INTERNATIONAL GEOLOGICAL CONGRESS

24th SESSION

(396) A COMPARATIVE STUDY OF MODELS TO PREDICT HYDROGEOLOGICAL RESPONSE

by

Teja Singh

ABSTRACT

Infiltration is an important hydrogeologic characteristic of a basin because of its influence on interflow, groundwater recharge and overland flow. The rate at which water enters the soil surface is the integrated net result of the many edaphic and vegetative influences that are often superimposed on the inherent geology of a watershed.

The present study was conducted in Streeter Basin, an experimental watershed situated in the southern foothills of Alberta, Canada. The bedrock, of Porcupine Hills sandstone, is overlain with silty to sandy till of varying thickness. Infiltration runs of 3-hour duration were obtained by using constant head double ring infiltrometers on 48 plots. A total of 13 edaphic and vegetative variables known to have influence on water intake rates of the soil surface were measured concomitantly at each plot. Such determinations were also repeated on another 80 plots in similar vegetation types in the vicinity to provide test of the predictor quality of the models. Stepwise regression and varimax rotation of the factor weight matrix were used to determine relative importance of the variables. Multiple linear regression models and principal components regression models were derived for extrapolation purposes.

Principal components prediction was better than or at least as good as the multiple regression prediction. Allied with the decision on the total number of variables is the equally important consideration regarding the total number of components to be included in the model. Optimum choice can make the difference in the superiority of one model compared with the other.

INTRODUCTION

There has been recently an increasing awareness that most hydrogeologic processes are multivariate in nature in that several variables operate concurrently in such systems. Simple deterministic "cause-and-effect" relationships are easy to discern when hydrogeologic problems are isolated and considered in relatively small segments. However as the number of variables increases considerably, numerous interactions and partial dependencies complicate the picture and blur the sharpness of the classical cause-and-effect approach (Krumbein, 1969), making it exceedingly difficult to determine precisely what actually "drives" the system.

One possible solution to such complex situations, of course, is to select a minimum number of the most important operative variables and to confine the study to a few interactions and combinations. Further, as in a greenhouse study, some of the variables can be kept constant and others allowed to vary according to the range of observations needed on each input variable. Quite often such mechanisms do not have counterparts in nature where most of the environmental factors change frequently and concurrently.

Multivariate data analysis provides a possible alternate analytical tool to deal with complex ecosystems. The variables can easily be considered concurrently and can be allowed to be numerous, subject to the data handling capacity of the modern age computers. By using information on a large number of variables it is possible

to avoid personal bias and preference for a few selected characteristics, thereby obviating the need to attach special importance or weight to only a few chosen variables (Gittins, 1965).

The purpose of this study is to compare the multivariate models in predicting hydrogeological response in areas similar to those from which they are derived. The study was conducted in Streeter Basin, a gauged watershed established for studying the hydrology of aspen forests and associated grasslands of southern Alberta (Jeffrey,1965).

METHODS

Constant head, double ring infiltrometers were used to obtain direct measurements of water intake at each site during a 3-hour run. A number of concomitant measurements were also made on edaphic and vegetative variables as listed in Appendix 1. Sixteen sites were taken in each of the eight vegetation types (Appendix 2) present on the watershed. As the entrance of water at the soil surface is primarily determined by the conditions existing near the surface layer most affected by management practices, the edaphic variables included in the study were those measured within the top 3 inches of the mineral soil.

Vegetation units 3, 4, and 8, each with 16 samples, were used in the derivation of models. Stepwise regression and varimax rotation procedures were used to assess the relative importance of variables included in the study. Multiple linear regression models were obtained

for each of these vegetation types and also similarly derived for a combination of the three sub-sets when grouped into a single set representing the aspen parkland vegetation in general. In the latter case dummy variables GR, SH, FOR were included to indicate the presence or absence of the vegetation type represented by each.

Prediction capability of the derived models was tested when applied to the experimental data obtained for the remaining 5 vegetation units (1, 2, 5, 6 and 7). Prediction errors and related statistics were calculated to determine the extrapolation capacity of the models in each case.

