METHODOLOGY FOR BUILDING SOIL BASED VEGETATION PRODUCTIVITY EQUATIONS: A STATISTICAL APPROACH[1]

by:

Jon Bryan Burley[2]

Abstract. Reclamation specialists have been interested in developing predictive equations to assess reclamation efforts in reconstructing soils to support vegetation growth. One predictive effort is associated with a statistical approach examining somewhat large data sets containing plant growth yields and soil variables. While the results from such procedures have been reported for the last seven years, a description of the methodology has not been described since 1987. This paper describes this statistical vegetation productivity model building process. To complete the basic analytic steps in the process, a statistical computing software package is required to conduct principal component analysis (PCA) and multiple regression analysis. The field data required to conduct the analysis are extensive. All crops and woody plants of interest (source of dependent variables) should be grown on all soil profiles (source of independent variables) for a period of approximately 10 years. The time period should include dry years, wet years, and average moisture years and should ideally include reclaimed and undisturbed soils. For any individual investigator, this type of data set would be expensive and time consuming to generate; however, the former United States Soil Conservation Service (SCS) has conducted similar work for a fair number of counties in the United States of America and can provide a substantial portion of the undisturbed soil database for investigators interested in developing a vegetation productivity equation for their region. Different units of measure across crop types are not necessarily an issue in vegetation productivity equation modeling, because each vegetation type is standardized to a mean of zero and a variance of 1. Then the various crop types and woody plants are examined with PCA. This statistical treatment allows the investigator to determine the number of dimensions necessary to explain the variance across all vegetation types. Ideally, if all of the crops of interest covary together, they can be combined into one dimension generating one dependent variable; otherwise an investigator may have to develop an equation for each significant dimension indicated in PCA. Soil factors suitable for regression analysis are calculated by employing a soil profile weighting formula. Before conducting regression analysis, a regression screening procedure may be employed to search for the most promising main effect, squared terms, and two variable interaction terms. The Maximum-R-squared improvement technique has been determined to be the best stepwise selection procedure to search for the best equation. Once a regression equation is selected, it can be further analyzed with bootstrap, subsampling, and jackknife statistical procedures. Finally, developed equations should be evaluated with results from reclaimed soils. These procedures for the basis of the methodology.

Additional Key Words: landscape planning, biometric statistics, soil science, prime farmland reclamation, agroecology

Introduction

This paper describes the fundamental procedures specific to the vegetation productivity model building process. To complete the basic analytic steps in the process, a statistical computing software package is required to conduct principal component analysis (PCA) and multiple regression analysis. In my efforts, I used the Statistical Analysis System (SAS) software for the microcomputer (1985); however, for

[1]Paper presented at the 1996, Annual Meeting of the American Society of Surface Mining and Reclamation, Knoxville, Tennessee

[2]Landscape Architecture Program, Department of Geography, College of Social Science, Michigan State University, E. Lansing, MI 48824 517/353-7880

Proceedings America Society of Mining and Reclamation, 1996 pp 789-798
DOI: 10.21000/JASMR96010789

investigators wishing to apply the methodology similar statistical software packages may suffice.

Table 1. Dependent variables and units of measurement as recorded and published by the U.S. Soil Conservation Service (Jacobson 1982).

Abbreviation	Vegetation	Measured Average Yield
Evergreen Trees		
JV	*Juniperus virginiana*	feet/20 years
PG	*Picea glauca densata*	feet/20 years
PP	*Picea pungens*	feet/20 years
PS	*Pinus ponderosa scopulorum*	feet/20 years
CO	*Celtis occidentalis*	feet/20 years
Deciduous Trees		
FP	*Fraxinus pennsylvanica*	feet/20 years
PD	*Populus deltoides*	feet/20 years
ST	*Salix alba tristis*	feet/20 years
UP	*Ulmus pumila*	feet/20 years
Deciduous Shrubs		
CA	*Caragana arborescens*	feet/20 years
CR	*Cornus sericea*	feet/20 years
PA	*Prunus americana*	feet/20 years
PV	*Prunus virginiana*	feet/20 years
SV	*Syringa vulgaris*	feet/20 years
Agronomic Crops		
SW	Spring Wheat	bushels/acre
BA	Barley	bushels/acre
OA	Oats	bushels/acre
SF	Sunflowers	pounds/acre
SB	Sugarbeets	tons/acre
SN	Soybeans	bushels/acre
GE	Grasses/Legumes	tons/acre

1 meter = 3.281 feet; 1 foot = 0.3048 meter
1 hectoliter = 2.837 U.S. bushels;
1 U.S. bushel = 0.363 hectoliter
1 hectare = 2.471 acres; 1 acre = 0.405 hectare
1 kilogram = 2.2046 pounds avoirdupois;
1 pound = 0.4536 kilogram

