Lecture Notes: "Land Evaluation"

by

David G. Rossiter

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

August 1994

Part 3: Modeling

Contents for Part 3 : “Modeling”

This unit presents some approaches to fully-*quantified* land evaluation. Van Diepen (1991) and others (Beek, Burrough & McCormack, 1987, Bouma & Bregt, 1989) consider that land evaluation must be as quantified as possible in order to meet the demands of land-use planning. The biggest obstacle, as we will see, is lack of quantified *knowledge*, and secondarily, lack of quantified *data*.

The FAO Framework is fundamentally a *classification* system, working with classified land data, inferring classified land qualities, and resulting in suitability classes. Historically, this made sense because data was collected over map units. Nowadays, with the advent of computers and more-or-less continuous sampling methods (e.g., remote sensors), it is possible to collect, store and process large quantities of more-or-less *point* data in space and time (actually, the sample unit is some small area, but for convenience is usually called a sampling 'point'). This level of detail allows us to *model* the *response* of the land to various land uses, thus fulfilling the fundamental definition of land evaluation: "the process of assessment of land *performance* when [the land is] used for specified purposes" (Food and Agriculture Organization of the United Nations, 1985).

There are two principal modeling approaches: *empirical* (also called *statistical*) and *dynamic simulation*. As we will see, there is much empirical content in dynamic models as well.

A major use of modeling for land evaluation purposes is to predict *yields* (either average yields or time sequences). The 'value' of the land is directly reflected by its productivity. The modeled yield, along with a price for the product, gives the *gross return*, i.e., the 'return' or 'input to the producer' side of the cost-return equation. Since yields vary with management level (e.g., type and level of inputs, timeliness of operations), modeling yield requires a careful specification of the *input levels* of the farming system.

Modeling can also be used to predict some *land qualities* that are important components of yield, e.g., moisture supply, nutrient supply, radiation balance, as well as land qualities important for the land use but not directly affecting yield, e.g. trafficability and workability.

In this unit, we will first consider *statistical* characterization of the land-yield relationship, paying special attention to the problem of yield *prediction*. We will consider univariate and then multivariate methods. Then we will consider the problem of modeling the *dynamics* of the production system. Finally, we will consider the problem of how to obtain model parameters from field observations.

1. Statistical modeling for land evaluation

Basic idea: quantify *observed* relations and use these to *predict* future situations. It will not work unless there is sufficient *data* on which to base the *statistical inferences*, so is not appropriate for new land uses or areas with insufficient samples. For land evaluations of established land uses with sufficient historical or experimental data it can be quite useful, in fact often the preferred method.

Example of a reasonable application of this method: to predict yield of maize grain under typical management conditions in NY State, as a function of climate and soil characteristics. There is a complete, detailed soil map of the state, a huge number of reliable laboratory analyses of soil fertility levels, detailed long-term climate records in a dense network, accurate records on management practices and yield levels by field, detailed and accurate market price records (both inputs and outputs). Maize grain is widely grown. For purposes of tax equalization and assessment, a statistical prediction of the time-series distribution of likely yield levels of a commonly-grown field crop such as maize would be a reasonable approach to a an equitable tax system.

Definition of statistics (Steel & Torrie, 1980) p. 2: "Statistics is the science, pure and applied, of creating, developing, and applying techniques such that the uncertainty of inductive inferences may be evaluated." Note the emphasis on *inductive inference*: we have made some observations and now wish to generalize them. Also the *uncertainty* (or, degree of confidence in the inference, and by extension, of predictions based on the inference) must be assessed.

The most common application of statistical modeling in land evaluation is *yield prediction*, so we will look at various statistical tools in this context. Keep in mind that the same tools can be used to model individual land qualities.

There are many excellent textbooks on statistics; especially recommended is (Steel & Torrie, 1980) for agricultural experiments and (Davis, 1986) for general descriptive statistics and a very accessible introduction to regression and multivariate methods. A comprehensive, practical approach to regression is (Draper & Smith, 1981).

1.1 Yield estimation

Yields evidently vary, and so do many production factors, both natural resources (e.g. soil characteristics) and management options (e.g. amount of fertilizer), and there are good reasons (both observational and theoretical) to believe that the variation in yields has at least some of its underlying cause in the production factors. Statistical methods have been used since the beginning of land evaluation (e.g. (Simonson, 1938)) to quantify these relations

and to determine how much of the observed variability can be explained by the production factors (and how much remains to 'chance', i.e. unexplained).

The 'production factors' mentioned above are roughly equivalent to *land characteristics* as we have defined them.

The *dependent* or *effect* variable ('y') is *predicted* by one or more *independent* or *causal* variables ('x'). So the basic idea has been: *observe* the effects (e.g. yields), *record* the supposed causes (e.g. amount of fertilizer applied, amount of a chemical element in the soil), and *infer* the causal relation between these by statistical inference.

There are two kinds of datasets:

1. *controlled* (usually from field experiments): the experimenter controls the levels of the independent variable

2. *observed* (usually from surveys): the levels of the independent variables are not controlled, only observed.

Controlled data is a more reliable way to understand cause and effect. Observed data is a cheap way to acquire data with a wide (numerical) range of each causal factor. However, nature may not provide us with the full range of effects that we could impose in an experiment (e.g. can only observe soil N from 40 to 120 kg ha^{-1}; but if we impose fertilizer treatments we could extend the upper end of the range to 300 kg ha^{-1}).

In land evaluation, controlled experiments can be used to choose diagnostic LCs and their relation to levels of LQs, although this is usually the job of applied agricultural researchers. Observed data is typically used directly in land evaluation, in an attempt to predict yield from observed LCs.

Strictly speaking we observe *correlations*, e.g., more fertilizer is correlated with greater yields in a certain range. We can turn this into a *causal* relation in two ways: (1) if the fertilizer is applied under the control of an experimenter, and we observe the yield at some later time, we have a *temporal sequence*, so that the yield could not have caused the level of fertilizer, any causal relation must be from fertilizer to yield. Still, to establish that the fertilizer actually affected yield, we need (2) some knowledge of the role of mineral elements in plant nutrition to establish a plausible mechanism.

Advantages of statistical methods of land evaluation:

1. They use actual observations either random or imposed ('the past is the key to the future');

2. More observations should lead to more reliable predictions, in a predictable and quantifiable manner;

3. The predictions are on a continuous scale and allow for fine distinctions with enough data;

4. We can consider input levels as well as natural resources as predictor variables;

5. There is a rich set of *analytical methods;* and

6. Each prediction comes with an estimate of its *reliability*.

Disadvantages:

1. The form of the statistical relation is not obvious *a priori* and several forms may give similar results in terms of goodness-of-fit;

2. They make assumptions about the distribution of the predictor and response variables that may be impossible to verify and in fact may be obviously a convenient fiction. Problems of unusual observations (outliers or a valid member of the sample?);

3 Extrapolation to values of the independent variable outside the domain of calibration is not justified;

4 It is not obvious how to define the sample space, and errors here invalidate the inferences;

5 It is not obvious which independent variables should be included in the relation; and

6 The precision of the statistical relation may not useful enough for meaningful predictions (especially with observational data).

1.2 What is the sample population?

The *population* is all possible values of a variable. 'Possible' is to be understood as limiting the *universe* or *sample space*. In land evaluation we definitely must specify the *geographical area* about which we want to make a statistical statement.

We may also choose to divide the area according to some factor that greatly affects the *degree* or *form* of response.

Basic rule: if the difference in response across an area is one of *degree*, don't divide the area (the regression equation will account for the factor); if the difference in response is one of *kind*, divide the area into sub-populations and analyze each one separately (or use 'dummy' variables if the form or response is the same but the degree differs).

(See example of rainfall vs. yield, below).

1.3 Simple linear regression

The simplest approach to statistical modeling is to assume a *linear* relation, within a certain range, between a *single* independent predictor (hence a 'simple' regression) and the dependent variable 'yield'. Examples of predictors: fertilizer level, moisture supply. This makes sense in the 'usual' response range: it is statistically impossible to distinguish the 'flat' part of an exponential or a higher-order polynomial from a straight line. Also, a single predictor may be enough to explain most of the observed response.

The methods of linear regression are very well-understood (Draper & Smith, 1981, Webster & Oliver, 1990), (Steel & Torrie, 1980, ch. 10). Assumptions and pitfalls are numerous. See especially (Webster, 1989).

Basic relation: $y = b_0 + b_1 x$, with two *parameters*: b_o is the *intercept* (e.g., yield level with no input) and b_1 is the *slope* or *gradient* of the linear relation. Can also be written in terms of the mean: $y = \bar{y} + b_1(x - \bar{x})$.

(Note: the intercept must logically be zero or positive, since we never have negative yield, but in fact the statistically-best regression may have a negative intercept; this relation is not to be used in this range!)

