Application and Integration of Multiple Linear Regression and Linear Programming in Renewable Resource Analyses[1,2]

GEORGE M. VAN DYNE[3]

Ecologist, Radiation Ecology Section, Health Physics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, and Associate Professor of Biology, University of Tennessee, Knoxville.

Highlight

This paper presents preliminary results of formulating quantitatively the influence of site factors on various nutrient production measures and using these relationships in linear programming models to determine the optimum protein production on a foothill range. Site characteristics for optimum protein production were constrained to fall within the range of variables measured, and were constrained to satisfy certain inherent relationships known about these variables. This example shows a useful application of an operations research technique to resource evaluation problems.

[1] *This research was supported by the Atomic Energy Commission under contract with the Union Carbide Corporation. J. S. Olson and B. C. Patten are acknowledged for their suggestions in this research and for review of the manuscript.*

[2] *Paper presented at 19th annual meeting, American Society of Range Management, New Orleans, La., February 1-4, 1966.*

[3] *Present address: College of Forestry and Natural Resources, Colorado State University, Fort Collins.*

Large, fast digital computers have become available in the last 15 years and have allowed the development of special methods of analyzing and studying complex systems in industry and government. Range ecosystems are good examples of complex systems, and it is inevitable that mathematical analysis will become increasingly important in the future in range research and range management, as well as in many phases of renewable resource management. To take advantage of the methodological and conceptual advances from operations research and systems analysis means we will have to give increased attention to formulating and studying range problems in mathematical terms.

This paper reports only an *introductory* approach in applying and integrating multiple linear regression and linear programming methods in studying what I call the "optimum site problem." The work at present is neither exhaustive nor complete but will serve to show, with realistic examples, the potential of these techniques for learning more about range ecosystems.

The purpose of this paper is (i) to show the development of the quantitative formulation of site relationships to vegetation productivity, (ii) to use multiple linear regression equations as objective functions in, and to develop constraints for linear programming models, and (iii) to show by example and discussion where these approaches have application in analysis of renewable resource management problems.

The Range Site

Foothill ranges are good examples of complex and diverse environments. A schematic simplification is given in Fig. 1. Different geologic formations may outcrop at different levels providing various parent materials for residual soils, and parent materials for some soils may be transported onto the site. Variations in degree of slope and exposure also are characteristic of foothill rangelands. Important variable climatic influences in-

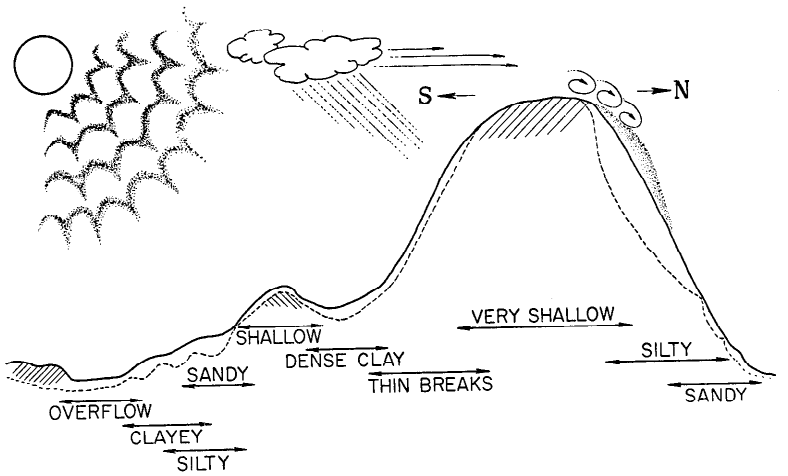

FIG. 1. A diagrammatic representation of the foothill range site complex showing variations in parent materials, soil depths, elevation, exposure, and climatic factors. Double-ended arrows show that boundaries of range sites are not discrete.

clude the angle at which sunlight strikes the soil surface and the exposure to the prevailing winds, which may be especially important in drifting snow onto leeward slopes. The ultimate result is the development, over a long period of time, of varying soils and topographic features which, when considered together with precipitation zones, we group arbitrarily into range sites, and to which we can ascribe a characteristic kind and amount of vegetation. The boundaries of range sites usually are not distinct, but they tend to intergrade and overlap in part. The range site name as such is of value, especially for large-scale surveys, but adds little to our quantitative knowledge about the relationships of the vegetation to the site factors.