RESULTS AND DISCUSSION

Table 1 provides a summary of data from the three vegetation units (3, 4, and 8) each representing a main vegetation type of the aspen parkland on the watershed. Multiple correlation coefficient (R) and standard error of estimate for the models derived from these units are listed in Table 2. R ranged from 0.96 to 0.99 for the stratified populations, and 0.80 to 0.81 for the combined units using, in addition, the 3 dummy variables. Without the dummies, the R is 0.73 for aspen parkland vegetation when so grouped. Errors of estimate were lower for stratified than combined populations. High R and low errors of estimate for the stratified populations actually did suggest that these would be good models for extrapolation, although the regression coefficients are likely to be relatively unstable because of the few

degrees of freedom left for the error term.

The stepwise multiple regression program (Kozak and Smith, 1965) showed the relative importance of variables as listed by their rank in Table 3. Obviously the rank is influenced by the vegetation unit to which stepwise regression is applied because in each of these types the variables are ranked differently. This is mainly due to the multicollinearities existing among the independent variables (Cavadias, 1964) as evident from some of the eigenvalues of the unrotated factor weight matrix (Table 4) which are zero or estimates of zero (Krumbein and Graybill, 1965). Components 12 to 16 taken together, for example, contain less than 1% of the total information content of all the variables.

Tables 5, 6 and 7 present the results obtained, using the models derived from vegetation units 3, 4 and 8, when extrapolated to vegetation units 1, 2, 5, 6 and 7. Although they differ considerably in terms of predictions within and among the five vegetation units, a comparison can be made of the mean error terms (i.e. average prediction error per observation unit for a total of 80 such units):

Model	Mean prediction error	
	Absolute	Actual
A. Stratified multiple regression	30.5	-16.5
B. Combined multiple regression (including 3 dummy variables)	9.2	+ 4.0
C. Combined multiple regression on principal components (including 3 dummy variables)	7.7	- 2.9

Both actual and absolute error accumulations are lower in the principal components regression model than in the multiple regression models. As prediction error is the difference between the actual and the predicted value $(Y-\hat{Y})$, the positive errors represent an underestimate and the negative errors an overestimate. The difference between the average predicted value and the actual value is a measure of bias and the results therefore show that the model with least bias in the present case is that incorporating regression on principal components.

Varimax rotation of the principal component (factor weight) matrix showed that three variables, namely SAND, SILT, and CLAY, can be omitted without any adverse effect on the predictive quality of the models. The mean prediction error when this was done amounted to:

Model	Mean prediction error	
	Absolute	Actual
A. Stratified multiple regression	16.1	- 4.5
B. Combined multiple regression (including 3 dummy variables)	7.7	- 2.7
C. Combined multiple regression on principal components (including 3 dummy variables)	7.7	- 2.7

Although individual predictions were not identical, the models B and C gave equal mean prediction errors when rounded and these were superior in prediction to models containing all the variables.

Using a different total number of components obviously gives different estimates of the mean prediction error for the same number of variables in the model. Prediction of the principal component model was

inferior when the number of components was dropped from 12 to 10 in case of the first category of models (Table 7) where all variables had been included in the model building process. Evidently much depends on the choice of the total number of components to be incorporated in the prediction equation.

Thus, allied with the decision as to the number of variables to be used in the predictor model is the equally important consideration regarding the choice of the total number of components to be retained in the model. Leaving more components out means loss in total information content and consequently a reduction in extrapolative power to some extent, and including more than the optimum necessary has a deleterious rather than a useful effect; increased use of multivariate techniques in future will undoubtedly lead to development of more exact and rigorous criteria to decide this critical question. Wallis (1968), for example, has suggested that most experimental hydrologic data have sufficient multi-collinearity for the 0.995 explained information content (variance) to be effective. This appears to hold good in the study reported here. Including less than the optimum number of components would generally make the principal component prediction inferior to that of the multiple regression.

A wide range of models is currently being investigated and are expected to shed further light on improved criteria to do an effective prediction job while extrapolating.