The parameters are determined uniquely from the data once a *goodness-of-fit* criterion is established; this is almost always to *minimize the sums of squares* of the deviations between the observed and predicted (by the regression equation) values of the dependent variable (although there are many other possible estimators). The easiest way to compute, and the most illuminating, is from the sample *covariance* between the dependent and independent variables:

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Then the slope is computed as $b_1 = s_{xy} / s_x^2$. The important thing to notice here is that the slope increases with increasing covariance, i.e., when large deviations from the mean in the dependent variable *y* are matched with large deviations from the mean in the independent variable *x*. If the deviations are in the same direction numerically, the covariance and slope are positive, otherwise they are negative.

The degree of association (*goodness-of-fit*) is best estimated by the *correlation coefficient*:

$$r = \frac{s_{xy}}{s_x s_y}$$

This ranges from -1 (perfect negative relation) to +1 (perfect positive relation). The square, r^2, is called the *coefficient of determination*, ranging from 0 (no explanation) to 1 (perfect explanation). The *statistical significance* of the regression depends on r^2 and the sample size, because the following statistic is distributed as Student's *t* with (*n*-2) degrees of freedom.

$$\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

Note that a high r^2 leads to a small denominator and large numerator, hence a large value of *t*, hence high significance. Also, this is a *signed* statistic, so the *t*-test is two-tailed. Also notice that the significance increases as the square root of the sample size, so that, roughly speaking, to double the significance we would have to quadruple the sample size.

Caution! a large slope b_1 does not necessarily indicate a strong causal relation; the *statistical significance* of the coefficient of determination r^2 is what matters.

We can also estimate the *confidence limits* for the population mean, for the slope of the regression, and for any prediction, using any *risk level* that we wish. These all require that we calculate the standard error of the estimate:

$$s_{y \cdot x} = \sqrt{\sum (y - \hat{y})^2 / n - 2}$$

where $\hat{y}$ is the predicted (estimated) value (on the regression line). Note that the closer is the actual to predicted value, the smaller will be the standard error of the estimate. This must be multiplied by the appropriate *t* to obtain a confidence limit for a given prediction. This is an advantage of linear regression: it provides a confidence in any prediction. Again, the relation of sample size to significance is as the square root.

1.4 Calibration vs. validation (postdiction vs. prediction)

The process of fitting a regression equation to observed data is *calibration*, i.e., we are calibrating the general linear regression model with specific values of the parameters. This yields a goodness-of-fit measure such as r^2. This measures how well we were able to calibrate from the *available data*.

> Another name for calibration is postdiction (as opposed to *pre*diction, see below), from the Latin 'post' (after) and 'dicere' (to say). We have made a statement about the *past*, i.e., what we observed in our original experiment or data set. We have made statistically-significant statements about its mean, variance, and degree to which the observed variance was explained by the calibrated regression.

If the sample was truly *representative* of the desired sample space, we would expect to obtain the same parameters, within experimental and observational error, in similar repeated studies.. However, we can't be sure that the sample with which we calibrated is representative of the situation we want to predict. In other words, the r^2 we obtained by calibration may not apply to the situation we had in mind for *prediction*.

A better measure of the predictive success of a regression is a *validation*. Here we use the equation to predict results for a *second* set of observational data, and compare the observed vs. predicted. The regression of the validation data vs. predictions should lie along a 45° (diagonal) line, within error limits. If not, the original model is not valid. If so, we can get an estimate of the predictive power from the r^2 of the *validation* regression (not the r^2 of the calibration regression).

> Another name for validation is prediction (as opposed to *post*diction, see above), from the Latin 'prae' (before) and 'dicere' (to say). We have made a statement about the *future*, i.e., what we expect to observe in later experiments or data sets. We have made statistically-significant statements about the amount of variance in a future data set that we expect to be able to explain by the calibrated regression.

One method of validation is to regress the predicted on actual variables, and test to see if we can detect an intercept $b_0 \neq 0$ or a slope $b_1 \neq 1$. If either of these can be proven, the original regression is not a valid predictor.

1.5 Problems with linear regression

1. Completely unjustified to *extrapolate*. For example, suppose that a regression of yield on rainfall was determined in a zone where the crop is adapted. If we go to a more arid zone, there is no assurance that the relation will hold, so that yields will gradually decrease with decreasing rainfall. In fact could be that the yield will immediately drop to zero (threshold effect). If we go to a more humid zone, there is no assurance that the yield will continue increase at the same rate. It could be that the yield will increase at a decreasing rate (see Mitscherlich's equation, below) or even decrease.

2. We may be attempting to model various causes with one equation, even when restricted to one predictor variable. The moisture example is good: if we have data from a large range of moisture regimes, we may be seeing multiples effects: plants need water to grow, but excess favors plant diseases (or at the best is wasted). In this case it would be best to divide the universe into *subpopulations*: high-, medium-, and low-rainfall regions, and develop linear regressions for each of these separately: a *piecewise* function. (Another solution is to perform a multiple regression with *dummy* variables to differentiate the subpopulations.)

3. The underlying relation may not be linear even in the zone of calibration, yet a linear fit may be statistically significant. Of course, a transformed fit would fit better, see next section. This is not much of a problem for prediction as long as we don't extrapolate beyond the zone of calibration.

4. There are many cautions regarding *outliers* and especially *observations with a large influence* on the results. Observations with high *leverage* on the equation can easily cause meaningless results, both in the equation itself and the computed level of significance.

1.6 Nonlinear regression

Often there are experimental results and/or theoretical considerations which suggest that the response is *nonlinear*, for example, an observed *diminishing returns* at high levels of a predictor, or *increasing returns* at low levels.

Many nonlinear effects can be *linearized*: new variables can be created from the original variables, and the resulting regression on the new variables is linear. Examples are polynomial regression and one-variable transformations (see next two sections). In this case we say that the regression is intrinsically linear even though it is non-linear in the original variables.

Some effects can not be linearized, in particular, interactions (such as cross-products) between two variables; these are intrinsically non-linear. These effects are rare in land evaluation (as opposed to agricultural experiments), because the datasets are usually too noisy to detect significant interaction effects.

1.6.1 Nonlinear regression: polynomials

Many kinds of nonlinear effects can be accounted for by the use of *higher powers* (than the first) of the predictor variable.

Here, the basic relation becomes: $y = b_0 + b_1 x^k$, which has the same parameters as simple linear regression but which is linear in some *power* 'k' of the predictor. Most commonly *k=2* (quadratic) to account for parabolic curve (e.g. concave upwards or downwards) or *k=3* (cubic) for curves with an inflection point (zero slope) in the range of interest.

Most commonly, higher powers are used as part of a multiple regression, see below.

1.6.2 Nonlinear regression: transformed variables

Another way (other than with powers) of dealing with suspected or measured nonlinear responses is by assuming that the underlying relation between predictor and result is linear if one or both of the variables is transformed to another scale. The analyst may try any number of *transformations* in an attempt to *linearize* the relation. Ideally these should correspond to some *theoretical basis*. If the transformation succeeds, the predictor is said to be *intrinsically linear* because we can now fit a linear equation (in the parameters) to the transformed data.

Why transform? Let me count the reasons...

(1) For *significance testing* in regression, it is required that experimental (observational) errors be independently and normally distributed with a *common variance*. Often we can see that this is not the case; certain transformations will restore these conditions.

(2) There may be some *physical* or *experimental* reason to expect a non-linear relation. A good example is hydrogen ion concentration in soil solution vs.

plant growth: generally the logarithm of the concentration (pH) shows a linear relation.

(3) A plot of the *residuals* (fitted vs. observed values of the dependent variable) show a *pattern* when plotted against the predictor. For example, increasing residual with predictor implies a quadratic relation.

(The engineer's golden rule: "Every relation looks linear when graphed on log-log paper".)

Some common transformations:

Logarithmic transformation: When standard deviations are proportional to mean, this transformation equalizes the variances.

Square root transformation: When enumeration (small integer) data follow a Poisson distribution where the mean and variance are equal; taking the square root of the response variable restores normality. This is equivalent to squaring the predictor variable.

Angular (inverse sine) transformation: Transforms percentile data to normal.

1.6.3 An example of nonlinear regression: Mitscherlich's equation

(see (van Diepen *et al.*, 1991) p. 141, (Wild, 1988) (p. 59-60), original work is (Mitscherlich, 1909))

This (optional) section shows how nonlinear response may be expected from theory and measurement, and how a transformation can linearize it.

In 1909, Mitscherlich formulated a general non-linear equation to describe yield response, based on *theoretical* considerations and careful *measurements*:

$$y = A \cdot (1 - e^{-cx})$$

where: y is the predicted yield (dependent variable), A is the maximum obtainable yield under perfect non-limiting conditions, x is the amount of the 'growth factor' (independent variable), c is a proportionality factor (parameter) which controls the steepness of the relation.

(From this equation we see that zero input predicts zero output; this assumption can be relaxed by adding another factor, '$+b$', which is the zero-input yield. In a sand or hydroponic culture the assumption of zero output if an essential element is completely missing is justified by experiment.)