Often it is desirable to be able to assess or to rank a given environmental complex such as a range site, a forest area, or some other unit according to some prescribed scheme of practical or theoretical importance. The assessment or ranking of a site implies that the various properties of this site have a functional relationship to criteria which are being ranked. Means of assessing the combined effects of site variables on some criteria have undergone long development originating with qualitative characterizations, and more recently turning to quantitative assessments.

Historical Development

An early qualitative statement, pertinent to the range site problem and attributed to Darwin, is that a particular plant community is selected from the available flora by the environment of a particular locale. This statement illustrates the early recognition that the various environmental factors, acting upon an original flora, lead to the development of a particular plant community. This notion was developed further by Dokuchaev (1898, referred to by Jenny, 1961), the Russian soil scientist, who formulated the following relationship.

$$S = f\,(Cl, O, P) \qquad (1)$$

where S refers to soils, Cl refers to macroclimate, O to organisms (presumably both plants and animals), and P to parent material. Later Jenny (1941) reformulated this relationship and added two new independent variables as follows:

$$S = f\,(Cl, O, R, P, t) \qquad (2)$$

where Cl, P, and S are defined as above, R refers to relief, and t to time. Jenny defines O as available flora and fauna so that it can be considered an independent variable rather than a dependent variable. This equation states that soil properties are dependent upon the influences of the climate acting over time on the original conditions of organisms, relief, and parent material. Similarly, Major (1951) has shown that vegetation is a function of the same state factors or independent variables. Later Jenny (1961) formulated a more general set of equations for an open system as follows:

$$l, s, v, \text{ or } a = f\,(L_o, P_x, t) \qquad (3)$$

where the dependent variables are any property of the total ecosystem (l), soils (s), vegetation(v), or animal community (a). The independent variables here are specified by the vector L_o which gives the initial state conditions, P_x which are the flux potentials, and t again referring to time. In the present sense, flux refers to the movement of matter and energy to and from contiguous ecosystems.

In all of the above formulations the time scale aproximates that of primary succession, evolutionary time, or geologic time. For a short time scale, such as much less than the time required for secondary succession, and for practical purposes, certain of the variables considered dependent variables in the above formulations may be considered to be independent variables. A change in terminology is introduced so that now independent and dependent are used in the conventional statistical sense rather than adhering strictly to Jenny's (1941) meanings. The statistical usage is denoted by asterisks. Thus, a new relationship may be formulated as follows:

$$V^* = f\,(Cl^*, R^*, S^*) \qquad (4)$$

or

$$Y = f\,(X_1, X_2, \ldots X_m | b_1, b_2, \ldots b_m)$$

where V* refers to some property of the vegetation which varies widely in a short period of time, for example, to the annual yield or composition of vegetation on a given site. The independent variables essentially are fixed in a short period of time and are Cl*, or macroclimate, R*, the relief features which would include such factors as elevation, slope, and exposure, and S*, the physical and chemical characteristics of the soil. A vegetation variable can be defined as a dependent variable, Y, and the site variables as independent variables X_i, in a multiple regression equation, and the b_i

are partial regression coefficients. The number of independent variables on any given range site is large, and their measurement becomes subject to practical considerations.

The Regression Model

The relationship of the vegetation variable, i.e., yield or composition, to any given independent variable may be nonlinear, and certain independent variables may have interacting influences. Development of a "mechanistic" model for predicting a vegetation variable, say productivity, no matter how interesting a modelling task, is unnecessary for the present purposes. As a first approximation and simplification for illustrative purposes, an empirical model for predicting a vegetation variable may be obtained by regression analyses. Multiple regression analysis techniques may be used to derive a first order model (linear terms only, without interaction) relating the independent site variables and the dependent vegetation variable, giving an equation as follows:

$$Y = b_o X_o + b_1 X_1 + b_2 X_2 + \ldots b_n X_n \quad (5)$$

$$\text{or } Y = \sum_{i=0}^{n} (b_i X_i)$$

where Y is the dependent variable, e.g., yield or composition of the vegetation, X_o is assigned the value one and the other X's are the independent variables, i.e., independent concerning time fixed to a narrow range. The value of such equations, of course, depends upon the sampling scheme in which the data were collected, the inherent variability of the population being sampled, and many other factors whose discussion is beyond the scope of this paper. Further information on the development and use of multiple regression models, both linear and nonlinear, may be found in statistical texts such as Ostle (1963), Hamilton (1964), and Keeping (1962).