I (Burley 1988) reviewed the historical development leading to the rise of predictive reclamation modeling as a tool to assist in pre/post-mining landscape planning and design. Significant contributions leading to the development concerning predictive reclamation equations include works by Neill (1979), Pierce *et al.* (1983), Lohse *et al.* (1985), Walsh (1985), Vories (1985), Doll and Wollenhaupt (1985) and Plotkin (1986). In addition to these investigations, Potter (1986) describes two methods to assess the vegetation productivity capacity of the landscape. In the first method an empirical "shot-gun" approach, labeled by Potter as "inductive," is used where a wide array of variables is examined by sampling the landscape and making statistical comparisons/ inferences. The second method is a heuristic approach, labeled by Potter as "deductive," where transects are sampled and the investigator develops generalities about the physical and chemical parameters governing vegetation potential of the landscape. Complementing Potter's descriptions, Reith (1986) provides an explanation of the fundamentals concerning reclamation models across a broad spectrum of reclamation applications ranging from the prediction of rill formation through multiple regression techniques to multi-equation stochastic ecosystem modeling of a grassland. These papers, published by the mid-1980s, represent the general knowledge base prior to the actual development of statistical vegetation productivity models.

Burley and Thomsen (1987) describe a discrete methodology to produce a quantitative reclamation productivity equation. Unfortunately, this approach is labeled by Potter (1986) as the "shot-gun" approach. I would like believe that scientists have been working for generations posing hypotheses and examining variables to determine their importance. Eventually, a somewhat small set of potential predictor variables can be examined in greater refined statistical detail. While some individuals may wish to consider this a shot-gun approach, I would characterize this approach as an indicator of the maturity of soil science and reclamation activities allowing multiple variable studies to be conducted. This advanced multi-variable complexity is evident in other disciplines such a econometrics, wildlife habitat modeling, water quality prediction models, and visual quality modeling.

In my reclamation modeling efforts, the basis for this methodology originated with multivariate statistical concepts presented by Kendall (1939), requiring computationally complex matrix algebra (see Johnson and Wichern 1988). With the advent of the computer to perform matrix algebra operations for dimensions greater than three, multivariate statistical techniques made reclamation productivity development possible. By computing eigenvalues and eigenvectors for all possible dependent variables such as crops and woody plants, an investigator could determine the extent of multi-variable covariance and develop an equation to represent a linear combination of variables to generate a single dependent variable. In other words,

if the first eigenvalue was relatively large and the coefficient loadings of the eigenvector for the first eigenvalue were relatively similar, a simple equation derived from the eigenvector would suggest a linear combination of dependent variables that can be combined into one value per soil type. With one dependent variable value per soil type, it is possible to perform multiple regression analysis using one dependent variable. Gersmehl and Brown (1990) employed Kendall's method to examine multiple crop productivity values across geographic regions in the United States of America. Their work suggests that in the Midwest and Northern Great Plains, multiple crop productivity values often covary across soil types. Their work substantiates the concept that the soil preferences of various agronomic crops are indeed similar. Table 1 lists the vegetation types in Clay County, Minnesota that have been employed by Burley and colleagues to generate dependent variables, generally covarying as a group.

These plant types are employed to predict vegetation productivity. However, the term "vegetation productivity" is a relatively weakly developed construct. In many respects vegetation productivity has been operationally expressed as vegetation yield, such as bushels per acre of harvested seed or feet of new apical terminal shoot growth per year and represents a certain anthropocentric perspective concerning plant growth. A plant physiologist may suggest that an abundance of seeds per acre does not necessarily mean that a vegetation type is internally healthy and an ecologist may suggest that unsustainable lush plant growth is not necessarily a sound ecological condition. Consequently I recognize that there exists the potential to develop new operational constructs for vegetation productivity. Nevertheless, in my vegetation productivity work, I have made that assumption that existing measures of vegetation yield and new plant growth are reasonable indicators of productivity and that my interests lie in the relationships between existing productivity measures and soil parameters. I also assume that these variables can be studied with multiple regression analysis.

In the multiple regression analysis portion of the model building methodology, a single independent variable value for each soil parameter was generated by applying a weighting formula (Equation 1 in Figure 1) suggested by Doll and Wollenhaupt (1985), where the soil parameters in the first foot of a soil profile contribute 40% of a plant's vegetation production, the second foot contributes 30%, the third foot contributes 20%,. the. fourth. foot.. contributes 10%, and the

[Eq 1]

$$V=[(\sum_{i=1}^{12} v_i)*0.4]+[(\sum_{i=13}^{24} v_i)*0.3]+[(\sum_{i=25}^{36} v_i)*0.2]+[(\sum_{i=37}^{48} v_i)*0.1]$$

Where:

V = Weighted Soil Variable Value

v_i = Value for Soil Variable in One Inch Layer of Soil Profile

i = Soil Layer in Profile

Figure 1. Weighting equation based upon soil depth.

remaining layers do not contribute to vegetation growth. With this formula, any soil parameter for a specific soil profile can be measured on a foot by foot basis (even inch by inch) and the investigator can generate a single value for each soil profile, such as a single weighted pH value or a single weighted bulk density value (see Burley and Thomsen 1987). Table 2 lists the typical soil variables employed by Burley and colleagues to generate independent variables. A document written by the Soil Survey Division Staff (1993) describe current methods to measure these variables.