Derivation of Mitscherlich's equation:

This is an integral form of the differential equation:

$$\frac{dy}{dx} = (A - y) \cdot c$$

which relates the *rate* of response dy/dx of the yield with respect to the factor to the *difference* between the maximum obtainable yield A and the actual yield y. The incremental increase in yield is proportional (with coefficient c) to the decrement from the maximum, i.e. the closer to the maximum yield, the less will be the increment in yield from an incremental amount of the factor. In other words, the 'law of diminishing returns'.

Linearlization of Mitscherlich's equation:

Transform by the logarithm and rearrange:

$$\begin{aligned} y &= A \cdot (1 - e^{-cx}) \\ y / A &= 1 - e^{-cx} \\ (y / A) - 1 &= -e^{-cx} \\ 1 - (y / A) &= e^{-cx} \\ \ln(1 - (y / A)) &= -cx \end{aligned}$$

which is simply a linear equation: $y'=c'x$, where y' is the transformed dependent variable normalized by the achievable yield A, and c' is the steepness parameter $-c$.

So, to calibrate (i.e., to determine the parameter c), we need a dataset of (x, y) pairs and an estimate of A, then we can use linear regression on the linearized form (logarithmic transform).

2. Multivariate statistical methods for land evaluation

The previous section explained simple regression on *single* predictor variables. It is rare that a single predictor variable by itself is very successful for yield prediction. In this section we consider the usual *multivariate* case when yield is predicted from several factors.

2.1 Multiple linear regression

Other than in a controlled environment or a very special situation in which only one factor is limiting, several factors usually limit plant growth and yield.

Many attempts have been made to quantify this relation, usually using *multiple regression*. At its worst this exercise results in a meaningless monster equation (exacerbated by easy availability of user-friendly computer programs for multivariate analysis and cheap computers), at its best it integrates the most important single factors and their interactions in a single predictive equation. General form:

$$Y = b_0 + \sum_{i=1}^{n} b_i \cdot x_i$$

where the x_i are the predictor variables. The parameters b_i are estimated by least-squares.

Very often *polynomials* are considered in multiple regression, either in the single-predictor case where several powers of the same variable are used to predict:

$$Y = b_0 + \sum_{k=1}^{m} b_{1k} \cdot x^k$$

or in the more general case where each of several predictors can have powers:

$$Y = b_0 + \sum_{j=1}^{n} \sum_{k=1}^{m} b_{jk} \cdot x^k$$

In both these formulas, some of the b_{*k} can be zero, if that power of the predictor doesn't enter in the equation. For example:

$$Y = b_0 + b_{11} \cdot x_1 + b_{13} \cdot x_1^{\,3} + b_{22} \cdot x_2^{\,2}$$

Here, x_1 is a predictor in both its first and third powers, and x_2 is a predictor only in its second power.

A good example of this approach for land evaluation is (De la Rosa, Cardona & Paneque, 1981), which was later incorporated into an automated land evaluation system for Mediterranean countries (De la Rosa *et al.*, 1992).

Another typical example, this from the great State of New York, is (Olson & Olson, 1986). They conceive of corn yield as being determined by a production function: *Y* = *f*(rainfall, temperature, management, site, topography, chemical characteristics, physical characteristics, mineralogy, biological organisms, time), which seems to cover most of the possibilities. They fixed the management level, essentially the level of fertilization, liming, pesticide use and tillage (in land evaluation terminology, this corresponds to evaluating for a *specific Land Utilization Type*) and approximated the other factors by measured variables (in land evaluation terminology, *land characteristics*). For example: the 'rainfall' conceptual variables was approximated by the actual measured variable 'total yearly rainfall', 'temperature' by 'growing degree days', and 'site' by 'drainage class (depth to mottles)'.

> Note: there is plenty of discretion in the choice of variables. For example, why not 'growing season rainfall' instead of 'total rainfall'? For growing degree days, there are several definitions; which to use? For a soil variable such as exchangeable bases, to what depth? or should values for each horizon be included as separate variables? One answer is to try lots of different variables and see which are better predictors. Another approach is to use principal components on all these variables (see below). Better still is to have some theoretical basis for your decision.

(Olson & Olson, 1986) measured yields and land characteristics for five sites in New York, for periods from 2 to 19 years, and attempted to fit the best multiple regression in the non-transformed variables. The best-fit equation was:

$$Y = -3156 + 116 \cdot rainstor + 485 \cdot temp + 9 \cdot bases + 45 \cdot ocarb$$

where *Y* is the yield in kg grain ha^{-1}, *rainstor* is an available-rainfall index in cm computed by another multiple regression (Olson & Olson, 1985), *temp* is the growing degree days (sum of °F greater than 50°F), *bases* is the sum of basic cations in meq m^3, and *ocarb* is the amount of organic carbon in g m^3.

The coefficients depend on the units of measure of the predictors and by themselves don't show how important is each factor; this is revealed in the stepwise regression (below). I.e., a large coefficient, by itself, is not necessarily more important numerically to the final result.

Note the negative intercept! This implies that corn yield would be negative, (i.e., we'd have to plow over 3T of corn grain into a fallow field to satisfy the equation!) which is absurd; however in the range of calibration of the equation the inclusion of this parameter gives a superior fit to a model with no intercept. The authors don't seem to notice this nor explicitly state the range of calibration (for which, presumably, yields would always be non-negative).

There are various ways to determine such an equation:

1. Try every possible combination of variables ('*all possible* multiple regressions') and pick the one with the best fit. Problems: (1) a lot of work, (2) several equations may be equally good, (3) the best equation may not be very understandable because it has too many predictor variables.

2. Use the best single predictor (linear regression), then keep adding predictors one at a time, always adding the one that most improves the fit, in order until the fit does not significantly improve (i.e. the improvement in r^2 is below a certain threshold) ('*forward* multiple regression'). Advantages: uses the minimum number of predictors, computation is easy, uses the least number of variables necessary to explain the result.

3. Do (1) then eliminate the least-important variables one at a time until the fit becomes significantly worse ('*backward* multiple regression').

4. Like (2) but re-examine all variables at each step, so that one may go into and then out of the equation ('*stepwise* multiple regression'). Generally the best compromise.

The problems with methods (1), (2), and (3) all stem from the *correlation* of predictor variables (see below).

The important point to remember is that very rarely is one multiple regression clearly 'best'. There is usually judgment and even arbitrariness in the process of selecting variables.

> In the example of (Olson & Olson, 1986), a single regression on *rainstor* had an r^2 of 0.35, adding *temp* increased this to 0.60, adding *bases* to 0.64, and adding *ocarb* to 0.66. This is about as good as these sorts of equations get. This does *not* mean than *ocarb* used as a single predictor in linear regression would have an r^2 of 0.02! If it were used first, it would probably have a higher r^2.

The *average* prediction may be better than the r^2 would seem to indicate. I.e., from the average values of the predictor variable to the average yield. Thus in long-term land evaluation (strategic planning) these methods may be acceptable even if they are not acceptable for year-to-year or near real-time tactical planning.

> In the example, the average difference in predicted and actual yields for a plot with 19 years of data was -194 kg ha^{-1} for a relative error of only 3%, even though the r^2 of the predicted-vs.-actual regression was only 0.52.

2.2 Problems of multiple linear regression

1. *Physical significance* of predictors, i.e., does the equation have any explanatory power? Some of these are fairly clear: for example, we know that warm temperatures favor higher biological activity, so there may be a fairly direct effect of growing degree days on maize yield, and its presence in an equation may be a realistic reflection of its biological effect. Others are

much less clear, for example, the effect of soil organic matter on yield. Even if this is shown statistically to affect yield, how does this occur? There are many possibilities: (1) improved soil structure leading to easier root growth, better aeration, faster infiltration, (2) higher water-holding capacity, (3) direct supply of plant nutrients, (4) faster soil warming in spring due to black color... (the benefits of soil organic matter are many). Some or all of these may be affecting yield, even synergistically. It is very unlikely that there is a simple relation between soil organic matter and yield, so its presence in an equation is not too helpful to our understanding.

2. *Correlation of predictors*, hence almost arbitrary choice of predictors. Example: minimum temperature, maximum temperature for a month. Both may be good predictors, and usually they are highly correlated. Once one enters the equation the other one will not. Which to use? One may be the correct physical predictor due to its effect on plant growth, the other may not really affect the plant but is highly correlated with the true predictor. The idea is to have a *meaningful* equation, not just a statistically-significant relation. Why? (1) aesthetics, (2) science, (3) probable predictive power.

2.3 Multiple nonlinear regression: accounting for interactions

To complicate matters, production factors will often *interact* either positively (synergistically) or negatively (compensatory). These interactions must be determined for each case. Linear regression can not account for these. The resulting equations are *intrinsically nonlinear*.

Example of *synergism*: High levels of one nutrient have little effect until other nutrients are at comparable levels.

Example of *compensation*: high nutrient levels lead to more efficient water use, so that the effect of added water at high nutrient levels is less than that at low nutrient levels.

General form of the equation, only considering two-factor interactions:

$$Y = b_0 + \sum_{i=1}^{n} b_i \cdot x_i + \sum_{i=1, j=1}^{i=n, j=n} b_{ij} \cdot x_i \cdot x_j$$

Note the interaction term, this is where the non-linearity occurs. These coefficients are related to the *covariance* between predictor variables.

