Journal of Plant Studies; Vol. 3, No. 2; 2014
ISSN 1927-0461 E-ISSN 1927-047X
Published by Canadian Center of Science and Education

Prediction From Archived Vegetation and Environmental Monitoring Data

David Scott[1]

[1] Lake Tekapo, New Zealand

Correspondence: David Scott, Lake Tekapo, New Zealand. E-mail: scottd_hc@xtra.co.nz

Received: June 24, 2014 Accepted: July 29, 2014 Online Published: July 31, 2014

doi:10.5539/jps.v3n2p91 URL: http://dx.doi.org/10.5539/jps.v3n2p91

Abstract

A comparison was made between analysis methods of archived monitoring data for extrapolating quantitative values or trends from additional partial measurements. These were done from both the 'species from environment and management' or 'environment and management from species' predictions. The sample data used was from two 19 year annually monitored grazing trials in a tussock grassland environment under 90 different controlled inputs, but treated as if from a random ecological survey. Quantitative estimation of trends from such a reference data set was best by a direct data search for closest matching observation, followed by response surface regression, simple multiple regression, multiple imputation, transformation regression, and least from principal component analysis.

Keywords: environment, vegetation monitoring, statistical analysis, prediction

1. Introduction

In recent decades there has been increasing interest in establishing the medium and long-term sustainability of different land-use practices and the desirability of establishing monitoring systems so that effects and trends can be quantified. This has been particularly true for conservation management and rangeland pastoral farming systems. Associated with this is a need to consider the form of the monitoring, its archiving, and how the resulting data can be used to determine trends or make predictions from monitoring in one situation to those in other like situations.

Trends in land ecosystem changes are probably detectable first in the state of surface litter and organic matter (Savory & Butterfield, 1999) on a time scale of months to years; secondly, from vegetation composition on a time scale of years to decades; and thirdly, from soil profile characteristics on the time scale of decades to centuries. Monitoring and the detection of vegetation changes are probably on a suitable time scale for practical land management to be able to detect trends in time to make management adjustments.

Quantification of trends is dependent on both the methods of initial monitoring of vegetation and environment, and analysis methods used to determine the relationships within those data. This can be either as a retrospective analysis of a particular data set, or in another situation as the basis for predicting some variables from measurement of others from a more complete reference data set. There has been a wide range of methods used for determining vegetation and environmental conditions differing greatly in their required skills, time resources, and resulting accuracy. A range of analytical biometric methods has been used to show the relationship or trends within such data. Generally only one method of analysis has been used in each investigation and there has been little comparison of the effectiveness of different methods on the same data set.

Interests in the present study were twofold: partly as a comparison of analysis techniques on the same data set, but more particularly how such base data can be used, in association with future measurements of some subset of the variables to make estimate/ predictions of the other variables in same or different situations.

Such estimates/predictions from a further subset of monitored variables can be from the perspectives of either "species from environment" (to predict species abundance from environmental variables), or "environment from species".

The opportunity was taken to investigate these aspects from the annual monitoring of vegetation from two long-term multivariate structured pasture trials in the tussock grasslands of New Zealand. These were on the effects of a factorial range of fertiliser and grazing managements on multi-species sown and resident species in

tussock grassland (Scott & Covacevich, 1987; Scott, 2001, 2005). The duration of the two trials resulted in a relatively large data set using one base method of vegetation assessment, but convertible into other forms of vegetation assessment. It was reasoned that a comparative analysis of a single base data set such as this, using different expressions of vegetation measurement and different forms of analysis, from a trial of the known but unused structure in the present context, would give some indication of the likely success and reliability of measurement and analysis method for unstructured general survey type monitoring. In one sense the relationship between variables were already known for this sample data set because of the experimental manipulative design of the trial. The question was more whether those relationships could have been determined if they had come from an unstructured random general survey.

The objective of the study was to compare the relative efficiency of different forms of analyses in making predictions from subsequent partial sample data. The scenario used was that, unknown to the investigator, another party had at some time conducted similar trials but had measured only the vegetation or environmental variables, or measured the vegetation on some of the present plots of the trials described but had failed to note the date or treatment from which they were taken, with the problem now being to correctly identify the treatment combination of the observation(s) and hence its environmental or vegetation conditions. A further restriction was placed on the present investigation in that only actual data from the present trials could be used to making those inferences. Thus the data was used as both "calibration" and "test" data.

The comparison was made from three different perspectives of: either having to predict species abundance from environmental (treatment) variables ("species from environment"); or treatment from species abundance ("environment from species"); or vegetation type from treatment. Following from that there is a discussion on the level of discrimination needed in monitoring and archiving.

2. Methods

2.1 Data Source

The sample data used derived from two trials which have been more fully described in references referred too. In brief, both trial were adjacent to each other at the AgResearch Mt John trial site, Lake Tekapo, New Zealand and were established by partial over-drilling of hieracium (*Hieracium pilosella*) infested fescue tussock (*Festuca novae-zelandiae)* grasslands in 1982, have been continued to the present with the data used from the first 19 years.

In one trial the main treatment blocks were two spatial replications of 5 fertiliser/growth regimes of nominally 0, 50, 100, 250 or 500 kg superphosphate ha^{-1} yr^{-1} (hereafter referred to as the Graze/fert trial). The superphosphate was sulphur fortified to 20-50% S at the 50 and 100 kg ha^{-1} yr^{-1} superphosphate rate, with double the fertiliser rate in the initial year. The 500 kg ha^{-1} yr^{-1} blocks also received fortnightly spray irrigation from late spring to autumn. Within the main treatment blocks were further grazing treatments of 3 relative stocking-rates (low, moderate or high in ratio of 1:2:4 sheep grazing days in Year 2-4, and 2:3:4 in subsequent years), by 2 grazing-methods (high sheep numbers for 5-7 days, i.e., mob-stocking; or few sheep for longer periods, i.e. sustained or set-stocking). Each fenced treatment combination was 8 x 50m.