It is also important to recall that most land-use disturbances do not typically affect some plant growth variables, such as climate. Instead, disturbances associated with surface mining activities usually affect the soil. Thus for effective reclamation, a reclamation success predictor such as an equation should focus upon the environmental feature that has been disturbed, the soil. Some investigators and reviewers of vegetation productivity papers have confused "real time crop-yield indexes" with reclamation productivity equations. While real time crop-yield equations can compute the predicted level of vegetation production for a particular year under specific field conditions experienced over the growing season, a reclamation productivity equation predicts the average expected yield across many years of cultivation. This average yield is produced by employing crop yield values that were measured over many years including drought years, wet years, warm growing seasons, and cold growing seasons. This averaging effect thereby negates the yearly variances upon crop yields produced by climate, allowing an investigator to study more closely the influences of soils upon vegetation growth over many growing seasons.

Burley et al. (1989) applied the multiple regression analysis statistical approach to produce a

Table 2. Main effect independent variables and units of measurement from the U.S. Soil Conservation Service (Jacobson 1982 and U.S. Department of Agriculture 1951).

Abbreviation	Factor	Unit of Measurement
FR	% Rock Fragments	Proportion by weight of particles > 7.62 cm
CL	% Clay	Proportion by weight
BD	Bulk Density	Moist Bulk Density g/cm cubed
HC	Hydraulic Conductivity	Inches/hour (1 inch = 2.54 cm)
PH	Soil Reaction	pH
EC	Electrical Conductivity	Mmhos/cm
OM	% Organic Matter	Proportion by weight
AW	Available Water Holding Capacity	Inches/inch, cm/cm
TP	Topographic Position	Scale 0 to 5 Where: 0=Low (Standing Water) 2.5=Mid-slope 5=High (Ridge Lines)
SL	% Slope	(Rise/Run)*100

productivity equation for seven agricultural crops: spring wheat, barley, oats, soybeans, sunflowers, sugarbeets, and grasses/legumes. The database for this investigation was the Clay County soil survey (Jacobson 1982). The result was a reclamation productivity equation with a coefficient of multiple determination R^2 = 0.740. In other words, the regressors explain 74% of the sum of squares variation in the regression model. This equation did not consider woody plants and thus is not an all inclusive vegetation productivity model. Since reclamation often includes woody vegetation for the development of housing or commercial/industrial sites, wildlife habitat, agricultural shelterbelts, and forestry post-mining land-use applications, the development of a productivity model which includes woody plants would be more universally applicable in reclamation planning and design, including the development of prime farmland where woody plants composed of shelterbelts and windrows can be intricate components of an agricultural landscape. An equation developed by Burley (1991) using Burley and Thomsen's (1987) methodology, is presented as the best universal reclamation equation, because it was suitable to a lage number of vegetation types. R^2 for this equation is 0.795, explaining approximately 80% in the sum of squares variation for vegetation from the regression model. Other equations reported include a Clay County, Minnesota sugar beet (*Beta vulgaris* L.) equation (Burley 1990), two equations for Polk County, Florida (Burley and Bauer 1993), a two county equation in the Red River Valley of the North (Burley 1995a), and an equation for Oliver County, North Dakota (Burley *et al.* 1996). Burley (1992) presented a series of issues associated with these equations that may merit further investigation.

In contrast to the approach developed by Burley and Thomsen (1987), an alternative productivity index has been applied by several investigators. This approach is termed the "sufficiency approach" and follows more closely the work of Neill (1979) and Pierce *et al.* (1983). Huddleston (1984) and Henderson *et al.* (1990) review the formative development of this approach. With this approach, single independent variable models are developed to predict soil vegetation productivity. The single variable models have some degree of statistical reliance and may be normalized as illustrated by Wollenhaupt (1985). However, each variable is then combined into a full order interaction term where each independent variable is multiplied together as a group and may be corrected with the geometrical mean (Gale 1987). Several investigators have reported experiments with these full interaction term sufficiency equations (Barnhisel and Hower 1994, Burger *et al.* 1994, Barnhisel et al. 1992, Hammer 1992, and Gale *et al.* 1991). My criticism of this approach is that the equations are heuristically derived and the functions are not statistically validated and are therefore less rigorous. It is not surprising to me that efforts to corroborate equations based upon this methodology have met with mixed results. The research conducted with this heuristic approach has several limitations. First, the variables presented in the

equations may potentially be highly over specified as investigators have not demonstrated the contribution of each variable within the equation and have not followed searching procedures associated with regression modeling. Investigators have not accomplished sufficient equation sifting to eliminate linear models, squared terms, second-order interaction terms, or any other equation configuration. In my opinion, investigators have prematurely jumped to full interaction terms and geometric means. Second, the soils employed in the models are often restricted in external validity applications, meaning that significant results may be limited to a small set of soils studied, sometimes only two or three soil types. The data sets are not broad and thus are not applicable to many soil types. Third the vegetation types studied often consist of one dependent variable type, meaning the reported results are actually only applicable to the crop studied, such as corn or Eastern white pine. These three issues are my current reservations about this heuristic approach.

