (1) a parent rock formation has differences in its grain size, chemical composition, fractures etc.; (2) the *climate* evidently varies over space (it doesn't rain everywhere on the earth at once); (3) organisms colonize certain areas preferentially; (4) the landscape does not have uniform topography. So it is only natural that the soil varies in space.

Climate also varies in space, because of the differential heating of the earth (tilted towards the sun) and its rotation, causing the general circulation of winds. Land masses also affect the general circulation (e.g., orographic effects), and there are local and micro variations in landforms, water bodies, vegetation etc. that cause climate to vary.

3.2 Key point: spatial dependence

The variability that we observe over space is usually *not completely random*, but has a *definite pattern*. In particular, we can often observe (or infer from our knowledge of the phenomena) that *nearby* locations will be *more similar* than *widely separated* locations. If I know it is raining at my home in Fall Creek, I can infer with some high degree of confidence that it is raining at Bradfield Hall about 1.5km distant. I have less confidence about the weather at Caldwell Field (3 km distant), even less about Dryden (15km distant) and almost none about the weather in Albany (about 200km distant). (Of course, my knowledge of the frontal patterns etc. might help me strengthen my inferences).

If knowledge about a variable at a sampling location allows us to predict with more-than-random accuracy the value at a ‘nearby’ location, we say that the variable exhibits *spatial structure* and that there is *spatial dependence* among the observations.

Note that knowledge of the underlying phenomenon may provide as much or more information than the relative location in space. For example, snow on South Hill will usually mean snow on East Hill but there may well be no snow in the Ithaca flats which is between; this because of the higher elevation and exposure to Lake Ontario. Here a *classification* of land areas based on those factors that we know to be related to climate (here, elevation and exposure) will give better predictions than just the location in space.

3.3 Dealing with spatial variability in land evaluation

There are various ways to account for spatial variability, as usual, each has its advantages and drawbacks. The fundamental problem is to *estimate* the value of a land characteristic at a *non-sampled location*, based on a set of existing samples. A related problem is to design an efficient sampling scheme to determine land characteristics over the entire study area with a prescribed accuracy.

3.3.1 Knowledge of the distribution of natural resources

We may know something about how the natural resource was caused. For example, orographic rain is caused by tradewinds being forced up mountains; the combination of elevation and aspect can be used to predict the rainfall, without nearby observation but from similar areas. The distance between the points is not so important as their similar environment. Another example is soil survey: if we can reliably map parent material and geomorphology, we can sample in *representative areas* of each map unit thus defined, and use those values for other delineations of the same map unit. Indeed this is the whole idea of soil survey: divide the soil cover up into more homogeneous areas.

Advantage: uses our knowledge about the causes of spatial variability in the natural resource

Disadvantage: no information on the distribution of values within each delineation.

3.3.2 Using a single nearby value

In the absence of information on the spatial dependence of a variable, and with no knowledge of any underlying causes, the most reasonable choice is to use the nearest measured value. A GIS can divide space up into *Theissen polygons*, also called *Voroni* or *Dirichlet* cells (can't they agree who invented it?), whereby each location in an area is associated with its *closest* observation point. (IDRISI module THIESSEN).

Advantage: uses actual data, does not make any assumptions about spatial structure

Disadvantage: abrupt change in values at the boundary between Theissen polygons, the tesselation depends on the sample points, polygons can have strange shapes, assumes no error within polygons.

3.3.3 Interpolating from nearby values

If we have reason to believe that the variable varies more-or-less *continuously* across space, it makes sense to *interpolate* at each point for which we want a value of the variable, from several 'nearby' points. The simplest case considers the *three* closest points, which define a response plane: $z = f(E,N)$ where E and N are the east and north coordinates (other systems could be used). Another reasonable choice is a *distance-weighted average* of any number of 'nearby' points; the advantage here is that closer points receive higher weight.

Advantage: uses more observational data

Disadvantage: assumes continuous behavior of variable, doesn't account for redundant observations, no way to determine the *error* of the estimate at an interpolated point.

3.3.4 Optimal interpolation or 'Kriging'

The problem with simple interpolation methods is that they can not account for *clustered observations*. Intuitively, if we have closely-spaced observations, they should not both be used in the distance-weighted formula, because their information is to some degree redundant. This concept can be formalized and solved by the use of *B*est *L*inear *U*nbiased *E*stimator ('BLUE') methods, more commonly known as *Kriging*, after the South African mining engineer who developed them. These methods provide optimal estimates for each point, and even better, the *error* of the estimate.

So, Kriging produces two maps for each variable: the estimated variable and its variance. These can be used directly in error-propagation GISs.

The mathematics of Kriging are beyond the scope of this course. See (Oliver & Webster, 1991) for motivation and a bit of the math, (Davis, 1986) pp. 383-405 for a good introduction with worked examples.

Advantages: mathematically optimal, provides an error estimate

Disadvantage: computationally intensive, problem of the ‘support’ (spatial area over which sample is taken vs. for which predictions are desired), assumes continuous variables; makes strong assumptions about the spatial structure, which must be inferred for each case.

Key issue with Kriging: A variogram must be estimated from sample data for each map to be produced by Kriging. There is no theoretical basis for a ‘best’ variogram, so its construction depends greatly on the analyst’s skills.