The second trial was a response surface design of 27 combinations of S fertiliser (as elemental sulphur) and P fertiliser (as triple superphosphate with 20% P) of nominally 0, 5, 10, 20, 50 or 100 kg element ha^{-1} yr^{-1} (hereafter referred to as the PxS trial). Four of the combinations were repeated with the addition of potassium and micronutrients (40 kg K ha^{-1} yr^{-1} as potassium chloride, and 4 kg ha^{-1} yr^{-1} "BASF Fertilon" as mixed micronutrients (Mo, B, Co, Zn, Co, and Fe). Each fenced treatment combination was 12.5 × 12.5 m. The fertiliser rates were doubled in the initial sowing year with half the sulphur applied as gypsum in that year.

The fertiliser was applied annually in early spring. Annual variations in the proximate analysis of P and S contents of the superphosphate were taken into account. The actual mean fertiliser rate for the PxS trial was given in (Scott 1999, Figure 4). For the Graze/fert trial they were 0 + 0, 4.1 + 17.6, 8.9 + 26.0, 22.7 + 54.5, and 46.8 kg ha^{-1} yr^{-1} P and 114.8 kg ha^{-1} yr^{-1} S for the zero, low, moderate, high and high + irrigation treatments respectively.

The trials were sown with a common mixture of 25 different legumes and grasses, using a rotary hoe drill which cultivated and sowed into about a third of the area over which it passed (Scott & Covacevich, 1987; Scott, 1999). The plots were given nearly two growing seasons to establish before commencement of the grazing treatments in the second autumn.

From the 3rd year on, plots were grazed from November to May as required. For the PxS trial that involved common mob-stocking on two or three occasions for 4-8 days each time. For the Graze/fert trial the number of grazings varied with fertiliser rate. The plots were grazed in groups of three for each of the different stocking-rate

and grazing-method combinations. The plots were grazed as judged from the moderate stocking rate plot, with the other two plots receiving their differential stocking rate for the same period. For the moderate stocking rate treatments (and all the PxS plots) this was when feed on offer was 1-2 t DM ha, and ceased when plots were grazed to a residual 1-2 cm height.

A high proportion of the treatments in the Graze/fert trial had been deliberately "mismanaged" both to determine which pasture-species tolerated those managements, what was the composition of the resulting vegetation, and the degree to which the vegetation and other data could be used to detect such management trends in other like situations. Those included the "under-utilised" low stocking-rate treatments, and the "overgrazed" high stocking-rate treatments. The set-stocked treatments would have been considered "mismanaged" based on the experience and practice in intensive farmed areas. The PxS trial treatments had been given "best" management.

2.2 Vegetation Monitoring

The vegetation composition in each plot was recorded in November of each year from the 2nd year on, after the common ungrazed regrowth period from mid winter. Relative species contribution and vegetation composition were determined by the species ranking method (Scott, 1989). Species in each plot were visually ranked in order of their contribution to herbage bulk within that plot (up to 10th species), and simultaneously the ratio of the relative contribution of two of the species (mostly 5th to 1st ranked species), so that the proportional contribution of each of the species could be estimated.

Records were kept of duration of sheep grazing on each plot. Annual grazing-days ha^{-1}, and cumulative grazing-days ha^{-1} for up to five years previously, was determined for each plot.

2.3 Representation of Monitoring Data

The soil fertility level of each of the plots was expressed as the annual application rate of S and P fertiliser (each generally on a square root scale). Grazing treatments in the Graze/fert trial were coded (relative stocking rate: "low" or "under-utilised" = -1, "moderate" or "best guess" = 0, and "high" or "over-grazed" =1; grazing method: "mob" = 0, and "set" = 1). The corresponding values for all of the PxS trial plots were 0, 0. Time was recorded as years from the start of development. Curvilinear and interaction effects were accommodated by introduction of up to cubic effects for main effects and quadratic interaction variables. These treatment variables and interactions were treated as continuous quantitative variables in all the analyses.

A number of alternative representations of the species composition data were investigated. The base method recorded species on an ordinal scale (1st ranked species = 1, 2nd ranked = 2, ------, up to 10th ranked = 10, and all other presence or absence = 11), analysed as continuous quantitative values.

There were also converted to three other representations of individual plant species data. The first was to re-rank them on a reduced ordinal scale. The dilemma in devising any rank scale is what value to assign to species present but not within the main ranking series. Intuitively a value of zero seems logical, but if that is to be part of an ordinal sequence then it has to be compatible with values assigned to other ranks. To accommodate this requirement the values assigned to a 2-ranked-species situation was "1st ranked = 2, 2nd ranked = 1, lower ranked or absent = 0", for a 3-ranked-species situation was "1st ranked = 3, 2nd ranked = 2, 3rd ranked = 1, and lower ranked or absent = 0", and similarly for 5 ranked and 10 ranked species.

A second form for the individual species rank data was conversion to semi-quantitative data of their proportional contribution to the total herbage mass using their rank values and the simultaneously field estimated ratio of 5th to 1st ranked species as a determinant of the gradient of an assumed log/linear relationship between abundance and rank (Preston, 1948; t'Majente, 1953). A third form was to determine species presence /absence (1/0) in each rank position. The three alternative representations were treated as continuous quantitative variables.

The rank order of species within a sample could also be used to define a nominal vegetation type or class e.g. for species A, B, C ---- etc. vegetation types based only on the most dominant ranked species could be B, C, A; combined 1st and 2nd ranked class A/B, A/C, C/A, etc; combined 1st, 2nd and 3rd ranked class A/B/C, C/B/A, etc. With the number of species present in the trials this resulted in a rapidly increasing and unworkably high number of vegetation types, with most being absent, or of very low observed frequency. An alternative to this was a conditional agglomerative hierarchical subdivision based on the observed frequency, with subdivision at each level confined to designated vegetation types that had greater than about 1% frequency in the combined data set, and other species in a "catch all" category e.g. A/B, A/others. This alternative was only pursued to the 3rd ranked species.

2.4 Comparative Analysis

The emphasis in the comparative analysis was towards methods that were capable of both retrospective analyses of the particular data set, and for their potential, with additional partial data, of carrying forward quantitative estimates into some other data base or general predictive "search" or "expert" system. The relationship between environmental conditions (= treatments), plant species and vegetation variables was evaluated by a range of multivariate methods.