ACKNOWLEDGEMENTS

The assistance of Messrs. T.A. Thompson and W.B. Chow of the Canadian Forestry Service in data collection and computer analysis is gratefully acknowledged. My sincere thanks also to Mr. Henry W. Anderson of the Pacific Southwest Forest and Range Experiment Station, Berkeley, California, for valuable advice in the initial stages of data analysis and interpretation.

REFERENCES

Cavadias, G.S. 1964. Methods of analysis and interpretation, in Research Watersheds, Hydrology Sympos., 4th, Subcomm. on Hydrology, Assoc. Comm. on Geodesy and Geophys., Nat. Res. Counc. Canada.

Gittins, R. 1965. Multivariate approaches to a limestone community. I. A stand ordination. Jour. Ecol., 53: 385-401.

Jeffrey, W.W. 1965. Experimental watersheds in the Rocky Mountains Alberta, Canada. International Association of Scientific Hydrology Pub. 66: 502-521.

Kozak, A. and J.H.D. Smith. 1965. A comprehensive and flexible multiple regression program for electronic computing. For. Chron., 41: 438-443.

Krumbein, W.C. 1969. Deterministic and probalistic models in geology, in Models of Geologic Processes, American Geophysical Institute, Washington, D.C.

Krumbein, W.C. and F.A. Graybill. 1965. An introduction to statistical models in geology. McGraw-Hill Book Co., New York.

Wallis, J.R. 1968. Factor analysis in hydrology—an agnostic view. Water Resources Res., 4: 521-527.

APPENDIX 1

List of Variables[1] included in the Analysis

Symbol	Code	Variable
a		Constant in the prediction model
X1	ST	Soil temperature (°C)
X2	AM	Antecedent moisture (% by weight)
X3	GL	Ground litter (fresh decomposed material on grass and forest lands, expressed as percentage of ground surface)
X4	BA	Basal area of grasses and forbs expressed as percent proportion of ground surface
X5	CC	Canopy cover (percent proportion of ground surface covered by the vertical projection of live aerial parts) of grasses and forbs
X6	OM	Organic matter (%) in soil
X7	WHC	Water holding capacity: moisture content (%) of an undisturbed and saturated soil after free drainage has practically ceased
X8	WP	Wilting point: moisture content (%) of a soil sample after reaching equilibrium with an applied pressure of 15 atmospheres
X9	BD	Bulk density
X10	Sand	Sand % (0.05-2.0 mm)
X11	Silt	Silt % (0.002-0.05 mm)
X12	Clay	Clay % (less than 0.002 mm)
X13	PORS	Porosity (total pore volume): percent by volume of total pore space of a soil sample, calculated from bulk density and specific gravity: $PORS = \frac{\text{Specific gravity - BD}}{\text{Specific gravity}} \times 100$

[1] All edaphic variables were determined from sampling within the top 3 inches of mineral soil

Symbol	Code	Variable
X14	GR	A dummy variable (0,1) indicating the absence or presence of grassland vegetation
X15	SH	A dummy variable (0,1) indicating the absence or presence of shrub vegetation
X16	FOR	A dummy variable (0,1) indicating the absence or presence of forest vegetation
Y	INF13	Total infiltration (in.) during the 3-hr run

APPENDIX 2

List of Vegetation Units

Code	Symbol	Vegetation Unit
1	Phpr	Phleum pratense, a grassland sub-type confined to valley bottoms
2	W	A grassland sub-type occurring on slopes and having a prominent forb ("weeds") component
3	Dapa-Feid	A grassland sub-type confined primarily to ridge tops and dominated by Danthnia payrii and Festuca idahoensis
4	Sal-Beoc	A shrub type dominated by Salix spp. and Betula occidentalis
5	Potrich	A forest sub-type dominated by black poplar (Populus trichocarpa)
6	Potr-Ros	A forest sub-type dominated by aspen (Populus tremuloides) and an understorey vegetation consisting primarily of rose (Rosa acicularis and Rosa woodsii)
7	Potr-Caca	A forest sub-type dominated by aspen and an understorey primarily of reed grass (Calamagrastis canadensis)
8	Potr-Asco-Epan	A forest sub-type dominated by aspen and having an understorey consisting primarily of showy aster (Aster conspicuus) and fireweed (Epilobium augustifolium)