In an *analysis of variance* (ANOVA), there would be one or more significant *interaction* terms. So one way to determine if there are interactions is to first do an ANOVA on the same data set. If there are no significant interactions, proceed with the multiple regression. If so, determine which interaction terms to compute for possible inclusion in the regression.

2.4 Principal components

(See (Webster & Oliver, 1990) for a good introduction in the context of soil science, (Davis, 1986) for a very accessible explanation of theory and computation in the context of geology.)

Principal component analysis is a method for transforming a multidimensional space to another one of the same dimensions (i.e., same number of axes or variables) but with two very interesting and important properties:

1. The first *component* (or *synthetic variable*) explains the highest proportion of the total variance, the second variable the second-highest proportion, etc. Therefore, the less significant variables (dimensions) can usually be discarded as insignificant noise, thereby *reducing* the effective dimensionality of the problem.

2. The axes are *orthogonal* (mutually perpendicular) in multi-dimensional Euclidean space, so that the principal components are completely *uncorrelated*.

Mathematically this is accomplished by finding the *eigenvectors* (synthetic variables) and *eigenvalues* (their variances, i.e. importance) of the *variance-covariance matrix* of the predictor variables. (These are sometimes called *characteristic* values and vectors, instead of using the German root 'eigen'.) The transformed matrix has zero covariances, i.e. it is diagonal with the eigenvalues on the diagonal in descending order. Major-league magic!

This is a theoretically-satisfactory way to handle the problem of correlated predictors. The first few components should be sufficient for a stepwise regression. Also, in a stepwise regression, additional synthetic predictor variables do *not* change the coefficients of variables already in the equation, because the predictors are uncorrelated.

Caution: the first component is *not* necessarily the best single predictor of yield, it only explains the most variance *among the predictor variables*. Other components may be better single yield predictors. Also, it may well be the case that some of the original variables are better single predictors than any of the principal components.

Problem: sometimes the new axes are not *interpretable* in terms of the original predictors, even though they are mathematically the best combination. The best situation is when the predictor variables naturally fall into highly-correlated interpretable groups, e.g. all the temperature-related variables in one group.

2.5 Use of principal components in yield prediction

1. In the *calibration* step, the original dataset is used to transform from *original* variables (land characteristics) to *synthetic* variables. In the process, we

obtain the principal component *coefficients* (i.e., the linear transformation from original variables to each of the synthetic variables).

2. Also in the calibration step, we establish a satisfactory regression from the first few synthetic variables (principal components) to the yield.

3. Now in the *validation* or *use* step, we observe values of the *original* land characteristics at a new set of data points.

4. Using the principal component coefficients from step (1)., we *transform* each observation from the original LC's to values of the synthetic variables.

5. We substitute the values of the synthetic variables into the regression from step (2), and calculate the predicted yield.

If the second data set is a *validation* data set, i.e., if we measure yield as well as the predictor LCs, we can determine the predictive power of the regression established in (2).

3. Dynamic simulation modeling for land evaluation

References: Overview: (van Diepen *et al.*, 1991) p. 184-188

Introduction to dynamic simulation as such: (Ferrari, 1982)

Principles of modeling: (Penning de Vries & van Laar, 1982)

The WOFOST approach to crop modeling: (van Diepen *et al.*, 1989, van Keulen & Wolf, 1986)

Other crop models: (de Wit & van Keulen, 1987, Dumanski & Onofrei, 1989, Jones & Kiniry, 1986, Wilkerson *et al.*, 1983), a cautionary note in (Varcoe, 1990)

Modeling land qualities: (Hanks & Ritchie, 1991); (Hutson & Wagenet, 1992) for risk of pesticide leaching

3.1 Why use dynamic models in land evaluation?

Statistical modeling attempts to describe the *static* relation between land characteristics and either yields or land qualities. In many situations, this may not give satisfactory results because of the *dynamic* (time-dependent) nature of either the land quality (e.g., available days for planting) or the land characteristics on which the evaluation is based (e.g., weather). If the land evaluation problem is fundamentally dynamic, the techniques of *dynamic simulation modeling* are appropriate.

Dynamic simulation can be used to model individual *land qualities*, e.g. water stress. This is appropriate if the *timing* of the quality is important. Water stress is a good example: the yearly moisture deficit often isn't as important as the deficit in specific parts of the crop growth cycle.

As with statistical methods, one of the main uses of dynamic simulation in land evaluation is to model *crop yield*, since this integrates most of the agro-ecological land qualities and is an important part of economic suitability. Dynamic models are used in preference to static models when they give better results, presumably because they better capture transient events like moisture stress.

(Dynamic simulation of yield levels under different stresses can provide insight into ALES 'S1 Yields' and proportional yield decision procedures such as decision trees, and limiting and multiplicative yield factors.)

Note that dynamic models have many uses in tactical planning (e.g., irrigation or pest management) where the time factor or transient phenomena are important. Perhaps their best use is to gain *insight* into the presumed workings of the system which they are modeling; i.e., make the 'black box' of the system more transparent.

3.2 Definitions: Systems, models, and simulation

See (de Wit, 1982, Ferrari, 1982)

System: a limited part of reality, with the connections between its elements and with the outside world (non-system) well-specified.

Model: a simplified representation of the system, usually in mathematical or computable form (as opposed to iconic or analog models).

Simulation: the art of building mathematical or computer models of a system and using these models to study the properties of the system.

Man-made (engineered) systems are easier to model than natural systems because of a drastically-reduced complexity. Agricultural systems are intermediate in complexity: greatly simplified from nature (fewer species, subtle effects often overwhelmed by management) but still biological not mechanical, electrical etc.

3.3 Definitions: Types of models

Dynamic models include *time* as an explicit element of the model, otherwise the model is *static*. In dynamic models, the *state* of the system at one time, plus the *driving forces*, follow definite *transformation relations* to reach the next state, and so on till the end of the simulation. This is sometimes called the *state-variable* approach.

Explanatory models attempt to explain how a system works, from some first principles. For example, crop growth based on photosynthetic reactions as influenced by temperature, light, vapor pressure etc.

Descriptive models simply attempt to characterize a system for predictive purposes, without pretending to explain. Statistical models are a subclass of these.

It would seem that the best explanation would give the best description, in practice the purposes of the models are quite different. In land evaluation we generally are presented with descriptive models with pretensions to explanation.

Don't be deceived by claims that a model is derived from 'first principles'. There is always another level of physical reality underlying the supposed physical laws (at least until we get to quantum mechanics). The question is whether these underlying phenomena have been correctly *aggregated* at the level we are studying. For example, the theory of chemical equilibrium hides the dynamics of the actual reactions, but if the time scale is long enough (e.g., milliseconds to millennia, depending on the reaction being modeled) these don't matter, 'it all averages out'. This must be established for each 'law'.

All realistic models contain large doses of subjectivity, judgment, and empirical parameters. Although it would seem that a dynamic simulation model, being more mechanistic and explanatory than a statistical model, would be better able to extrapolate, this is not always so, and must be established by *validation* over the expected range of inputs, just like a statistical model. There is no assurance, except accumulated evidence, that the physical basis of the model is correct.

3.4 Definitions: Model parameters vs. data

In common language these both might be considered 'data', but they have very different roles in dynamic simulation modeling.

Model parameters are *constants* during the execution of the model, but may be variable between executions. Describe the *static* context in which the model is being run. Analogous to the parameters of a regression equation. They *parameterize* the equations of the model, i.e., supply specific values that control their numerical behavior (n.b. the *form* of the equations are fixed). Example: number of heat units that must be accumulated before a plant will flower; this will vary among species and varieties. Example: exponent in a decay function.

Data are the time series of input *variables*, which cause *state changes* in the model. They *drive* the behavior of the model in a particular execution. Example: rainfall over time.

4. Dynamic simulation of crop yield: the WOFOST approach

Many dynamic simulation models have been developed to predict crop yield. For didactic purposes we follow the approach of the Center for World Food Studies (CABO) in Wageningen, the Netherlands. They developed a flexible model based on basic plant physiology, to predict yields in several *production levels*, which more-or-less correspond to Land Utilization Types.

The WOFOST approach (van Diepen *et al.*, 1989, van Keulen & Wolf, 1986) considers three (here expanded to five) levels of increasingly more realistic limitations. This approach allows us to understand the production system in increasing detail. You will appreciate as you read this list how much harder it is to model increasingly-realistic production systems than to model simple, controlled systems.

First we will define the production levels are, then we will study production level 1 in some detail.

4.1 Production levels

4.1.1 Production level 1: Radiation and temperature limited

Growth occurs in conditions with ample plant nutrients, water, and oxygen (if necessary) all the time. The growth rate of vegetation is determined by weather conditions and the response of the plant to these. This can be approached in practice with very intensively managed irrigated crops. The model is basically one of photosynthesis, partition of carbohydrate, and physiological growth stages (e.g., flowering, senescence). The only inputs to the model are temperature and radiation (perhaps inferred from cloudiness).