The variables in the sample data set were of four types. The S and P fertiliser rates, relative stocking rates or grazing methods could be regarded as independent variables in determining species and grazing capacity responses in a forward "species from environment" analysis. But alternatively they might be regarded as dependent variables to be predicted in a monitoring situation of an "environment from species" analysis. Time or years from start of development, is in one sense similar to the other variables, but in another sense is peculiar in that it can only increase and is likely to have an autoregressive or lag component. The species rank variables could be regarded as dependent variables within the context of the particular trials, but as independent variables in an "environment from species" analysis. Their other characteristic is that they are relative or compositional in nature, even though they have been treated as though independent i.e. the value for one species is dependent on the relative value of other species. On the shorter time scale of yearly increments the achieved sheep grazing capacities could be viewed as dependent variables relative to the pasture species and treatment variables. But on longer time scales, of five or so years, there was an inter-action of past grazing capacity and current pasture species variables. Thus there was ambiguity in the relative treatment of pasture species and grazing capacity variables as independent or dependent variables.

A range of multivariate methods were compared for making estimates or predictions. The approach used was to add partial observations of some known variables to the data set (as if coming from a query) and, using missing values methods, to estimate the values of other unknown or query variables that were wished to be predicted.

Table 1. Mean and standard deviation of variables for combined data from both trials. Species variable = rank abundance: 1 =1st, 2 =2nd, -----, 10 = 10th, 11= absent. (n=1418) (from Table 3 – Scott, 2005)

Code	Species variables	Mean	STD	Code	Environmental variables		
Lp	*Lupinus polyphyllus*	3.89	3.66	t	Years from start of development	9.0	4.9
Ta	*Trifolium ambiguum*	5.47	3.45	t^2	- quadratic years	105	90.8
Tr	*T. repens*	8.64	2.95	t^3	- cubic years	1377	1518
Th	*T. hybridum*	7.74	3.61	r	Stocking rate (1= low, 0 = mod., 1 = high)	-0.022	0.65
Tp	*T. pratense*	9.00	2.90	m	Stocking method (0 = mob, 1 = set)	0.31	0.46
Tm	*T. medium*	10.36	1.25	i	Irrigation (0 = nil, 1 = used)	0.13	0.34
Lc	*Lotus corniculatus*	9.7	2.05	p	Annual P fertiliser rate (kg P ha^{-1} yr^{-1})	22.0	27.7
Dg	*Dactylis glomerata*	7.33	3.37	√p	- square root P rate	3.69	2.90
Fr	*Festuca rubra*	5.60	2.78	s	Annual S fertiliser rate (kg S ha^{-1} yr^{-1})	43.1	44.5
Ae	*Arrhenatherum elatius*	8.03	3.39	√s	- square root S rate	5.48	3.63
Pp	*Phleum pratense*	9.62	2.23	g1	Grazing in previous year (SU ha^{-1})	3.43	2.86
Hl	*Holcus lanatus*	10.08	1.87	g5	Cumulative previous 5yr grazing (SU ha^{-1})	14.8	12.8
Ac	*Agrostis capillaris*	8.57	2.70	x1	Interactions = t^2*√s	575	707
Ao	*Anthoxanthum odoratum*	9.87	2.67	x2	= t^2*√p	387	523
Bt	*Bromus tectorum*	9.68	4.90	x3	= t*g1	34.5	36.8
Fn	*Festuca novae-zelandiae*	8.77	2.87	x4	= t*g5	162	175
Hp	*Hieracium pilosella*	5.32	3.29	x5	= r*√s	-0.140	4.27
Pp	*Poa pratensis*	10.12	1.89	x6	= g5*√p	71.1	97.4
Pe	*Pyrranthera exigua*	10.51	1.66				

A case-based method was developed for a direct search for the closest matching observation in the reference data set using two different criteria for "closest". Its development is described further in the Results.

The other statistical methods evaluated were stepwise multiple regression using only the known variables; stepwise quadratic response-surface regression using the known variables and their implied quadratic and first order interactions; bivariate spline-regression; multiple imputations; transformation regression using 5-node splines with back transformation to original units; discriminant analysis, canonical discriminant analysis, correspondence analysis principal components (SAS, 1989, 1990); and closest matching observation in a table of a regular grid determined from one of the best fitting methods. These statistical methods differed in their assumptions on the nature of the variables, separation of variables into groups (independent or dependent), and where the experimental or random variability is assumed to reside.

For each method, each observation in the reference data set (n = 1418) was regarded in turn as missing in some variables, and having to be estimated from remaining observations. The mean correlation between observed and predicted values, after all iterations had been completed, was used as a measure of reliability of the method. This was done both for all remaining observations at each iteration, or for a reduced reference data set of a random 20% of the original.

3. Results

The description of the variables used, their means and standard deviation, are given in Table 1 together with their coding used in other tables. The first sections of the Results deal with various estimation methods used, while the later sections with their use in 'species from environment' or 'environment from species' analysis and general questions of sampling efficiency.

3.1 Case Based Data Search

The most direct way of using a reference data set is a search for the closest corresponding observation, and possibly interpolation with the second closest. It will be shown that this was among the best and most reliable prediction method.

The search for the closest correspondence between a query observation and a reference data set is based on the distance between the set of query independent variables and the corresponding set of independent variables in the reference data set and their associated values for the dependent variable(s). The distance can be defined as the Euclidian distance in multi-dimension space using either the difference between the query and reference value of each independent variable in their measured units (when variables are measured on the same scale as for the plant variables in the present study), or after standardisation of the independent variables using their respective standard deviations in the reference data set. Standard deviation adjustment is appropriate when the variables are measured in different units, and is probably also a better default option. It is helpful for subsequent calculations if the reference data set is sorted at that stage by the distance to the query values.

The next determination to make is whether the query is within or outside the ranges of the reference data. A criterion is that at least one observation in the reference data set must be within some pre-specified critical distance of the query. The critical distance would be specified in either the measured units, or a proportion of standard deviations of the independent variables, corresponding to that used for distance calculations. The critical distance is defined by the known variables.