4. References

1. Bevington, P.R. & Robinson, D.K. 1992. *Data reduction and error analysis for the physical sciences*. 2nd ed. New York: McGraw-Hill. xvii, 328 pp. QA278.B63 1992 LC

2. Buol, S.W., Hole, F.D. & McCracken, R.J. 1989. *Soil genesis and classification*. 3rd ed. Ames, IA: The Iowa State University Press. xiv, 446 pp. S591 .B941 1989 Mann

3. Burrough, P.A. 1986. *Principles of geographical information systems for land resources assessment.* New York: Oxford University press. xiii, 193 pp. HD108.15 .B97 1986 Mann

4. Burrough, P.A. 1989. *Fuzzy mathematical methods for soil survey and land evaluation.* J. Soil Sci. 40: 477-492.

5. Burrough, P.A. 1993. *Soil variability: a late 20th century view.* Soils and Fertilizers 56(5): 529-562.

6. Burrough, P.A., MacMillan, R.A. & van Deursen, W. 1992. *Fuzzy classification methods for determining land suitability from soil profile observations and topography.* J. Soil Sci. 43: 193-210.

7. Cressie, N. 1991. *Statistics for spatial data*. Somerset, NJ: John Wiley & Sons. 928 pp.

8. Davis, J.C. 1986. *Statistics and data analysis in geology*. New York: Wiley. x, 646 pp. QE48.8 .D26 1986 Engineering

9. Eastman, J.R. 1993. *IDRISI Version 4.1 Update Manual.* Worcester, MA: Clark University Graduate School of Geography. 211 pp.

10. Eastman, J.R., Kyem, P.A.K., Toledano, J., and Jin, W. 1993. *Explorations in Geographic Information Systems, Volume 4: GIS and decision making*. Geneva (Switzerland): United Nations Institute for Training and Research (UNITAR). 112 pp.

11. Forbes, T.R., Rossiter, D. & Van Wambeke, A. 1982. *Guidelines for evaluating the adequacy of soil resource inventories*. 1987 printing ed. SMSS Technical Monograph #4, Ithaca, NY: Cornell University Department of Agronomy. 51 pp. S592.14 .F69 Mann

12. Goodchild, M.F. & Gopal, S. (ed). 1989. *The accuracy of spatial databases*. London: Taylor & Francis. xviii, 290 pp. + G70.2 .A17 Olin

13. Heuvelink, G.B.M. 1993. *Error propagation in quantitative spatial modelling: applications in Geographical Information Systems*. Netherlands

Geographical Studies 163, Utrecht: Faculteir Ruimtelijke Wtenschappen Universiteit Utrecht. 151 pp.

14. Heuvelink, G.B.M. & Burrough, P.A. 1993. *Error propagation in cartographic modelling using Boolean logic and continuous classification.* Int. J. of GIS 7(3): 231-246.

15. Heuvelink, G.B.M., Burrough, P.A. & Stein, A. 1989. *Propagation of errors in spatial modelling with GIS.* Int. J. of GIS 3(3): 303-322.

16. Isaaks, E.H. & Srivastava, R.M. 1989. *Applied geostatistics.* Oxford: Oxford University Press. 561 pp.

17. Kandel, A. 1986. *Fuzzy mathematical techniques with applications.* Reading, MA: Addison-Wesley. xiv, 274 pp. QA248 .K36 Engineering

18. Klir, G.J. & Folger, T.A. 1988. *Fuzzy sets, uncertainty, and information.* Englewood Cliffs, NJ: Prentice-Hall. xi, 355 pp. QA248.K49 1988 LC

19. Law, A.M. & Kelton, W.D. 1991. *Simulation modeling and analysis.* 2nd ed. New York: McGraw-Hill. 759 pp. QA76.9.C65 L41 1991 Engineering reserve

20. Morgan, M.G. & Henrion, M. 1990. *Uncertainty : a guide to dealing with uncertainty in quantitative risk and policy analysis.* New York: Cambridge University Press. x, 332 pp. pp. HB615 .M665x 1990 Olin

21. Oliver, M.A. & Webster, R. 1991. *How geostatistics can help you.* Soil Use Manag. 7(4): 206-217.

Abstract: “Geostatistics is basically a technology for estimating the local values of properties that vary in space from sample data. Research and development in the last 15 years has shown it to be eminently suited for soil and ripe for application in soil survey and land management. The basic technique, ordinary kriging, provides unbiased estimates with minimum and known variance. Data for related variables can be incorporated to improve estimates using cokriging. By more elaborate analysis using disjunctive kriging the probabilities of deficiency and excess can be estimated to aid decision. The variogram is crucial in all geostatistics, it must be estimated reliably from sufficient data at a sensible scale and modelled properly. Once obtained it can be used not only in the estimation itself but also to choose additional sampling sites, improve a monitoring network or design an optimal sampling scheme for a survey. It may also be used to control a multivariate classification so that the resulting classes are not too fragmented spatially to manage.”

22. Taylor, J.R. 1982. *An introduction to error analysis: the study of uncertainties in physical measurements.* Mill Valley, CA: University Science Books. 270 pp. QA275.T24 1982 Uris

23. Veregin, H. 1989. *Error modelling for the map overlay operation,* in *Accuracy of spatial databases,* Goodchild, M.F. & Gopal, S., Editor. London: Taylor & Francis. p. 3-18. +G70.2 .A17 Olin oversize

24. Vieira, S.R., Hatfield, J.L., Nielsen, D.R., and Biggar, J.W. 1982. *Geostatistical theory and application to variability of some agronomical properties*. Hilgardia 51: 1-75.

25. Zimmerman, H.-J. 1991. *Fuzzy set theory and its applications*. 2nd ed. Boston: Kluwer Academic. 399 pp. QA248.Z55 1990 LC