4.1.2 Production level 2: Water limited

Growth is limited by water shortage at least part of the time, but when sufficient water is available the growth rate increases up to the maximum rate set by the weather. This can be approached in practice by intensively managed dryland crops. The model must determine water stress (so, needs to model soil water, the plant root system, and plant transpiration) and its effect on the photosynthetic and growth processes. Another input to the model is precipitation, and the soil profile must be modeled at least for the water balance.

4.1.3 Production level 3: Nitrogen limited

Growth is limited, at least part of the time, by shortage of nitrogen (N) and water or weather at other times. This is usually more limiting than other nutrients because N transformations in the soil are much more rapid than for other nutrients. This is common in dryland crops even if 'well-fertilized' and is especially relevant in systems which use animal or green manures. The model must determine soil N dynamics, plant uptake, N use in the plant, and effects of N stress on photosynthesis, partition and growth. Soil temperature can have a large effect on microbial populations, so this must usually be modeled. Soil organic matter cycling must be modeled.

4.1.4 Production level 4: Nutrient limited (other than N)

Growth is limited, at least part of the time, by shortage of mineral elements other than nitrogen, e.g. phosphorus (P) or potassium (K), or by other soil chemical conditions, e.g., soil reaction (pH), and by N, water or weather at other times. This is the usual situation in 'improved traditional' agricultural systems which use tillage and pesticides. Same model requirements as level 3 but for all the elements (although the soil dynamics are easier for elements other than N).

4.1.5 Production level 5: Weeds, pests & diseases

Growth is limited, at least part of the time, by competition from weeds and attack by pests or diseases. This is the usual situation in 'traditional' agriculture where fertilizer use is low or none and there is no use of agrochemicals. The model must incorporate the ecology of pests and diseases and multi-species competition.

4.2 Governing equations at production level 1

Let's go through the governing equations for the *simplest* case, to see the general form of the model, the kind of information that is needed to run the model, and where the *feedback* (dynamics) comes in.

(1) The underlying plant-physiological processes are *photosynthesis* and *respiration*. A single healthy leaf has a *maximum* CO_2 assimilation rate F_g that depends on the efficiency of the photosynthetic process as a photochemical reaction, and subsequent steps on the photosynthetic pathway. For C_3-type plants (e.g., temperate grasses, like wheat), this is on the order of 40 kg ha^{-1} (leaf) h^{-1}. For C_4-type plants (e.g., tropical grasses, like maize), this is on the order of 70 kg ha^{-1} (leaf) h^{-1}.

These rates have to be established by experiment, and depend on *temperature*. In the simplest case we would have a simple linear regression from temperature to maximum assimilation rate, with two parameters: the constant and the linear terms (established by experiment):

$$F_g = \beta_0 + \beta_1 \cdot T$$

Data need: temperature (dynamic)

Parameter need: parameters in regression of assimilation on temperature

(2) We must *integrate* (sum) the hourly rates over the hours of daylight (depends on latitude and date) to obtain F_{cl}, the gross assimilation rate (kg ha^{-1} d^{-1}).

Data need: latitude (static), day of year

(1b, 2b) The same analysis holds for overcast days, using assimilation rates for diffuse radiation, which are typically about 38% of the assimilation rates for clear days.

Parameter need: gross diffuse-lit assimilation rates as a function of temperature, or as a function of gross sunlit assimilation rate (can be a constant function).

(3) The *gross canopy assimilation rate* of CO_2, F_{gc}, (kg ha^{-1} d^{-1}) can then be calculated as a weighted sum of the direct and diffuse assimilation rates:

$$F_{gc} = f_o \cdot F_{ov} + (1 - f_o) \cdot F_{cl}$$

where f_o is the fraction of the day when the sky is overcast, and F_{ov} and F_{cl} are the gross assimilation rates (kg ha^{-1} d^{-1}) on completely overcast (diffuse radiation) and clear days (direct radiation).]

Data need: fraction of the day overcast and clear. Can sometimes be estimated from rainfall occurrence and amount.

(4) If the canopy is not completely closed (e.g. at the beginning of the crop cycle), some of the light is wasted, i.e. is not intercepted by a leaf. We must reduce F_{gc} accordingly by a factor f_h:

$$f_h = (1 - e^{-k_e \cdot LAI})$$

where *LAI* is the *Leaf Area Index* (m^2 leaf m^{-2} ground) and k_e is an *extinction coefficient* which measures how deeply light penetrates in the canopy. This can vary from 0.5 to 0.8 depending on crop geometry.

Parameter need: extinction coefficient

(*Interesting point*: this empirical relation is derived and calibrated by regression analysis, i.e. the supposed 'physically-based' model contains this (and many other) empirical equations. It is too complicated to derive from the geometry of the plant canopy. Another possibility would be to *simulate* the light and its interception by individual leaves in the canopy, as well as the geometry of each leaf over time. Major supercomputer power required! and unclear whether we know enough to do this.)

Simulated state variable: LAI

Note how the high uncertainty in k_e can greatly affect the results. For a LAI of 2 (fairly early in crop development), f_h varies from 0.63 to 0.80; for a LAI of 5 (full leaf), f_h varies from 0.76 to 0.98. These then multiply the predicted gross assimilation rate:

$$F_g = F_{gc} \cdot f_h$$

(5) The gross assimilation must be reduced by the *maintenance respiration*, i.e., the amount of carbohydrate that must be burned to maintain the plant at the same state, without any growth (this because enzymes and proteins break down with time, and cells must be repaired). This depends on temperature, having a Q_{10} of about 2 (i.e., doubles with a 10°C increase in temperature). This is the *relative maintenance respiration* rate R_m (kg CH_2O kg^{-1} dry weight d^{-1}).

Parameter need: relative maintenance respiration rate as a function of temperature

Data need: air temperature

(6) Not all the carbohydrate that is not used for maintenance respiration actually becomes structural carbohydrate (i.e., increase in dry weight). This *efficiency factor* E_g must be determined experimentally. So at last we have an equation for ΔW, the rate of increase in structural dry weight in kg ha^{-1} d^{-1}:

$$\Delta W = E_g \cdot (F_g - R_m \cdot W)$$

where W is the current dry weight of the live parts of the crop (kg ha^{-1}), i.e., the part that needs to be maintained with the maintenance respiration. This equation is valid at every (daily) *time step*.

Parameter need: efficiency factor of conversion of carbohydrate to structural dry weight.

Simulated state variable: Structural dry weight. Complication: above- vs. below-ground, very hard to measure the latter.

(7) Then we can derive the *difference equation*: the *rate* of growth depends on the *actual size* of the crop, which in turn is increased by the new growth:

$$W_{t+1} = W_t + \Delta W_t$$

Boundary (initial) condition need: W_0, original dry weight of seeds or transplants.

(8) There is also a *feedback*: notice that we used the *LAI* to adjust the gross assimilation rate for incomplete light interception. So, we must express LAI as a function of dry weight:

$$LAI_{t+1} = W_{t+1} \cdot sla \cdot (10^{-4} ha\ m^{-2})$$

where *sla* is the *specific leaf area* (m^2 kg^{-1}), which converts from leaf *weight* to leaf *area*.

Parameter need: specific leaf area.

So now we can run a simulation with two state variables: *W* and *LAI*.

This model contains an implicit limit to plant growth: as plant weight *W* grows, so does maintenance respiration. After a certain point, which depends on the extinction coefficient, extra *LAI* does not improve radiation interception, so that photosynthesis is at a maximum. When photosynthesis equals maintenance respiration, growth stops.

Note that we haven't attempted to deal with growth stages, partitioning, grain or other economic product formation, etc.

Simple program in a procedural language (Pascal-like pseudocode):

```
{ data arrays: temp[first_day..last_day],
         hrs_overcast[first_day..last_day] }
{ parameters r_m_fact, ov_fact, f_cl_fact, k_e }
{ not shown: real function daylight_hrs(day: integer) }
{ initial conditions w_0 and lai_0 }
w := w_0; lai = lai_0
for day = first_day to last_day by 1 do begin
         f_cl := (f_cl_fact_1 +             { sunlit assimilation rate }
                f_cl_fact_2 * temp[day]) * daylight_hrs(day}
         f_ov := f_cl * ov_fact             { shade assimilation rate }
         f_gc := hrs_overcast[day] * f_ov + (daylight_hrs(day) -
                hrs_overcast[day])* f_cl    { gross canopy assimilation }
         f_h := (1-exp(- k_e * lai)         { light reduction factor }
         f_g := f_gc * f_h                  { actual gross assimilation }
         r_m := r_m_fact * temp[day] * w    { maintenance respiration }
         w_new := e_g * (f_g – r_m)         { compute new growth }
         w := w + w_new                     { update dry weight }
         lai := w * sla * 10**-4            { update LAI }
         end { for day }
print "The final weight is ",w, "The final LAI is ",lai
```

4.3 Model assumptions

The production levels each incorporate a set of *assumptions*. For example, in all levels except 5, weeds, pests and diseases are not important to yield. Any such assumption must be explicit, and the model *user* (e.g., the land evaluator) must determine whether the assumptions are likely to be true before using the model.