When the query is outside the data field the best estimate for the dependent variable(s) from the reference data set is those from the observation(s) closest to the query in its distance for the combination of values of independent variables.

The combination of values of independent values in the query may correspond exactly with some observation(s) in the reference data set, as indicated by a zero distance and be regarded as a point estimate of the corresponding dependent variables.

Where the combination of query independent variables lies within the reference data field then an initial estimate of the dependent variables can be based on the observation(s) closest to the query. However this can be improved by consideration of the next closest observation(s) and possible interpolation between them.

To use such interpolation for improvement of estimates, the next closest observation(s) has to be on the "opposite side" of the query values to that of the closest observation(s). This requires reduction of the reference data set by removing observations that are on the "same side" of the query as the closest observation(s). There seems to be no simple precise algorithm for doing this in multi-dimensions. An approximation is to consider each independent variable in turn and remove observations which are not on the same side of the closest

observation as the query value. This method does delete some observations which may be "adjacent" to the closest and where some interpolation might have been possible, but retains all observations which are clearly "opposite" to the closest relative to the query. The next closest observation is then selected from the reduced data set, and interpolation of query dependent variables made from their respective values in the two reference observations in proportion to their distances from the query observation. Where there is no next closest point (as when the closest point is at the edge of the data field) then the best estimate remains the closest point. Even though only a few observations may be used to make these final estimates, it is not possible to reduce what may be a large reference data set, much before selection the second closest observation(s).

In all appraisals allowance has to be made for the possibility from either chance or by design that the query may be equidistant from a number of observations in the reference data set. Such coincidences may be relatively common, as with empirical data from factorial design experiments, or tables constructed from data, or where there are more combinations of different dependent variables in the reference data set than asked for in a query. For example in the present study a simple query in relation to a particular S fertiliser level would elicit all the associated observations of variation due to P fertiliser levels, stocking-rate, stocking methods and years of development. Where such coincidences occur then one can probably do no better than average the respective values of the independent and dependent variables. Where such independent variables are averaged for observations from the references data set it is necessary to recalculate the distances for the second closest observations.

Such coincident points allow some estimate of the reliability of the estimate of the dependent variables estimates. If there is only a single closest point then there can be no estimated of standard deviation.

The final estimates for each of the dependent variables can be coded for the various conditions of the closest matching observation(s) – as being either "point", "in", or "out" of the data field; whether the closest observation(s) represent single (sX) or multiple (mX) combinations of the independent variables; or single (sY) or multiple (mY) values of each dependent variable; and an estimate of their standard deviation (., 0, or an estimate).

It is suggested that the nature of the closest observation in the reference data set relative to the known variables in the query observation is useful information even if other methods are subsequently used to estimate the query values. The potential disadvantage of the data search approach is that most estimates will be determined solely by the two closest observations, with no indication of how representative those are of the local data field.

3.2 Missing Value Methods

Estimates using "missing value" techniques are present in many of the statistical methods. All missing value methods fit a general model to the known or reference observations of variables and use this to estimate the values for the query variables of an added observation(s) from the combined measured variable in the reference and added observations, and thus have the advantage, or disadvantage, of being determined by the whole data set and the assumed response function.

Multiple regression, giving the relationship between groups of independent and dependent variables, established from the reference data set, could be used to make a prediction from the known variables to the query variables in these situations. The present analysis used a standard set of predictor variables common to each group and also included comparison with two other methods: inclusion of all significant predictor variables independently for each response variable, and for transformation regression with splines. The resulting reliability, as given for percentage variation accounted for, increased in order of the three methods (for "environment from species" plant species only- mean 52%, 52% & 65% respectively; for "environment from species" plant and other non-dependent variables - 62%, 71% & 81%; and for "species from environment" - 52%, 61% & 68%).

A wider range of response fitting method was evaluated in the present context, and also repeated for each observation in turn as the missing values and using the remaining observations as the reference data set. They were done both for just a simple listing of the independent treatment variables and for the addition of a combination of time and created interaction variables.

A potential weakness of the multiple regression approach as a general search method is that it assumes a linear additive relationship for the effect of independent variables. That implies the need for some pre-analysis if specific polynomial and product variables are required to accommodate curvilinear or interaction effects. Also the fitted function is determined from the whole reference data set without reference to the query values. An alternative method evaluated was to limit the reference data set to the 20% of the observations closest to the known variables for the query and to limit the independent variables to those known, without inclusion of

curvilinear polynomial or interactions variables. This was done in the hope that by so limiting the reference data field, a linear multiple regression function would be sufficiently robust as a general search strategy, while still incorporating some smoothing of local reference data variability.

Another approach used to accommodate possible curvilinear or variable interactions was response surface regression with its automatic generation of quadratic and first order interaction terms for the known independent variables.

A more general approach for potential curvilinear relationships (but not variable interactions) is the use of spline regression which has the potential to more closely reflect the response in each section of the data field and limited effect of data in other sections of the data field. This option was evaluated in transformation regression with a 5-knot spline for the dependent variables and with back transformation to original scale units.

The above regression methods assume that the fitted function, once determined, is correct and only make one estimate of value for each of the query variables, so can give no real estimate of how reliable that estimate is. The multiple imputation method incorporates both the variation there may be in the fitted function and multiple estimations of the query value to give an estimate of its reliability.

The above statistical methods differentiate between independent and dependent variables and where the measurement- without-error and random variation lay. As indicated earlier this results in different assumptions about the same variable when used in different contexts. Hence there is attraction to the more even-handed treatment of principal component analysis which treats all variables as being in the same group, but the disadvantage of an assumption of multi-normal distribution.

The combined data set in the present study was relatively large (n = 1418) by comparison to many vegetation studies, and that may have biased the success of some methods, particularly the data search option, towards more favourable consideration than might be present in many data sets. The comparative analyses for all methods were repeated where the reference data set was randomly reduced to about 20% of the original to simulate smaller reference data sets.