The sufficiency approach and the multiple regression modeling approach are two current vegetation productivity approaches. Regardless of the shortcomings for either approach they are models that may merit potential use in landscape planning. Numerous states require quantitative reclamation assessment procedures (primarily for coal surface mining reclamation on prime farmlands), suggesting that soil productivity equations are potentially compatible with these quantitative assessment demands and could make a contribution in evaluating the post-disturbance soil environment. Burley and Thomsen (1990) have described the application of a soil productivity equation for reclaiming surface mines. The application illustrates how these equations may be used to interpret landscape reconstruction configurations and how to evaluate the effectiveness of various reclamation treatments.

Discussion

The Dependent Variables

The field data required to conduct the analysis are extensive. All crops and woody plants of interest should be grown on all soil profiles of interest for a period of approximately 10 years to gain a perspective of vegetation performance on soils across climate variability. The time period should include dry years, wet years, and average moisture years. For any individual investigator, this type of data set would be expensive and time consuming to generate; however, the United States Soil Conservation Service (SCS) has conducted similar work for a fair number of counties in the United States of America and can provide a substantial portion of the database for investigators interested in developing a vegetation productivity equation for their region. Since the SCS operates under a county administrative structure, the data sets are naturally organized by county. These data sets are derived from soils on non-mined land. Ideally, a investigator should include results from reclaimed soils also. Neverthelss, I am somewhat disappointed in the conviction expressed privately by some colleagues that a statistical equation derived from soils not disturbed by mining is not valid for reclaimed soils, especially when many of these soils formed within the last 12,000 years, deposited by glacial and related surficial activity. Consequently, the soils that Burley and colleagues have used are relatively new and originated from relatively freshly disturbed materials. In addition, many of these soils are highly disturbed due to recurring fluvial processes, wind erosion, deep tilling, and soil amendments. Thus, these soils are not necessarily undisturbed material weathering in one location for millions of years, such as the soils that may be found in the tropics. Compared to millions of years, 12,000 years is a rather short time span, meaning the glacially deposited soils may have much in common with reclaimed surface mine soils. Concurrently, other than variations in physical parameters such as bulk density and chemical parameters such as nitrogen levels, no investigator has presented any information to suggest that reclaimed soils are intrinsically any different than pre-mine soils. No investigator has suggested that unmined versus mined soils should be a qualitative independent regression variable. If a reclaimed surface mine soil originated from substrate within the bounds of the soils studied to develop a soil productivity equation, and if the mine soil is within the physical and chemical bounds of soils studied to develop the soil productivity equation, then in theory, the soil productivity equation based upon non-reclaimed soils should be applicable to the reclaimed surface mine soil. The work of Burley et al. (1996) suggests that a such an equation may be able to predict productivity on reclaimed soils. Unless an investigator can find the physical or chemical variable that indicates reclaimed soils are substantially different, I will maintain there is no fundamental difference. In fact, within the horticulture and construction industry, soils are rebuilt everyday from highly disturbed areas to support turf growth, wood plants, and vegetable gardens. If these reconstructed soils were really any different from what we know and apply from non-mined soils, then theorctically, landscape construction

activities to build soils for lawns and gardens should not be successful. However, careful soil profile development on construction sites results in successful soil conditions for plant growth. Unless the reclaimed soils generate unsuspected toxic conditions such as high selenium values, surface mine soils are really no different than the collective properties of non-mined soils. For example, in the study areas I have examined, I have found dense clays, acidic clays, alkaline clays, clays on slopes, clays in wetlands, permeable clays, deep clays, and shallow clays, replicating almost any non-toxic soil condition found on reclaimed clay soils. This broad variability of the data set allows reclaimed soils with properties that fit within the parameter bounds of this data set to estimate vegetation performance. I would suggest that those studies that have been unsuccessful at predicting vegetation productivity have either been working with heuristic equations or have not generated a data set with substantial breadth across many years with many soil types and many vegetation types to develop statistically predictive results.