Example: The GAPS soil-water model: does not account for snowfall (precipitation that can be stored at the surface); the GAPS wheat growth and yield model does not account for vernalization. For both these reasons, this model should not be used to model winter wheat production in Kansas.

Another problem is the *range of calibration* of the model. It is usually unclear how far the model can extrapolate. Example: application of today's crop growth models for global climate change studies. Many models assume constant CO_2 concentration in the atmosphere, as this changes too slowly (40 years to see appreciable changes) to affect the success of the model for today's yield predictions, yet clearly the model should include CO_2 effects if we're modeling the year 2100. Major problem: very hard to get realistic parameters in today's atmosphere.

5. Key issues in dynamic simulation modeling

After the previous section, you may think that models are impossibly complex. In many cases, however, they give good results. There is no *a priori* way to know this; each model in each location must be *calibrated* and *validated* to see if the model assumptions are satisfied. In this lecture we examine some of the key issues which result in more or less successful models.

5.1 The time step

In the state-variable approach, the differential equations of growth are solved by *numerical integration*, in effect, computing the growth in one step from the state in the previous state and the amounts of some inputs (e.g., temperature, precipitation) in the current step.

The *time step* controls the *temporal resolution* of the integration and its *accuracy*. Exactly the same problem as numerical integration by quadrature: the finer the quadrature, the more accurate the result. Depending on the rate of change of the growth function, a longer or shorter time step is needed to reach a given accuracy.

For example, the maximum carbohydrate assimilation rate can be calculated every hour, ten minutes, minute, 30 seconds, second... using the average temperature *for that time step* (in practice, usually the temperature at the mid-point of the time step), then summed over the day.

After a certain number of divisions, further refinement in the time step is insignificant (e.g., when the temperature doesn't change significantly between time steps).

In practice, the time step is also controlled by the temporal resolution of the *data*. For example, if rainfall is only available on a daily or hourly basis, this would seem to limit the resolution (the time step couldn't be any finer than the data). There exist some methods for decomposing a single amount into shorter time steps:

1. Instantaneous air temperature can be inferred from daily Min/Max by fitting a sine wave with period of 24 hours and assuming that the Max occurs at 1500 and the Min at 0300; then the temperature in any discrete time step can be obtained by integrating that portion of the sine wave and dividing by its duration to obtain an average. This works fairly well except when frontal air masses pass through, or if there are local diurnal air movement (e.g. sea or mountain breezes) which are not modeled by the sine wave.

2 Precipitation in a sub-daily time step can be inferred from daily precipitation by dividing the 24-hour precipitation by the number of hours in the time step. If the precipitation has a known temporal pattern (e.g. almost always afternoon thundershowers in the southeastern USA in summer) an approximate rule based on this prior knowledge can be used instead.

3. Daily precipitation can be inferred from monthly precipitation by fitting historical data to a statistical distribution and then sampling from it. E.g., Poisson or Weibull distributions. This is a good approach for *simulated future climates* e.g., for climate-change studies.

Any such *data generators* should be viewed with caution and verified both theoretically and practically before use.

5.2 Data sources

As you can appreciate from the discussion of model levels and time step, *detailed* and *frequent* data is required for dynamic simulation.

An example is the 'minimum data set' for the CERES and GRO series of models.

(1) *Soil properties as a function of depth*: horizon thickness, upper and lower limits of volumetric water, volumetric water at saturation, bulk density, pH, organic carbon, total nitrogen

(2) *Daily weather data*: radiation, precipitation, max/min temperatures

(3) *Crop parameters*: maturity type, photoperiod response, yield components

(4) *Initial conditions*: water content by depth, nitrates and ammonium by depth

(5) *Management conditions*: sowing date, plant population, irrigation amounts and dates, fertilizer amounts and dates, residue management, plowing depth

(What do we do when we don't have the required data? See the section on *parameter estimation*, below.)

GAPS is somewhat adaptable: there are various ways to model the same phenomenon, differing in their data requirements (and assumptions).

Example: potential evapotranspiration (ET) by: Priestly-Taylor, Penman, pan, and Linacre. The more complicated model doesn't always give the best results!

Penman-Monteith (model vapor diffusion and energy budget): *parameters*:
height above canopy of wind speed measurement; shortwave absorptivity of

leaves, resistance of canopy when stomates open, aerodynamic resistance of the canopy, latitude; *data*: wind speed, min/max air temperature, solar radiation, relative humidity, saturated vapor density (can be estimated from air temperature if not available); *simulated variables*: height of canopy

Priestly-Taylor (simpler version of Penman-Monteith, does not simulate aerodynamic resistance to vapor transport): *parameters*: alpha, shortwave absorptivity of the soil surface, latitude ; *data*: min/max air temperature, solar radiation; *data or simulated variables*: cloudiness

Pan (based on direct measurements of evaporation): *parameters*: pan and crop coefficients; *data*: daily measured pan evaporation

Linacre (simple empirical formula based on easily-obtainable data): *parameters*: elevation above sea level, latitude; *data*: min/max air temperature, dew-point temperature (optional, min. air temperature can approximate this)

5.3 Calibration of model parameters

As you can appreciate from the discussion of model levels, dynamic simulation models have from 10s to 1,000s of parameters ('magic numbers'), similar to a regression equation's parameters. Each of these must be established by calibration or from 'first' principles. The big problem is that very rarely does an experiment give information on only one parameter, and most of the possible interactions are unknown. Also, many adjustments of parameters can lead to the same final results. What is needed is insight into the physical system, and again many of the interactions are poorly-understood.

Example: calibration of the CERES and GRO series of models for the IBSNAT project: very large, detailed experiments for years, even then it is unclear how to adjust the model.

Again, GAPS is somewhat adaptable: there are various ways to model the same phenomenon, differing in their parameters to be calibrated. Example: soil water balance by Richards' Equation (requires saturated hydraulic conductivity and moisture release curves for each layer) vs. the 'tipping bucket ' (only requires saturated water content, field capacity water content, and wilting point water content; except for the last these are very easy to measure).

5.4 Sensitivity analysis

A major issue with simulation modeling is the large number of *model parameters* (calibration values) and *input data* that are required. The question naturally arises: what happens if we get some of these wrong? The correct question is: how *sensitive* is the model to variations in parameters or data? Especially since parameter calibration is largely a black art, sensitivity analysis

allows us to see where we should concentrate our calibration and modeling efforts, i.e., where the model is most sensitive.

Definitions

Sensitivity: rate of change in *output variable* per unit change in *input variable* or parameter.

Absolute sensitivity: in terms of units, e.g. 'kilograms of yield' per 'mm of precipitation'.

Relative sensitivity: both factors *standardized* to zero mean and unit standard deviation.

Example of sensitivity to data: In a global climate change study in midwestern USA, Rossiter & Riha determined that large uncertainty in rainfall events did not substantially affect simulated yields (using the GAPS model) in soils with high water-holding capacity (i.e. the soil buffered the rainfall inputs); in sandy shallow soils there was a correspondingly large effect. So the fact that the actual climate change is uncertain didn't have much effect on our results in the first kind of soils.

Basic method for sensitivity analysis

Vary the parameter in some predictable way, run the model, and record the output. Plot the output vs. the parameter value, and perform a *regression analysis* to quantify the effect of the parameter on the results. The *absolute sensitivity* is the slope of the regression line. The *relative sensitivity* is the slope of the regression line if both the independent and dependent variables are standardized; this is the *correlation coefficient*.

How to vary the parameter?

Method 1: random sample from a known or assumed probability distribution. This gives an unbiased estimate of the sensitivity to the parameter, *if* the underlying distribution is known.

Method 2: random sample from a known or assumed empirical distribution (e.g. historical time series)

Method 3: non-random sample from a known or assumed probability distribution, selecting e.g. ±1, ±2 standard deviations. This is an efficient way to get at the sensitivity to *extreme* values of the parameter.

Method 4: stratified random sample from either a probability or empirical distribution; the sampling scheme has certain desirable theoretical properties that allow an estimate of sensitivity with a smaller sample size, because the sample is more evenly spread out over all possible values. Example: Latin hyper-cube sampling (LHS) (Morgan & Henrion, 1990 pp. 204-205)

6. Transfer functions and parameter estimation

(Bouma, 1989, 1986, Hutson & Wagenet, 1992, Ritchie & Crum, 1989)

A major problem with applying simulation models to land *areas* (as opposed to single sites) is that the required model parameters are not usually available over wide geographical areas. There is a mismatch between *model* and *resource inventory*.

Supposing the required parameters are not available from routine survey. The modeller has two choices:

1. Make 'reasonable' guesses, then tune these parameters by some calibration trials. The parameter is never measured.

   This matching is strictly trial-and-error and has no physical basis. It may be justified if the model is not sensitive to the parameter and we have sufficient confidence that the estimated parameter is within a defined range.