While it was anticipated that in treating species rank abundance as continuous quantitative variables, and in their estimation by different method under different assumption, would yield continuous rather than discrete numbers, that the methods may be more robust in maintaining the relative rank order of species, as would be used in defining different nominal vegetation types. The different estimation methods were also compared in this regard.

3.3 Comparison of 'Environment From Species' Estimates

The prediction of response variables was highest for years of development (t) and irrigation (i), moderate for grazing capacity (g1, g5) and fertiliser level (p, s) (Table 2). The lowest predictability was for relative stocking rate (r) and grazing method (m) - which are probably the two most desirable for prediction for management in a monitoring program in the sample context. Predictability was increased slightly as the number of plant species considered was increased from 10 to 30.

Direct search of the reference data set for the closest corresponding observation in species composition was generally one of the three best prediction methods for environmental variables. In this, determination of closeness distance after normalisation of the independent variables by their standard deviation was slightly better than linear distances for species in their original scale units (which were common for all species) were generally slightly better than when only the 10 most common species were used as independent variables and the reverse when 30 species were considered. However, the differences were small, and in combination with the next section the implication would be that it is probably better and more consistent to automatically normalise variables by their standard deviation in any general purpose search method.

The second most consistent prediction method was multiple regression using only the specified independent species variables and confining the reference data set to the 20% of observations closest to the query observation. Inclusion of all observations decreased the predictability probably because interaction and curvilinear effects within the whole data not being adequately described by a linear additive model.

When only 10 species were used, response surface regression also gave good predictability and better than multiple regression with all observations, as would be expected because of the inclusion of generated quadratic and first-order interaction terms. However, when the number of species was increased to 30, the implied generation of additional variables in response surface regression, overwhelmed the computing process, and gave negligible prediction.

Table 2. Estimation of 'environment from species' composition. Percentage variation accounted for ($100*R^2$) in comparison of missing-value multi-variate methods in estimating query values (for each observation in turn) from base data set (remaining observations) (n= 1418). Variable codings as in Table 1. Using either 10 species (Ac, Ae, Dg, Fn, Fr, Lp, Hp, Ta, Th, and Tr) or 30 species

	10 species								30 species							
	t	p	s	i	m	r	g1	g2	t	p	s	I	m	r	g1	g2
(a) All data																
Data search – linear	79	51	49	84	20	24	57	74	88	63	55	94	26	31	61	81
- std	87	57	54	85	24	28	62	77	86	57	49	92	26	23	59	77
Simple regression	73	51	43	59	22	19	57	70	86	57	52	77	26	23	67	80
Reg. –closest 20%	82	61	50	78	33	32	68	81	91	62	57	87	36	33	73	85
Response surface	83	60	50	77	31	30	68	81								
Mult. imputation	68	44	32	52	12	11	48	64	86	57	49	92	26	23	59	79
Trans.regression	62	8	16	4	1	1	59	74	83	40	43	55	1	0	67	81
Principal component	0	0	0	2	1	2	0	0	54	29	42	48	1	0	57	62
Smoothed table	73	52	42	74	18	23	58	67								
(b) Random 20%																
Data search – std	70	43	35	73	13	13	51	68	76	47	44	83	13	16	50	73
Simple regression	76	58	54	81	17	15	63	77	78	49	44	86	14	18	51	71
Reg. –closest 20%	69	65	61	55	29	14	70	74	84	46	48	74	18	17	64	77
Response surface	81	53	50	79	24	20	67	8								
Mult. imputation	68	46	43	57	12	10	52	67	67	35	20	69	8	1	42	72
Trans. regression	61	0	5	4	1	1	63	69	79	1	0	6	0	0	6	27

Multiple imputations, using five estimates per missing values, gave good predictability when 30 species were considered and moderate predictability with 10 species, and had an advantage that it also generated an estimated standard error for each of the predicted values. Transformation regression gave moderate predictability of the more predictable variables, but poor predictability of other variables.

Predication, using principal component missing value methods, gave the lowest predictability of the analysis methods.

With a large data set, with observations as successive yearly measurements on the same plots of slowly changing vegetation, it is probable that in making each observation missing in turn, that direct data search of the remaining data set should locate similar observation close to the query observation. However, when the remaining data set was randomly reduced to 20% of the original to simulate a smaller reference data set, the direct data search still gave predictability similar to other methods. With a reduced reference data set the response surface regression gave the best predictability when only 10 species were included, but failed for a large number of species. The predictability by the simple regression method was less affected by the reduction in the reference data set than the further reduction implied by reducing it to the closest 20% (i.e. 4% of reference data set).

3.4 Comparison of 'Species From Environment' Estimates

The prediction of species rank abundance was moderate for the 10 species considered (Table 3) with predictability tending to increase for species with higher standard deviation in rank values in the base data set ($P \leq 0.05$). Predictability was highest for *Lupinus polyphyllus* and *Hieracium pilosella*, the two species most universally present in the reference data set.

The three better methods of species prediction were direct data search, simple regression on the closest 20% observations, and response surface regression.

In the direct data search, as would be anticipated, the normalisation of distances to independent environmental variables by their standard deviation gave slightly better fit that leaving variables in their original different scale units. The reduction of the reference data set to the 20% closest observations gave clearly better predictability than the use of whole data set in multiple regression. Similar predictability was given from 7 environmental variables by response surface regression with its implied generation of quadratic and first order interaction variables of the independent variables. The calculations were only occasionally overwhelmed when the number of environmental variables was increased to 15, with the implied duplication of some of the interaction variables.