The Linear Programming Model

The question may be asked, *"How can we select values for the site variables which will maximize the value of the vegetation property?"* The above multiple regression equation for predicting a vegetation variable can be used in a linear programming model as an objective function. Then we are interested in learning the values of the X's which would give us the maximum or minimum value, depending on which is desired, of the vegetation variable. If there were no constraints on selecting the values for the X's and we desired to maximize our vegetation variable, one could simply take an extremely large value for each site factor which has a positive partial regression coefficient and an extremely small value for those with negative coefficients. However, in real life this is not possible. Often there is a functional relationship between the variables which can be shown by a set of inequalities as follows:

$$\begin{aligned} &a_{11} X_1 + a_{12} X_2 + \ldots + a_{1m} X_m \leq c_1 \\ &\vdots \\ &a_{i1} X_1 + a_{i2} X_2 \ldots a_{ij} X_j \ldots a_{im} X_m \leq c_i \\ &\vdots \\ &a_{n1} X_1 + a_{n2} X_2 + \ldots + a_{nm} X_m \leq c_n \end{aligned} \quad (6)$$

where the a_{ij} and c_i are constants. In these inequalities the coefficients a_{ij} may be zero for many of the terms providing that at least one a_{ij} is greater than zero. An additional set of constraints in the linear programming model is that $X_i \geq 0$ for all i. Further background on linear programming models and applications may be found in such texts as Spivey (1963) for introductory treatment and Hadley (1962) for more advanced treatment. Further considerations about constraints pertinent to the optimum site problem follow.

Constraints on the Solution

There are three general types of constraints: (a) inherent relations, (b) those constraints to make the solution realistic, and (c) those imposed to evaluate economical or biological factors.

Constraints which are inherent in the nature of the independent variables include the following examples: (i) sand + silt + clay = 100, where mechanical composition data are expressed in percent; (ii) A horizon depth + B horizon depth = depth to C horizon; and (iii) depth to B horizon $\leq$ depth to C horizon. Here, for example, depths of the horizons have functional or predictable relationships following from their definitions.

Certain constraints are imposed upon the selection of values for the site factors in order to keep the solution realistic. Thus, for example, the following conditions represent a first approximation of some boundary conditions for the selection of each site variable in the solution vector:

$$\text{min. in field} \leq \text{site variable}_j \leq \text{max. in field} \quad (7)$$

Constraints may be imposed when certain economical or biological factors are to be considered and which have a functional relationship to the dependent variable which is being maximized or minimized. In general,

$$\mathbf{X^*}_{1xm}\,\mathbf{B^*}_{mx1} \geq \mathbf{Y^*}_{1x1} \quad (8)$$

where $\mathbf{X^*}$, $\mathbf{B^*}$, and $\mathbf{Y^*}$ are components of regression functions for other dependent variables, whose minimum or maximum values are being set according to some heuristic decision about the nature of the solution. An example of such an imposed constraint follows.

Assume heights and ages of two species of trees are measured in plots along with site variables. Multiple regression

equations are developed to predict height of each tree species from the set of site variables. Let the regression equation for species 1 be used as the objective function in the linear programming model. Assume we would like to find the site conditions to maximize height of species 1, yet we want these site conditions to provide at least better than average height for species 2. This can be accomplished by using the regression equation for species 2 as an inequality to be greater than or equal to the mean height of species 2. Four constraints of this type, developed from regression equations for dependent variables other than protein yield, were included in this problem and are discussed in more detail in the section on the optimum site.