2. Derive the required parameters from other, *easier-* (or cheaper-) *to-measure* parameters. The derivation is known as a *transfer function*. In the case of soil-survey data transformed to model parameters, it is known as a *pedotransfer function* (this term introduced by (Bouma & van Lanen, 1986)).

   The derivation should have a physical basis, or at least, even if just a regression, the calibration of the transfer function should be independent of the calibration of the model as a whole.

6.1 Key questions for evaluating transfer functions

(1) *How good is the relation?* This is often measured by the coefficient of determination (r^2) of the original calibration regression, however it should really be measured by the coefficient of determination (r^2) of the regression of predicted vs. measured for a set of *calibration* data (very rarely done this way). This is the same issue as the postdiction vs. prediction discussed under statistical yield modeling.

(2) *Is the relation good enough?* Even if the relation is not particularly good, it may be good enough for modeling, depending on the *sensitivity* of the model to this parameter, which can be determined by sensitivity analysis as explained in the previous section.

(3) *Over what population is the relation valid?* For example, if no organic/volcanic/vertic etc. soils are included the relation should *not* be assumed to hold! P.M. Driessen, pp. 217-221 in (van Keulen & Wolf, 1986)

is a good example of the many cautions to be observed when estimating water relations in 'unusual' soils.

(4) *Does the relation seem to have a physical basis?* We should have more confidence if the relation seems at least plausible.

6.2 Example: modeling the soil-water regime

The classic example of pedotransfer functions (e.g. (Bouma, 1989)) is modeling the soil water regime, because this is the critical dynamic land characteristic needed to evaluate land qualities such as 'moisture availability', 'trafficability', and 'oxygen availability to roots'.

Soil water is often modeled on the basis of the *energy* of water and the *forces* acting on it in the soil profile. The following sorts of parameters are required:

- √ saturated hydraulic conductivity { K_{sat} }
- √ air-entry potential (water potential at which air enters the pores) { $K(\theta)$ }
- √ hydraulic conductivity as a function of water content { $K(\theta)$ } or slope of the moisture release curve on a log-log scale { b }

This set of parameters (to model by water potentials) is only measured on a few profiles in routine survey, and often is not measured at all. Its experimental determination is difficult, expensive, and error-prone.

A simpler way of modeling soil water is on a *capacity* basis (volumetric water content, e.g., the so-called 'tipping bucket' approach) as opposed to a water-potential basis. In this case, we need to know things like:

- √ field-capacity water content or 'drained upper limit', % of volume
- √ wilting-point water content or 'lower limit', % of volume
- √ saturated water content, % of volume

This set of parameters (to model water content) is a much easier set of measurements than those for water potentials, but still is only measured on a set of profiles in routine survey.

But, there is a lot of simple geographically-based soils data collected by soil survey organizations. Nowadays much is available in databases and/or GIS format. So if we can somehow determine the parameters we need from the data we have, we can apply the model across the entire geographic range of interest.

In both cases, to apply the model across the *entire range* of soils found in a survey area, we must *infer* the required parameters by *pedotransfer functions*.

6.3 Estimating parameters for a capacity model

The simpler case is the capacity model. (Ritchie & Crum, 1989) is typical of many efforts that have been made to derive critical volumetric water contents from soil survey data. The pedotransfer functions are usually simple equations from available land characteristics to the required parameters. These equations are obtained by regression analysis on a large number of samples where both field and laboratory values are measured.

6.3.1 A simple relation: saturated water content

We assume that the saturated water content θ_{sat} is 0.85% of the total porosity, which is completely determined by the bulk density D_f, assuming the bulk density of the mineral particles to be 2.65 (valid for quartz and aluminosilicate-dominated soils):

$$\theta_{sat} = (1 - D_f / 2.65) * 0.85$$

This is an example of a *physically-based* related which only requires some particular values. Here, the physical basis is that saturated water almost fills up the pores.

6.3.2 A more complicated case: extractable water

(Ritchie & Crum, 1989) base much of their presentation on the study by (Cassel, Ratliff & Ritchie, 1983) on estimating potential plant-extractable water for a wide variety of soils in the USA. (Cassel, Ratliff & Ritchie, 1983) considered a large number of samples with the following *dependent* variables:

(1) DUL (drained upper limit, i.e., 'field capacity'), volume %;
(2) LOL (lower limit, i.e., 'wilting point'), volume %;
(3) PLEXW (extractable (by plants) soil water) ≡ DUL − LOL, volume %

They considered the following 22 *independent* (predictor) variables:

<u>Group 1: routine physical measurements (particle-size distribution)</u>
(1) % Sand (50-2000 µm), % Silt (2-50 µm), % Clay (<2µm)
(2) % very fine, fine, medium, coarse, and very coarse sand
(3) % through #200 sieve (= % silt + % clay + 1/2 % very fine sand)
(4) % coarse fragments (>2000 µm) by weight

<u>Group 2: routine chemical measurements</u>
(4) % organic carbon
(5) $CaCO_3$ (weight)
(6) Cation Exchange Capacity (CEC) by NH_4OAc
(7) pH in 1:1: soil:water (weight)

<u>Group 3: non-routine physical measurements</u>
(8) Weight % water at -0.06, -0.10, -0.33, -2, and -15 bars (PW-15)
(9) bulk density at -0.33 bar (considered to be at DUL water content)

All measurements were made at the same laboratory (USDA/SCS Lincoln)

They developed single regression equations to fit all the soils in the sample, but found that the *universe* of all soils was best *divided* according to major textural class in order to obtain reasonably-significant regressions. The three classes were (1) s, ls; (2) sl, l, scl, sil, si; (3) sicl, sic, cl c.

They provided four levels of multiple regressions: using two, four, ten and nine measured soil properties, respectively. This allowed for situations with only a few variables as well as better predictions where more variables were measured. The final choice of variables was:

	Database level			
Variable	1	2	3	4
% clay	√	√	√	√
% sand		√	√	√
% silt		√	√	√
% fine silt			√	√
% very fine sand			√	√
% fine sand			√	√
% medium sand			√	√
PW-15		√	√	√
CEC			√	
% through #200 sieve	√		√	√

Their best results for level one, for all soils, were:

$$DUL = 8.682\,922 + 1.447\,671 * (\%\ clay) - 0.018\,783 * (\%\ clay)^2 - 0.003\,282 * (\#200\ sieve)^2 + 0.000\,033 * (\#200\ sieve)^3,\ r^2 = 0.76$$

$$LOL = 1.659\,522 + 0.930\,216 * (\%\ clay) - 0.000\,197 * (\%\ clay)^3 - 0.003\,849 * (\#200\ sieve)^2 + 0.000\,036 * (\#200\ sieve)^3,\ r^2 = 0.70$$

Note the arbitrariness of these equations. Why use a squared-clay term for DUL but a cubed-clay term for LOL?

The most sophisticated method, i.e., dividing the soils into three groups and using up to ten predictor variables, resulted in better results for sandy soils (r^2 of 0.91 and 0.90), results comparable to the all-soils approach for medium-textured soils (r^2 of 0.77 and 0.71), and poorer results for clayey soils (r^2 of 0.62 and 0.69).

To *validate* their results, they regressed the *calculated* extractable water CALPLEXW = DUL–LOL on the measured extractable water PLEXW. Obviously, this line should be 1:1. In fact, the following results (values of r^2) were obtained:

	Number of groups	
Number of predictors	1	3
2	0.34	0.39
4	0.36	0.46
10	0.56	0.74
9	0.43	0.64

Evidently, the only satisfactory prediction was from the 10-variable predictors on the three textural groups taken separately. The simple predictor from % clay and % passing the #200 sieve, for all soils, only explains 34% of the observed variance in extractable water.

The *standard error* of the estimate depended on the estimate itself, i.e., the error increased with water content. This is not so bad, since an error in calculating available water is more critical in soils with low water-holding capacity. Unfortunately the authors only show standard errors for the relations with 10-variable predictors. In the best case (10 variables, 3 groups of soils), the errors are on the order of ±3% available water, i.e. in 1m deep soil profile, ±30mm of available water, equivalent to a 5-day supply at a transpiration rate of 6mm day^{-1}. So the practical uncertainty is about ±5 days without rain! This seems like a lot in sub-humid conditions.

Cautionary notes: These relations are probably 'as good as it can get'. They were calibrated with field and laboratory data with the best possible methods. They emphatically do *not* apply to soils with 'unusual' laboratory of field behavior, for example, those with coarse fragment >10% by weight, organic carbon > 4%, soils with significant partially-decomposed organic matter, soil material that is difficult to disperse in the laboratory (silica-enriched soils), soils where clay increases with energy level in the determination (highly-micro-aggregated Oxisols), and soils dominated by amorphous clay minerals (Andisols), . They must be adjusted for root behavior for poorly-drained soils, soils with root-limiting layers, soils with toxicities or high salt contents (> 2 dS m^{-1}). They do not apply in soils with groundwater influence.

Seems like we've excluded many soils...

(One possible conclusion: we are measuring the wrong things in the laboratory.)