Table 3. Estimation of 'species ranking from environment'. Percentage variation accounted for ($100*R^2$) in comparison of missing-value multi-variate methods in estimating query values (for each observation in turn) from base data set (remaining observations) (n= 1418). Variable codings as in Table 1. Analysis using either 7 simple treatment variables (p, s, i, m, r, g1, and g5) or all 15 combined time (t) and time by treatment interaction variables

Method	Predicted species ranking																			
	7 environmental variables										15 time and environmental variables									
	Lp	Ta	Tr	Th	Dg	Fr	Ae	Ac	Fn	Hp	Lp	Ta	Tr	Th	Dg	Fr	Ae	Ac	Fn	Hp
(a) All data																				
Data search -lin	65	33	35	38	42	42	52	31	46	64	66	54	48	60	51	52	60	36	47	66
- std	72	38	38	43	49	45	54	31	46	66	75	65	50	69	62	59	63	37	58	76
Simple regres.	58	32	35	31	34	25	38	23	19	62	60	52	44	66	53	43	50	24	37	65
Reg.–close 20%	67	45	42	49	46	44	57	34	42	70	69	67	53	73	65	61	65	37	54	74
Response surf.	65	45	40	45	46	39	59	29	42	69	71	70	54	74	66	62	70	36	56	75
Mult. Imput.	51	21	26	24	25	16	29	16	14	54	52	42	35	60	46	33	41	17	26	59
Trans. Regres.	56	36	31	25	31	22	40	17	22	64	57	52	40	69	42	37	45	18	64	31
Principal comp.	0	3	1	2	0	2	1	4	0	5	0	0	0	0	1	0	1	0	1	0
Smoothed table	75	65	50	72	65	61	63	35	56	72										
(b) Random 20%																				
Data search –lin	55	30	24	32	32	37	41	27	32	55	51	28	25	38	33	32	39	25	26	50
– std	61	27	27	28	35	31	44	22	32	60	64	52	33	60	49	47	56	20	46	66
Simple regres.	56	31	34	30	33	24	36	22	19	60	57	50	40	64	51	41	49	20	35	63
Reg.–close 20%	56	31	34	29	33	24	36	21	17	60	53	47	36	64	48	44	49	16	34	62
Response surf.	61	41	37	41	39	33	54	21	37	66	53	53	29	53	49	43	56	14	38	58
Mult. Imput.	48	23	27	21	26	16	29	13	9	57	46	39	27	57	40	35	43	12	25	54
Trans. Regres.	41	29	0	23	25	16	25	0	0	62	34	47	0	65	32	35	44	0	0	63

While multiple imputation and transformation regression gave moderate predictability, being good for the more predictable species, were inconsistent, with predictability decreasing for the less predictable species. Principal component analysis was the least effective method.

When the reference data set was randomly reduced to 20% of the original, to simulate a reduced reference data set, the response surface regression gave a slightly better predictability than direct data search, regression, or regression of closest 20% observations, when only 7 environmental variables were considered. There was little difference between these four methods when the number of environmental variables was increased to 15.

In the "species from environment" analysis the principal components and transformation regression method failed in many instances by producing "outrageously large missing value estimates" for some combinations of variables. This had to be anticipated with the algorithm used in the method (SAS, 1989 – p.1311) and related to the gradients in the multi-dimensional space near the estimation points.

The problem probably arises from the nature of the species variables. The observations within an empirical data set will not be equally dispersed among possible combinations of values. Indeed, within the species data of vegetation composition components, there will be a strong tendency for the data combinations to be confined to a multi-dimensional response surface or narrow layer data cloud. This is because of the nature of many vegetation sampling methods are compositional. The present study used ranking of species according to their contribution to herbage bulk. In this only one species could be 1st ranked in a particular sample, only one could be 2nd ranked, etc., and it was not possible for several species to hold the same ranking within a particular observation (with the exception of common ranking for rare or absent species). Such ordinal relationships, or requiring components to sum to a constant value, occurs in other sampling methods e.g. species ground cover or proportional contribution needing to sum to 100%. Also as species in vegetations tend to segregate into proportions showing a log/linear relationships between relative abundance and rank the dispersion of value combinations may be less than anticipated. While such narrowing of dispersion of variable combination should not be a limitation in any query based on a "real" single sample, it can cause problems in queries based on the

means of several samples, or where speculative combinations are being investigated, where there may be a higher frequency than expected of "the query not being within the data field".

While it easy to detect obvious miss-predictions like negative values, or values well outside a variables likely range, the presence of some in any prediction method undermine confidence in some of the results. One of the virtues of the direct data search method is that predicted values can only be within the range of those of the reference data set.

3.5 Comparison of 'Vegetation From Environment' Estimates

The vegetation types used in this comparison were based on 10 species, based on their rank order either that in the original data, or predicted by one of the above methods, and with vegetation type defined by their order in either 1, 2, ---- 5 species vegetation types, and with their reliability estimate being their exact correspondence between the original and predicted types (Table 4).

There was reasonable predictability of vegetation type if it was only defined by the most dominant species, but rapidly decreasing predictability as vegetation types were subdivided by consideration of ranking of further species.

The highest predictability for vegetation type was from direct data search for most closely corresponding species composition in the reference data set. The next best methods were response surface regression, or simple regression on the closest 20% of observations. There was only slightly less predictability, and the same ordering of the methods, when the reference data set was reduced to 20%.

3.6 Smoothed Table or Response Matrix

Estimation or prediction by the previous described methods involved re-analysis from the reference data set where variables were at their measured and generally irregular levels, and still including any experimental error. For a hard copy 'table look-up' approach, or as a summarized response matrix for inclusion in a field computer or expert system it may be desirable to have the variables and responses on some regular grid. The process of translation from the raw data dispersion to a more regular grid could also include the element of data smoothing to remove some components of experimental error. A regular common grid would also aid in combining data from different sources. Estimation from such a smooth response function would be by data search or one of the other methods.

The cautions in producing such regular grid tables is that they can carry with them an implication of separation into independent and dependent variables, that they will depend on the interpolation methods used, and that it is too easy to produce a large number of variable combinations outside the raw data range. A conservative approach would suggest that a table should not contain more entries and combinations than reasonably present in the raw data.

The method used was to use a regular grid as the known variables, and then estimate the query variables for each of them from the original reference data set.