In my studies, crop harvest data and woody plant growth rates are the typical dependent variables employed in a vegetation productivity equation study. Crop harvest data may be in bushels per acre, tons per acre, pounds per acre or any other quantitative harvest value typical for the crop of interest. The woody plant growth rates have been presented by the SCS in feet per years of growth, although other forest measurements or horticultural vegetation growth rates or plant volumes could potentially be employed. In addition, the crop logical types do not have to be consistent. For example, a woody crop species and a mixed crop such as a grasses and legume mix can be employed in the study. The analysis will indicate the statistical relationship between the various vegetation types. This is a difficult concept for some investigators to intellectually grasp because apples and oranges plus many other types of numerical information can actually be compared and combined. Different units of measure across crop types are not necessarily an issue in vegetation productivity equation modeling, because each vegetation type is standardized to a mean of zero and a variance of 1. Then the various crop types and woody plants are examined with PCA. This statistical treatment allows the investigator to determine the number of dimensions necessary to explain the variance across all vegetation types. Ideally, all of the crops of interest covary together and can be combined into one dimension; otherwise an investigator may have to develop an equation for each significant dimension indicated in PCA. In other words, PCA is a data reduction tool that may allow an investigator to determine whether corn, soybeans, wheat and *Fraxinus pennsylvanica* can be combined together or must be analyzed separately. This is an important issue. If all vegetation types do not covary in productivity, then the investigator must develop a large number of individually tailored vegetation productivity equations and the reclamation specialist may reclaim a landscape suitable for one crop but not suitable for another, thereby excluding future production options for the farmer. If the crops do covary, a universal vegetation productivity equation may be possible for the study site.

PCA results typically begin with presentation of an eigenvalue for each dimension in the data set. The largest number of dimensions is equal to the number of variables present in the data set. For example if three crop types are presented for PCA, then the largest number of dimensions is three. This approach is extremely useful for a large number of variables, because the mathematics employed in the technique can examine a data set in a multidimensional space greater than three dimensions. Until the development of the computer, PCA was often limited to three dimensions as the matrix algebra required to compute the eigenvalues was difficult to compute by hand. The important feature of the eigenvalue is that each dimension is orthogonal to every other dimension, meaning that the information associated with each eigenvalue is independent of every other dimension. PCA assumes that the latent roots of the data set are definite and real and the data set is composed of multivariate-normal variables.

The largest eigenvalue is presented as the first eigenvalue in PCA. With standardized variables, the eigenvalue can be no larger than the sum of the variables under study. For example if five crop types are being studied, the largest eigenvalue can be no greater than 5.0. In addition, the sum of all the eigenvalues can be no greater than 5.0. This means that if the largest eigenvalue is 4.8, the sum of the remaining eigenvalues must be equal to 0.2. The proportion of the sum for any combination of eigenvalues indicates the amount of variance explained by those eigenvalues. Suppose two eigenvalues from standardized variables sum to 3.0 and there are four variables in the analysis. These two eigenvalues represent 75 percent of the variance in the data set.

Eigenvalues greater than 1.0 originating from standardized variables are considered to represent significant dimensions. The significant dimensions are then inspected by examining the eigenvector

coefficients. Each variable in the analysis contains an eigenvector coefficient associated with each eigenvalue. The eigenvector coefficient indicates the strength of association the variable has with the eigenvalue. In vegetation productivity analysis studies, the investigator is interested in which variables are associated with which dimensions. Burley *et al.* (1989) discovered that crop variables in their study covaried together and that the eigenvector coefficients were relatively equitable across the first eigenvalue. Thus Burley *et al.* (1989) were able to develop a linear equation to compute the generation of one combined plant growth productivity value per observation case. The eigenvector coefficients indicate the weighting of each crop variable for a specific dimension and range in value from 1.0 to -1.0. The crop value multiplied by the weighting coefficient and then summed with the results from the other crops forms a linear combination equation to build a combined vegetation productivity value for each observation case. Consequently, an investigator may be able to build a data set containing one dependent variable per observation case from a set of many dependent variables.

The Independent Variables

The independent variables are comprised of soil characteristics from each soil profile of interest. A major issue associated with soil characteristics is the variability of any one soil parameter within any soil profile. Where should the investigator measure soil reaction? Doll and Wollenhaupt (1985) suggest that any soil parameter should be measured through the first four feet of the profile and weighted according to the equation identified in Equation 1. While some researchers may question the accuracy of the weighting formula, this formula represents the current understanding and state-of-the-art of profile contribution in vegetation growth, and therefore this is the weighting equation employed in my work.

Analysis

Once the soil profile weighting formula has been employed and the dependent variables have been constructed, the data set is ready for regression screening. In SAS, regression screening can be accomplished with the RSREG procedure. This procedure examines main effect, squared terms, and two variable interaction terms for association with the dependent variable. The most promising variables can then be entered into a regression stepwise procedure to evaluate various combinations of regressors.