6.4 Estimating parameters for a water-potential model

(This follows (van Genuchten *et al.*, 1989), see also (Vereecken *et al.*, 1989).)

The water potential model of soil water is more complicated both theoretically and in the parameters that it requires. The basic 'continuity' equation is:

$$\frac{d\theta(h)}{dh}\cdot\frac{\partial h}{\partial t}=\frac{\partial}{\partial z}K(h)[\frac{\partial h}{\partial z}-1]$$

where h is the water pressure head (soil water potential), θ(h) is the volumetric water content as a function of pressure head, $K(h)$ is the (unsaturated) hydraulic conductivity of the soil as a function of pressure head, z is depth below datum (usually the soil surface), t is time. In words, this equation says that the rate of change of water potential with time times the slope of the water release curve equals the rate of change with depth of the unsaturated hydraulic conductivity, as water flows down under the influence of gravity.

The key factors in this equation are the functions θ(h) and $K(h)$ and the slope of the water release curve $d\theta(h)/dh$. Since entire functions must be determined, rather than just single parameters, the estimation problem is much greater. Ideally these are derived by field or soil-column experiments.

However they must often be estimated from simpler soil properties. For example, (Vereecken *et al.*, 1989) estimated the moisture release curve θ(h) with transfer functions for eight parameters (see their p. 123 for graph and equations), with the adjusted r^2 for the parameters varying from 56% to 93%.

7. References

1. Beek, K.J., Burrough, P.A. & McCormack, D.E. (ed). 1987. *Quantified land evaluation procedures: Proceedings of the international workshop on quantified land evaluation procedures held in Washington, DC 27 April - 2 May 1986*. International Institute for Aerospace Survey and Earth Sciences (ITC) Publication No. 6, Enschede, the Netherlands: ITC.

2. Bouma, J. 1989. *Using soil survey data for quantitative land evaluation*, in *Advances in Soil Science*, Stewart, B.A., Editor. New York: Springer. p. 177-213.

3. Bouma, J. & Bregt, A.K. (ed). 1989. *Land qualities in space and time. Proceedings of a symposium organized by the International Society of Soil Science (ISSS), Wageningen, the Netherlands 22-26 August 1988*. Wageningen: Pudoc. 352 pp. S593 .L25 1989

4. Bouma, J. & van Lanen, H.A.J. 1986. *Transfer functions and threshold values: from soil characteristics to land qualities*. in *Quantified land evaluation procedures: Proceedings of the international workshop on quantified land evaluation procedures held in Washington, DC 27 April - 2 May 1986*. Washington, DC: ITC.

5. Cassel, D.K., Ratliff, L.F. & Ritchie, J.T. 1983. *Models for estimating in-situ potential extractable water using soil physical and chemical properties*. Soil Sci. Soc. Am. Proc. 47(4): 764-769.

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

7. De la Rosa, D., Cardona, F. & Paneque, G. 1981. *Crop yield predictions based on properties of soils in Sevilla, Spain*. Geoderma 25: 267-274.

8. De la Rosa, D., Moreno, J.A., Garcia, L.V., and Almorza, J. 1992. *MicroLEIS: A microcomputer-based Mediterranean land evaluation information system*. Soil Use Manag. 8(2): 89-96.

9. de Wit, C.T. 1982. *Simulation of living systems*, in *Simulation of plant growth and crop production*, Penning de Vries, F.W.T. & van Laar, H.H., Editor. Wageningen: PUDOC. p. 3-8.

10. de Wit, C.T. & van Keulen, H. 1987. *Modelling production of field crops and its requirements*. Geoderma 40: 253-265.

11. Draper, N.R. & Smith, H. 1981. *Applied regression analysis*. 2nd ed. New York: John Wiley. xiv, 709 pp. QA276 .D76 1981 Mann, Engineering reserve

12. Dumanski, J. & Onofrei, C. 1989. *Techniques of crop yield assessment for agricultural land evaluation.* Soil Use Manag. 5(1): 9-16.

13. Ferrari, T.J. 1982. *Introduction to dynamic simulation,* in *Simulation of plant growth and crop production,* Penning de Vries, F.W.T. & van Laar, H.H., Editor. Wageningen: PUDOC. p. 35-49.

14. Food and Agriculture Organization of the United Nations. 1985. *Guidelines: land evaluation for irrigated agriculture*. Soils Bulletin 55, Rome, Italy: FAO. 231 pp. S590 .F68 no. 55 Mann

15. Hanks, J. & Ritchie, J.T. (ed). 1991. *Modeling plant and soil systems.* Number 31 in the series Agronomy, Madison, WI: American Society of Agronomy. xix, 545 pp.

16. Hutson, J.L. & Wagenet, R.J. 1992. *LEACHM: Leaching Estimation and Chemistry and Model. A process-based model of water and solute movement, transformation, plant uptake and chemical reactions in the unsaturated zone. Version 3*. SCAS Research Series No. 92-3, Ithaca: Cornell University, Department of Soil, Crop and Atmospheric Sciences. 127 pp. pp.

17. Jones, C.A. & Kiniry, J.R. (ed). 1986. *CERES-Maize: a simulation model of maize growth and development.* College Station, TX: Texas A & M University Press. 194 pp.

18. Mitscherlich, E.A. 1909. *Des Gesetz des Minimums und das Gesetz des abnehmended Bodenertrages.* Landwirsch. Jahrb. 3: 537-552.

19. 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

20. Olson, K.R. & Olson, G.W. 1985. *A soil-climate index to predict corn yield.* Agric. Syst. 18: 227-237. S540.A2 A27 Mann

21. Olson, K.R. & Olson, G.W. 1986. *Use of multiple regression analysis to estimate average corn yields using selected soils and climatic data.* Agric. Syst. 20: 105-120. S540.A2 A27 Mann

22. Penning de Vries, F.W.T. & van Laar, H.H. (ed). 1982. *Simulation of plant growth and crop production*. Simulation Monographs, Wageningen: PUDOC. 308 pp. QK731. S61 Mann Reserve

23. Ritchie, J.T. & Crum, J. 1989. *Converting soil survey characterization data into IBSNAT crop model input,* in *Land qualities in space and time. Proceedings of a symposium organized by the International Society of Soil Science (ISSS), Wageningen, the Netherlands 22-26 August 1988,* Bouma, J. & Bregt, A.K., Editor. Wageningen: Pudoc. p. 155-167. S593 .L25 1989

24. Simonson, R. 1938. *Methods of estimating the productive capacity of soils.* Soil Sci. Soc. Am. Proc. 3: 247-251.

25. Steel, R.G.D. & Torrie, J.H. 1980. *Principles and procedures of statistics: a biometrical approach*. 2nd ed. New York: McGraw-Hill. 633 pp.

26. van Diepen, C.A., Van Keulen, H., Wolf, J., and Berkhout, J.A.A. 1991. *Land evaluation: from intuition to quantification,* in *Advances In Soil Science,* Stewart, B.A., Editor. New York: Springer. p. 139-204.

27. van Diepen, C.A., Wolf, J., van Keulen, H., and Rappoldt, C. 1989. *WOFOST: a simulation model of crop production.* Soil Use Manag. 5(1): 16-24.

28. van Genuchten, M.T., Kaveh, F., Russell, W.B., and Yates, S.R. 1989. *Direct and indirect methods for estimating the hydraulic properties of unsaturated soils,* in *Land qualities in space and time. Proceedings of a symposium organized by the International Society of Soil Science (ISSS), Wageningen, the Netherlands 22-26 August 1988,* Bouma, J. & Bregt, A.K., Editor. Wageningen: Pudoc. p. 61-72. S593 .L25 1989

29. van Keulen, H. & Wolf, J. (ed). 1986. *Modelling of agricultural production: weather, soils and crops.* Simulation Monographs, Wageningen: PUDOC. x, 479 pp. S494.5.M3 M68 1986 Mann

30. Varcoe, V.J. 1990. *A note on the computer simulation of crop growth in agricultural land evaluation.* Soil Use Manag. 6(3): 157-160.

31. Vereecken, H., Maes, J., Van Orshoven, J., and Feyen, J. 1989. *Deriving pedotransfer functions of soil hydraulic properties,* in *Land qualities in space and time. Proceedings of a symposium organized by the International Society of Soil Science (ISSS), Wageningen, the Netherlands 22-26 August 1988,* Bouma, J. & Bregt, A.K., Editor. Wageningen: Pudoc. p. 121-124. S593 .L25 1989

32. Webster, R. 1989. *Is regression what you really want?* Soil Use Manag. 5(2): 47-53.

33. Webster, R. & Oliver, M.A. 1990. *Statistical methods in soil and land resource survey.* Oxford: Oxford University Press. S591 .W38 1990 Mann

34. Wild, A. (ed). 1988. *Russell's Soil conditions and plant growth.* 11th ed. London: Longman Scientific & Technical. xi, 991 pp.

35. Wilkerson, G.G., Jones, J.W., Boote, K.J., Ingram, K.T., and Mishoe, J.W. 1983. *Modeling soybean growth for crop management.* Trans. ASAE 26: 63-73.