Based on the above results and considerations, different methods of data smoothing and table formation were tried. In the raw data the S and P fertiliser rates, while design variables and treated as continuous variables, because of variation in the proximate analysis of the actual fertilisers used, had actual irregular values. For table formation a grid for 0, 5, 10, 20, 50 and 100 kg element ha^{-1} yr^{-1} was established for both S and P fertiliser. The years from start of development had fixed regular discrete coding, which for the table formation were determined for 2 yearly intervals to 18 years. In the raw data set the treatment variables of stocking-rate, grazing-method, and irrigation in the Graze/fert trial, had fixed regular discrete coding but treated as continuous variables (-1, 0, 1; 0, 1; and 0, 1 respectively) and for the table were determined for the bounding S and P fertiliser rates of the corresponding treatments.

In the raw data the species rank data were treated as continuous variables so as means of sub-plot measurements had irregular continuous values, and after interpolation for the table were rounded back to discrete rank values of 1---10. Sheep carrying-capacity is likely to always be a response variable with continuous irregular values appropriate and similarly treated in a table. Species rank abundance and sheep carrying capacity for this table grid (n = 1320) were determined from the full data set using the estimation method of either direct data search with interpolations, simple regression, simple regression of closest 20%, or transformation regression. An assessment of use of each of the tables to back-estimate the values in the reference data set is included in the comparison of prediction methods (Tables 2-4).

Table 4. Estimation of vegetation type from environment. Percentage correct predictions in comparison of missing-value multi-variate methods in estimating query values (for each observation in turn) from base data set (remaining observations) (n= 1418). Variable codings as in Table 1. Using vegetation based on rank order of 10 species (Ac, Ae, Dg, Fn, Fr, Lp, Hp, Ta, Th, and Tr) from either 7 simple treatment variables (p, s, i, m, r, g1, and g5) or 15 combined time (t) and time by treatment interaction variables. Confidence interval c ±3.1% at 50%

Method	Predicted vegetation type									
	Veg-1		Veg-2		Veg-3		Veg-4		Veg-5	
	7	All	7	All	7	All	7	All	7	All
(a) All data										
Data search – linear	62	69	33	38	16	20	8	11	4	4
- std	62	72	32	40	16	21	8	11	4	4
Simple regression	56	63	19	27	8	14	2	5	1	2
Reg. –closest 20%	62	70	28	38	15	20	6	12	2	4
Response surface	63	68	28	38	14	18	6	9	2	4
Mult. imputation	50	57	16	23	6	11	2	4	0	1
Trans. regression	44	53	16	23	4	10	1	4	0	1
Smoothed table		73		43		26		16		10
(b) Random 20%										
Data search – std	59	66	27	40	13	25	5	19	2	15
Simple regression	56	52	19	23	7	11	2	4	1	2
Reg. –closest 20%	57	50	21	22	9	11	3	4	1	1
Response surface	62	49	26	26	13	13	5	6	2	2
Mult. imputation	51	44	17	18	6	8	1	2	0	1
Trans. regression	37	36	12	15	4	6	1	1	0	0

The method of table formation with a regular grid which gave the highest subsequent predictability was from direct data search of the original data set, followed by simple regression of the closest 20%, then response surface regression, then transformation regression, and in each case direct data search and interpolation to make back estimates from the table.

The predictabilities of environment factors from species using the smoothed table were slightly lower than for the best three of the previous comparisons. For estimation of species ranking and vegetation from environment the estimated predictabilities are higher than for the other methods. However there was not a strict comparison with the previous methods. In the table formation all observations were used in estimating values at each grid point, so each particular observation had a contribution to estimation at a particular point, hence contributing to its own estimation in the back estimations, whereas in the other comparisons each observation in turn was specifically excluded from its own estimation. It is probably sufficient to conclude that the table so formed with a more regular grid of values retain almost all the distribution characteristics of the raw irregular data set.

3.7 Vegetation Monitoring Efficiency

The above comparison of estimation methods on a sample data set used the base method of rank order of a species as a measure of their relative abundance and using either 10 or 30 species in determining 'environment from species' relationships. The present section examines the effect of the number of species that need to be assessed in monitoring and alternative ways in which their base rank order data might be expressed in estimating management or environmental factors. The two dependent variables considered were P fertiliser rate, and the relative stocking rate, the two variables which in the analysis above had the highest and lowest predictability respectively (Table 2, Figure 1).

For all forms of the species data the predictability of the regression equations increased greatly when 2 or 3 species were ranked relative to that if only the most dominant species was used. In the analyses the species type 'other' was excluded, so as not to cause singularity in matrices. Presence /absence of species (1/0) data had the lowest resultant predictability of the two management variables and predictability decreased if further species were considered. The transformation of ranks to percentage contributions gave generally greater predictability than the ranks on an ordinal scale, though plateauing in predictability at about 5 species being ranked. The data used in that analysis considered all species in the data, not just the fourteen more frequent, so the percentage

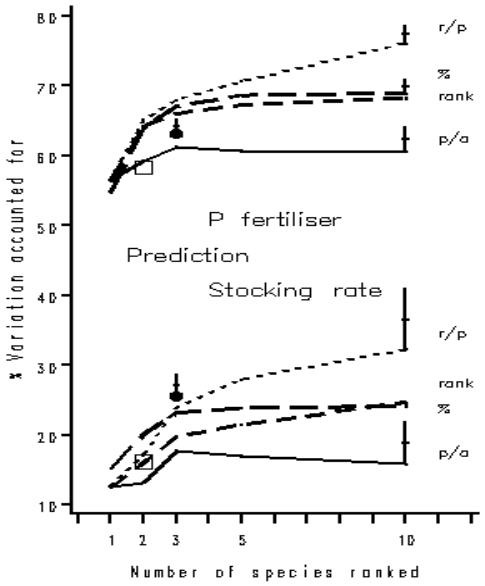

Figure 1. Percentage variation accounted for in predicting P fertiliser rate and stocking rate treatment in Graze/fert trial from alternative representations of plant parameters for 14 species. Rank = rank in either 1st, 1st-2nd, 1st–3rd, 1st–5th, or 1st–10th species group; p/a = presence /absence of species in the same groups; r/p = presence /absence in each rank position for same groups; % = percentage contribution of ranked species in same groups; ● = 44 vegetation types; and □ = reduced 14 vegetation types. Bars show increase when other non-dependent treatment variables included

variation accounted for is higher than reported in the table. The highest predictability of the two management variables came from expressing the vegetation data as presence /absence of each species in each rank position, with predictability increasing with number of species considered. In the analysis of the latter method, the solution did not take account that rank classes were mutually exclusive e.g. a species could not appear as both 2nd and 3rd ranked in a sample.