The Maximum-Rsquared improvement technique has been determined to be the best stepwise selection procedure (SAS 1982). The investigator must then examine the results of the stepwise procedure to assess which equation is the best from those presented. Typically the investigator will examine the multiple coefficient of determination, preferring larger values which explain a larger percentage of the variance in the data set. Values near 0.4 are considered weak and values near 0.9 are considered strong. In addition, the investigator will examine the p-value for the overall regression and the p-value for each regressor. I recommend that one examine p-values that originate from Type II Sum of Squares, a technique which computes the p-value for the regressor assuming that all other regressors are already in the equation. This is a conservative approach to determining p-values and is consistent with not overestimating the p-value's significance. P-values less than or equal to 0.05 are considered statistically significant. P-values less than or equal to 0.01 are considered highly significant. Therefore, the investigator is searching for an equation which has at least significant p-values and explains as much of the data set's variance as possible.

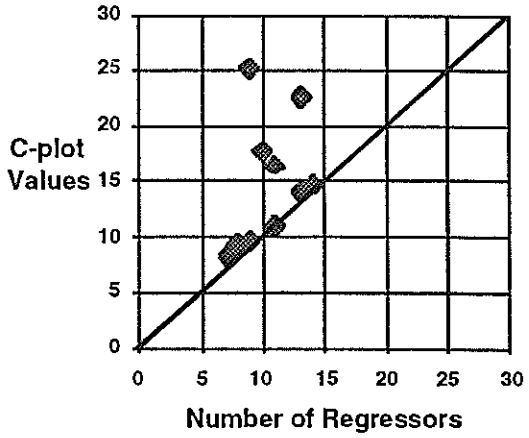

Figure 2. Plot of regressors and C-plot values for equations presented by Burley (1995b).

However, the investigator must be concerned about over specifying and multicollinearity issues associated with regression modeling. One approach in evaluating various equations is to prepare a C-plot of the equations, where the number of regressors are plotted against the C-plot value (Younger 1979). The equations above the 45 degree line are considered equations which are not over specified. Equations that approach the 45 degree line without crossing into the lower portion of the quadrant are preferred over

equations distant from the 45 degree line (Figure 2). Those equations which are further from the origin and along the 45 degree line are preferred over equations close to the origin. This technique can remove much of the collinearity problems inherent in regression modeling. One can also conduct collinearity diagnostics to determine the collinearity associated with any regression equation (SAS Institute 1982). The diagnostic procedure produces a condition index number indicating the degree of collinearity associated with dimensions of a regression equation. A condition index of less than 10 is considered to indicate relatively little or no collinearity (Rawlings 1988). A value greater than 10 indicates weak collinearity. A condition index number between 30 to 100 indicates moderate to strong collinearity. Values above 100 indicate serious collinearity problems. With each condition index, weighting coefficients indicate the level of association for regressor variables. Regressors containing coefficients approaching 1.0 are strongly associated with that particular dimension. More than one regressor with substantial values for a single condition index identifies the potential collinear variables.

[Eq. 2]

Pseudo-Value = k * (parameter estimate calculated from the whole sample) - (k-1)*(parameter estimate calculated from the group with the j sample removed)

Where: k = total number of observation cases

Figure 3. Nonparametric pseudo value jackknife equation.

[Eq. 3]

Standard Error = (Variance of the Pseudo-values/(k-1))$^{0.5}$

Where: k = total number of observation cases

Figure 4. Nonparametric standard error estimate equation for jackknife procedure.

Analysis of the Analysis

There are several methods to evaluate the estimates of parameters and validity of a particular regression equation. One method is the jackknife approach (Efron 1982). With this technique a nonparametric statistic is produced for each regression parameter by applying Equation 2 in Figure 3. The variance for this sampling procedure is employed to calculate the standard error associated with the estimates of each parameter (Equation 3 in Figure 4, Mathsoft 1988).

In contrast to the jackknife approach, the bootstrap technique samples the data set with replacement to build a larger observational set to compute parameter estimates (Mathsoft 1988 and Efron 1982). For bootstrap analysis, the standard deviations of the coefficient estimates are the standard error estimates for the coefficients. In other words, an estimated distribution of the coefficients can be calculated. Wide distributions indicate unstable coefficients and narrow distributions indicate relatively stable coefficients.

Another procedure which can be employed to study the applicability of a regression equation is the subsampling procedure, where a portion of the data set is removed from the equation development process and then employed later to compare the results from the equation with the subsample.

In addition to the collinearity diagnostics, jackknife techniques, bootstrap techniques, and subsampling procedures, one can inspect the plots of the residuals associated with a specific regression equation. The plots can identify regression assumption violations concerning constant variance and histogram plots can identify violations concerning the normality of the residuals.

Finally, selected vegetation productivity equations developed with the procedures described by Burley and Thomsen (1987) should be assessed with data sets from reclaimed soils. Considering the number of reclamation research centers in North America, one might expect that suitable data sets exist for conducting a reclaimed soil investigation. However, these potential data sets often do not contain observations according to depth, nor do they contain all of the desired soil variables needed by the developed equation. While equations derived from non-mined soils are certainly possible to construct, the methodology described in this paper is directly applicable to unreclaimed data sets also, providing the data set is comprised of all vegetation types across all disturbed soil types. However, it appears that currently, no reclamation center or reclamation investigator has such a large and extensive data set.