Another realistic consideration concerning constraints is that all of the variables in the regression function, i.e., the objective function, are not equally important. Site factors having highly significant relationships with the vegetation parameter could be given additional consideration in the solution, i.e., the solution can be weighted for these variables. A preliminary suggestion on a method to accomplish this would be to use factor or principle component analyses to get an equation which would be a new linear combination of the more important independent variables. Such an equation could be used as a constraint to be satisfied in the linear programming solution.

From Data to Models

The above equations show how a property of the vegetation may be related quantitatively to measurable site factors, and they show how these relationships can be used to formulate an objective function and constraints in a linear programming model. The regression model is based on experimental data for the dependent vegetation parameters and the independent site factors collected under an appropriate experimental design or sampling plan. To provide a realistic example, site data and nutrient production data, collected from plots located by multistage randomization, are taken from range experiments of Van Dyne and Kittams (1960) and the following matrices are defined:

$$\mathbf{Y}_{nxj} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1j} \\ Y_{21} & Y_{22} & \cdots & Y_{2j} \\ \vdots & & & \\ Y_{n1} & Y_{n2} & \cdots & Y_{nj} \end{bmatrix}$$

$$\mathbf{X}_{nxm} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} \\ X_{21} & X_{22} & \cdots & X_{2j} \\ \vdots & & & \\ X_{n1} & X_{n2} & \cdots & X_{nj} \end{bmatrix} \quad (9)$$

In both $\mathbf{Y}$ and $\mathbf{X}$, n = 1,2 . . . 66 plots in one year and 151 plots in another. Each plot or location is considered a site and independent and dependent variables were measured at each. In $\mathbf{Y}$, j = 1,2, . . . 5 dependent variables: protein yield, grass and sedge composition, perennial grass yield, phosphorus yield, and lignin composition. In $\mathbf{X}$, m = 1,2, . . . 11 topographic and edaphic variables: elevation, exposure, and slope and the soil variables of concentration or content of sand, rock, clay, phosphorus, organic matter, conductivity, and pH (Table 1). Many other variables could have been measured in the field, such as microclimatic variables, if unlimited funds were available. Many additional variables could be generated from powers and products of the existing 11 variables, however, for purposes of illustration only these 11 variables will be considered in this introductory study.

In the multiple linear regression analyses the $\mathbf{Y}$ matrix was considered columnwise so that in each univariate multiple regression analysis a vector, $\mathbf{B}$, of regression coefficients was selected so as to minimize the function

$$Q = (\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB}), \quad (10)$$

and was accomplished for each column vector by finding

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (11)$$

For the following discussion, each dependent variable is considered separately.

The relationship between the linear regression model and the linear programming model is as follows. The regression equation (5),

$$\mathbf{Y}_{1x1} = \mathbf{X}_{1xm} \mathbf{B}_{mx1}, \quad (12)$$

Table 1. Dependent and independent variables measured in individual plots on foothill range and used in regression and linear programming analyses.

Dependent		Independent	
Y_1	Protein yield	X_1	Elevation
Y_2	Grass + sedge composition	X_2	Exposure
Y_3	Perennial grass yield	X_3	Sand content of soil
Y_4	Phosphorus yield	X_4	Clay content of soil
Y_5	Lignin composition	X_5	Rock content
		X_6	Phosphorus in soil
		X_7	Organic matter in soil
		X_8	pH of soil
		X_9	Conductivity of soil
		X_{10}	Slope
		X_{11}	Soil depth

Objective function:

$$Y = 14. + .02X_1 - 11.X_2 - .78X_3 - 1.6X_4 + .47X_5 + .02X_6 + 3.4X_7 - 4.6X_8 - 22.X_9 - .09X_{10} + 6.0X_{11}$$

Constraints:

4780.	$\leq X_1 \leq$	5980.	.3	$\leq X_7 \leq$	56
0	$\leq X_2 \leq$	2.0	5.8	$\leq X_8 \leq$	8.0
46.	$\leq X_3 \leq$	93.	0.	$\leq X_9 \leq$	18.
2.	$\leq X_4 \leq$	19.	1.	$\leq X_{10} \leq$	55.
1.2	$\leq X_5 \leq$	62.	2.	$\leq X_{11} \leq$	24.
33.	$\leq X_6 \leq$	222.			