The predictability of the equation for the relative stocking rate increased markedly if other environmental variables were included. However, there was little effect of additional information on the predictability of P fertiliser rate.

Expressing the species data as 14 2-species vegetation types or as 44 3-species vegetation types as in (Scott, 2005 -Table 1), gave lower predictability of P fertiliser rate than use of species data. While the predictability of relative stocking rate with the 2-species vegetation type was lower than retention as individual species data, and that for the 3-species vegetation type was higher.

4. Discussion

The objective of the study was to investigate methods of how to carry forward archived quantitative results from vegetation and environment monitoring surveys to their utilization in predicting trends from further partial measurements, either in the same or other situations. That involved a comparison of the predictive power of different statistical methods, and the extent to which the different aspects in the original measurements contributed to the predictive power.

The study used the experimental methods and species of controlled agricultural studies, but the analytical methods more usually associated with ecological studies. The features of the base data were a wide range of controlled soil fertility and grazing management variables, the wide range of predominately agricultural pasture species with continuous annual records over two decades and with the formal structure of the trial design set aside for the present study.

However, it should be emphasized that the preference is that the study should be viewed in the context of the general possible techniques for obtaining, analysing and utilizing of archived monitoring data, rather than the specific biological and ecological features of the sample data used.

The comparison used the "missing value" techniques present in many of the statistical methods. Researchers are often ambivalent about using such methods as they seem to imply some deficiency in the base data (from the development of the method in estimating missing observations in otherwise factorial experimental designs). But missing value methods are prediction methods. If a base data set is complete, (e.g. there were measured values for all designed treatment combinations and responses), and one adds "new" observations relating to known measured values for some variables and unknown desired query values of others, then missing value methods are predictive methods for that situation and better described as "prediction from partial additional data".

The present study raises a number of implications to the general question of vegetation monitoring for predicting management or environmental trends.

The first is the degree of detail and accuracy needed in the initial samples. These were that most of the predicative information was present in the first 3-5 most abundant species in a sample; that ordinal ranking is only slightly less useful in predictability than quantitative measures; and that in multi-species situations there is no advantage in amalgamating the species data into vegetation types. Ordinal ranking methods are also generally much easier and quicker to obtain, and consequent increase in the frequency of observations and the area or diversity that can be monitored. Also the accuracy of predictions in many methods depend more on the range of conditions sampled and the number of samples rather than the inherent accuracy of individual samples.

The general predictive superiority of the case-based direct data search for the closest corresponding match implies the desirability of archiving the base data in their simplest form.

The study also illustrated the different types of question that may be asked of the same base monitoring data set. Comparisons were given for both estimation of pasture species response given the treatment and environmental conditions, and of the reverse, of estimation of treatment or environmental conditions given the species response or vegetation type. Either may be required in different contexts.

The interchange between being independent and dependent variables results in associated changes in statistical assumptions within the analysis method used and there seems too little guidance in the general literature of their implication for prediction. This also relates to the change of the assumption of measurement-without-error of designated independent variables, when in practice probably all variables have a measurement error.

The comparison of methods was made on a single sample data set from a single site. The issue that needs to be investigated more fully is how, in archiving, to reference such monitoring data sets with adequate associated covariates to define the environmental domain to which they may apply or extrapolated too, thus enabling amalgamation with other data sets. This may be to the likes of the categorical classes of soil type or land environmental classification, which in turn would need access to the quantification and continuum of the defining ecological variables involved. The difficulty would be for modified systems, such as rangeland and agricultural systems where there are changed or have other inputs. We have attempted a semi-quantitative approach to that in Scott et al. (1995).

A particular difficulty will be how too usefully reference 'time' in such archived data sets. In one sense date, year, years-of-development etc. are irrelevant. The need is for some concept time parameter like 'ecological stage'.

Acknowledgements

The empirical field work was carried out while the author was with AgResearch, Lincoln, New Zealand, and supporting technical assistance is gratefully acknowledged.

References

Mannetje, L., & Haydock, K. P. (1963). The dry-weight-rank method for the botanical analysis of pastures. *Journal of the British Grassland Society, 18*, 268-75. http://dx.doi.org/10.1111/j.1365-2494.1963.tb00362.x

Preston, F. W. (1948). The commonness and rarity of species. *Ecology, 29*, 254-283. http://dx.doi.org/10.2307/1930989

SAS. (1989). SAS/STAT Users Guide, Version 6. Cary, NC, SAS Institute Inc.

SAS. (1990). SAS/GRAPH Software: Reference, Version 6, First Edition. Cary, NC, SAS Institute Inc.

Savory, A., & Butterfield, J. (1999). *Holistic management - a new framework for decision making* (p. 616). Washington, D.C., Island Press

Scott, D. (1989). Description of vegetation interactions using visual ranking of species. *New Zealand Journal of Ecology, 12*, 77-88.

Scott, D. (2001). Sustainability of New Zealand high-country pastures under contrasting development inputs 7. Environmental gradients, plant species selection, and diversity. *New Zealand Journal of Agricultural Research, 44*, 59-90. http://dx.doi.org/10.1080/00288233.2001.9513463

Scott, D. (2005). Sustainability of New Zealand high-country pastures under contrasting development inputs 9. Vegetation interaction. *New Zealand Journal of Agricultural Research, 50*, 393-406. http://dx.doi.org/10.1080/00288230709510307

Scott, D., & Covacevich N. (1987). Effects of fertiliser and grazing in a pasture species mixture in high country. *Proceedings of the New Zealand Grassland Association, 48*, 93-98.

Scott, D., Maunsell, L. A., Keoghan, J. M., Allan, B. E., Lowther, W. L., & Cosssens, G. G. (1995). A guide to pastures and pasture species for the New Zealand high country. *Grassland Research and Practice Series, 4*, 1-41.