Concluding Remarks

This technique represents the basic approach I have folllowed in developing vegetation productivity equations and presents a few more assessment procedures not initially described by Burley and Thomsen (1987). It is my belief that investigators across the globe and especially in North America will employ the methodology that I have described in this paper and build regionally specific equations to study the reconstruction of soils to support plant growth..

Bibliography

Barnhisel, R.I. and J.M. Hower. 1994. The use of productivity index to predict corn yields on restored prime farmland. International Land Reclamation and Mine Drainage Conference and Third International Conference on the Abatement of Acidic Drainage, Volume 3: Reclamation and Revegetation. United States Department of the Interior, Bureau of Mines Special Publication SP O6C-94:20-27.

https://doi.org/10.21000/JASMR94030

Barnhisel, R.I., J.M. Hower, and L.D. Beard. 1992. Development of a soil productivity model for use in prime farmland reclamation. Dunker, R.E., R.I. Barnhisel, and R.G. Darmody (Eds.) In: Proceedings of the 1992 National Symposium on Prime Farmland Reclamation. Department of Agronomy, University of Illinois, Urbana, Illinois 61801, :205-211.

Burger, J.A., J.E. Johnson, J.A. Andrews, and J.L. Torbert. 1994. Measuring mine soil productivity for forests. International Land Reclamation and Mine Drainage Conference and Third International Conference on the Abatement of Acidic Drainage, Volume 3: Reclamation and Revegetation. United States Department of the Interior, Bureau of Mines Special Publication SP O6C-94:48-56.

https://doi.org/10.21000/JASMR94030048

Burley, J.B. 1988. Development and Application of an Agricultural Soil Productivity Equation for Reclaimed Surface Mines in Clay County, Minnesota. M.L.A. Thesis, University of Manitoba.

Burley, J.B. 1990. Sugarbeet productivity model for Clay County Minnesota. Journal of Sugar Beet Research 27(3 & 4):50-57.

https://doi.org/10.5274/jsbr.27.3.50

Burley, J.B. 1991. A vegetation productivity equation for reclaiming surface mines in Clay, County Minnesota. International Journal of Surface Mining and Reclamation 5:1-6.

https://doi.org/10.1080/09208119108944279

Burley, J.B. 1992. Vegetation productivity equations: an overview. Dunker, R.E., R.I. Barnhisel, and R.G. Darmody (Eds.) In: Proceedings of the 1992 National Symposium on Prime Farmland Reclamation. Department of Agronomy, University of Illinois, Urbana, Illinois 61801, 259-265.

Burley J.B. 1995a. A multi-county vegetation productivity equation for soil reclamation. Hynes, T.P. and M.C. Blanchette (Eds.) In: Proceedings of Sudbury '95 - Mining and the Environment. CANMET, Ottawa, Volume 3, 1113-1122.

Burley, J.B. 1995b. Vegetation productivity soil reclamation equations for the North Dakota coal fields. PhD. dissertation, University of Michigan.

Burley, J.B. and A. Bauer. 1993. Neo-sol vegetation productivity equations for reclaiming disturbed landscapes: a central Florida example. Zamora, B.A. and R.E. Conally (Eds.) In: The Challenge of Integrating Diverse Perspectives in Reclamation: Proceedings of the 10th Annual National Meeting of the American Society for Surface Mining and Reclamation. ASSMR, Spokane, WA, 334-347.

https://doi.org/10.21000/JASMR93010334

Burley J.B., K.J. Polakowski, and G. Fowler. 1996. Vegetation productivity equation for reclaiming surface mines in Oliver County North Dakota. In: The 1996, Annual Meeting of the American Society of Surface Mining and Reclamation, Knoxville, Tennessee.

https://doi.org/10.21000/JASMR96010393

Burley, J.B. and C.H. Thomsen. 1987. Multivariate Techniques to Develop Vegetation Productivity Models for Neo-Soils. 1987 Symposium on Surface Mining, Hydrology, Sedimentology and Reclamation. University of Kentucky, 153-161.

Burley, J.B. and C.H. Thomsen. 1990. Application of an agricultural soil productivity equation for reclaiming surface mines: Clay County, Minnesota. International Journal of Surface Mining and Reclamation 4:139-144.

https://doi.org/10.1080/09208119008944181

Burley, J.B., C.H. Thomsen, and N. Kenkel. 1989. Development of an agricultural soil productivity equation for reclaiming surface mines in Clay County, Minnesota. Environmental Management 13(5):631-638.

https://doi.org/10.1007/BF01874969

Doll, E.C. and N.C. Wollenhaupt. 1985. Use of soil parameters in the evaluation of reclamation success in North Dakota. Bridging the Gap Between Science, Regulation, and the Surface Mining Operation. ASSMR Second Annual Meeting, Denver, CO, 91-94.

https://doi.org/10.21000/JASMR85010094

Efron, B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial Applied Mathematics.