$$X_3 + X_4 \leq 100$$

$$.01X_1 - 6.1X_2 - 1.7X_3 - .88X_4 - .36X_5 + .03X_6 - 3.4X_7 - 4.4X_8 + 3.2X_9 - .32X_{10} - .24X_{11} \leq -150.$$

$$.13X_1 - 79.X_2 - 16.X_3 - 12.X_4 - 6.9X_5 + .26X_6 - 39.X_7 + 140X_8 - 519.X_9 - 4.3X_{10} + 24.X_{11} \leq 73.$$

$$.01X_1 + .01X_2 - .02X_3 - .03X_4 + .01X_5 - .01X_6 - .07X_7 - .17X_8 + 4.7X_9 - .01X_{10} + .08X_{11} \leq -2.1$$

$$.01X_1 - 1.0X_2 + .05X_3 - .03X_4 + .01X_5 + .01X_6 + .59X_7 - .28X_8 - 20.X_9 + .04X_{10} + .12X_{11} \leq 3.4$$

Table 2. The objective function and constraints of the linear programming model for determination of site characteristics (X_i) for optimum crude protein yield (Y).

becomes the objective function,

$$f = \mathbf{X}_{1xm}\,\mathbf{B}_{mx1}, \qquad (13)$$

which is to be maximized according to the constraints (6),

$$\mathbf{A}_{nxm}\,\mathbf{X}_{1xm} \leq \mathbf{C}_{nx1}, \qquad (14)$$

where **A** and **C** respectively are a matrix and a column vector.

Also, the linear programming model requires the following constraints which are consistent with the values of variables measured in real life,

$$\mathbf{X}_{1xm} \geq \mathbf{O}_{1xm} \quad . \qquad (15)$$

The multiple linear regression model (Table 2) shows the relationship between protein production and 11 topographic and edaphic site variables. This equation, less the constant term, becomes the objective function for the linear programming model. Values for the site variables are selected to maximize this function subject to the constraints that the variables for each site are within the limits found in the field for that site (22 constraints), that inherent relationships among these site variables are satisfied (1 constraint), and that additional inequalities (described below) are satisfied so that certain nutritional and management criteria are met (4 constraints).

The Optimum Site

He have used an optimization technique to determine maximum protein yields under a given set of conditions. Specifically, the objective in this problem was to produce protein for utilization by cattle and sheep during the nonwinter period i.e., to maximize Y_1 (Table 1) subject to various constraints. Important economical and biological constraints were: (1) Sites having a higher than average grass and sedge composition in the herbage were being sought in contrast to those having a large percentage of woody vegetation. (2) A higher than average percentage of grass and sedge alone is inadequate for the selection of a site; an additional constraint was imposed that the site must have better than average grass and sedge yield. (3-4) Other constraints, based on nutritional criteria, were that the site must have better than average phosphorus yield as well as having herbage with less than average lignin concentration. The four multiple linear regression equations relating site factors to grass and sedge composition and phosphorus yield, and lignin concentration were used to derive these inequality constraints. This was accomplished by using the appropriate mean value of the parameter as the Y term in the regression function, and then the constant term was subtracted from both sides of the inequality.

Although highly simplified models were used in this illustrative example, the value of these methods of analysis is illustrated when comparing the predicted optimum protein yield with the average yield which was measured. The value of the objective function for the optimum solution was a protein yield of 129 lb/acre. This compares to the measured range of protein yield from 24 to 211 lb/acre, with a mean of 77 lb/acre.

Because important powers and products of independent variables were omitted from the regression functions, the values for the site factors of the optimum site may or may not be entirely realistic. The values for site factors for the "optimum site" for protein production were at or near the maximum values found in the field for soil phosphorus content, pH, and soil depth. The optimum site values were at or near the minimum field values for elevation, soil organic matter, and sand and clay (implying a relatively high silt content). The optimum site would be nearly level and would be on north to east exposures. Values for soil conductivity and rock content for the optimum site would be intermediate to the extremes measured in the field.

The above conditions apply, of course, to the constraints used in this particular model. Altering the constraints, even though using the same objective function, would lead to a different set of values for site factors. The impact of each constraint on the solution could be evaluated by adding one constraint at a time in