Gale, M.R. 1987. A forest productivity index model based on soil-root distributional characteristics. Ph.D. Dissertation, University of Minnesota, St. Paul.

Gale, M.R., D.R. Grigal, and R.B. Harding. 1991. Soil productivity index: Predictions of site quality for white spruce plantations. Soil Science Society of America Journal 55:1701-1708.

https://doi.org/10.2136/sssaj1991.03615995005500060033x

Gersmehl, P.J. and D.A. Brown. 1990. Geographic differences in the validity of a linear scale of innate soil productivity. Journal of Soil and Water Conservation, May-June:379-382.

Hammer, R.D. 1992. A soil-based productivity index to assess surface mine reclamation. Prime Farmland Reclamation. Prime Farmland Reclamation National Symposium, University of Illinois, :221-232.

Henderson, G.S., R.D. Hammer, and D.F. Grigal. 1990. Can measurable soil properties be integrated into a framework for characterizing forest productivity? Sustained Productivity in Forest Soils. Proceedings of the 7th North American Forest Soils Conference, University of British Columbia, Faculty of Forestry Publications, :137-154.

Huddleston, J.H. 1984. Development and use of soil productivity ratings in the United States. Geoderma 32:297-317.

https://doi.org/10.1016/0016-7061(84)90009-0

Jacobson, M.N. 1982. Soil Survey of Clay County, Minnesota. USDA, SCS and Minnesota Agricultural Experiment Station.

Johnson, R.A. and D.W. Wichern. 1988. Applied Multivariate Statistical Analysis. Prentice Hall.

Kendall, M. 1939. The geographical distribution of crop productivity in England. Journal of the Royal Statistical Society 102:21-48.

https://doi.org/10.2307/2980188

Lohse, J.S., P. Giordano, M.C. Williams, and F.A. Vogel. 1985 Illinois agricultural land productivity formula. In: Bridging the Gap Between Science, Regulation, and the Surface Mining Operation. ASSMR Second Annual Meeting, Denver, CO, 24-39.

https://doi.org/10.21000/JASMR85010024

Mathsoft. 1988. Statistics I: Tests and Estimation. Mathsoft, Inc, Cambridge, Massachusetts.

Neill, L.L. 1979. An evaluation of soil productivity based on root growth and water depletion. M.S. Thesis, Univ. of Missouri, Columbia.

Pierce, F.J., W.E. Larson, R.H. Dowdy and W.A.P. Graham. 1983. Productivity of soils: Assessing long-term changes due to erosion. Journal of Soil and Water Conservation 38:39-44.

Plotkin, S.E. 1986. Overseeing reclamation: From surface mine to cropland. Environment 28(1):16-20, 40-44.

https://doi.org/10.1080/00139157.1986.9929868

Potter, L.D. 1986. Pre-mining assessments of reclamation potential. Reith, C.C. and L.D. Potter (eds) In: Principles & Methods of Reclamation Science: with Case Studies from the Arid Southwest. University of New Mexico Press, 41-67.

Rawlings, J.O. 1988. Applied Regression Analysis: a Research Tool. Wadsworth & Brooks/Cole Advanced Books & Software.

Reith, C.C. 1986. Understanding reclamation with models. Reith, C.C. and L.D. Potter (eds) In: Principles & Methods of Reclamation Science: with Case Studies from the Arid Southwest. University of New Mexico Press, 85-107.

SAS Institute Inc. 1985. SAS Introductory Guide for Personal Computers, Version 6 Edition. SAS Institute.

SAS Institute Inc. 1982. SAS User's Guide: Statistics, 1982 Edition. SAS Institute.

Soil Survey Division Staff. 1993. Soil Survey Manual. U.S. Department of Agriculture, Handbook No. 18.

U.S. Department of Agriculture. 1951. Soil Survey Manual. U.S. Department of Agriculture, Handbook No. 18.

Vories, K.C. 1985. Proof of vegetative productivity: Research needs. 1985 Symposium on Surface Mining, Hydrology, Sedimentology, and Reclamation. University of Kentucky, :145-149.

Walsh, J.P. 1985. Soil and overburden management in western surface coal mines reclamation - findings of a study conducted for the Congress of the United States -Office of Technology Assessment. In: Bridging the Gap Between Science, Regulation, and the Surface Mining Operation. ASSMR Second Annual Meeting, Denver, CO, 257-264.

https://doi.org/10.21000/JASMR85010257

Wollenhaupt, N.C. 1985. Soil-water characteristics of constructed mine soils and associated undisturbed soils in southwestern North Dakota. Ph.D. dissertation, North Dakota State University, Fargo, North Dakota.

Younger, M.S. 1979. Handbook for Linear Regression. Duxbury Press, Wadsworth, Inc.