Table 1. Summary of data from which models derived

Variable	Minimum			Maximum			Mean			Standard deviation		
	Grassland	Forest	Shrub	Grassland	Forest	Shrub	Grassland	Forest	Shrub	Grassland	Forest	Shrub
ST	4.0	0.9	3.5	17.5	14.4	19.0	11.9	10.8	9.1	4.8	2.8	5.4
AM	19.5	0.6	3.9	100	83.1	88.2	52.2	58.8	50.2	24.2	13.1	20.2
GL	62.5	1.0	1.4	92.5	98.5	95.0	84.1	96.4	90.1	10.2	1.4	4.2
BA	5.0	1.1	2.3	20.0	5.0	22.5	9.3	3.2	9.6	3.7	1.1	4.2
CC	30.0	1.2	1.1	90.0	8.8	90.0	56.3	43.7	63.3	18.8	23.1	21.3
OM	9.1	2.2	2.1	15.4	15.7	15.2	13.2	12.2	10.8	1.7	2.1	2.5
WHC	46.7	2.3	1.9	106.2	208.0	235.6	70.5	81.6	87.4	17.5	48.4	63.6
WP	21.5	2.3	0.4	64.0	86.5	95.0	41.8	45.2	42.7	10.4	15.5	21.6
BD	0.3	0.2	0.3	1.0	1.0	1.0	0.7	0.6	0.7	0.2	0.2	0.2
SAND	35.0	2.1	0.3	71.0	60.0	62.0	51.3	47.0	44.2	10.0	6.3	10.9
SILT	15.0	1.9	0.5	52.0	54.0	64.0	37.6	40.0	41.5	10.0	7.3	11.7
CLAY	8.0	1.5	0.3	18.0	19.0	24.0	11.2	13.0	14.3	2.5	3.0	4.6
PORS	56.0	1.2	0.3	86.0	89.0	86.0	68.1	73.1	69.5	7.5	7.2	9.1
INF13	1.7	1.7	1.7	10.8	38.6	20.5	5.1	18.2	12.0	2.7	11.4	4.1

Table 2. Multiple correlation coefficient (R) and error of estimate of linear regression models for the populations from which derived

Population	Linear regression model consisting of	Total Samples	Mean (in.)	R	Standard error of estimate (SEE)	SEE as percentage of mean
1. Grassland	13 variables	16	5.1	0.99	1.3	25.1
2. Forest	13 variables	16	18.2	0.98	6.4	35.2
3. Shrub	13 variables	16	12.0	0.96	3.2	26.5
Aspen parkland	13 variables	48	11.8	0.73	7.1	60.4
Aspen parkland	13 + 3 dummies = 16 variables	48	11.8	0.81	6.4	54.6
Aspen parkland	10 + 3 dummies = 13 variables	48	11.8	0.80	6.2	52.5

Table 3. Ranking of variables by stepwise regression

Rank in / Variable	Grassland	Forest	Shrub	Aspen parkland	Average Rank
ST	6	10	1	3	2
AM	4	4	2	4	1
GL	11	3	7	11	7
BA	9	1	10	2	3
CC	1	2	12	5	2
OM	2	8	9	6	4
WHC	3	11	5	10	6
WP	10	12	6	7	8
BD	5	5	3	1	1
SAND	7	7	13	8	8
SILT	13	9	11	13	10
CLAY	12	13	8	12	9
PORS	8	6	4	9	5

Table 4. Individual and cumulative contribution of principal components

Principal component	1	2	3	4	5	6	7	8	9	10	11	12	13-16
A. All variables included:													
Eigenvalue	5.05	3.24	2.88	1.82	0.78	0.67	0.42	0.38	0.30	0.19	0.14	0.08	0.05
Contribution (%)	31.6	20.2	18.0	11.4	4.9	4.1	2.6	2.4	1.9	1.2	0.9	0.5	0.3
Cumulative (%)	31.6	51.8	69.8	81.2	86.1	90.2	92.8	95.2	97.1	98.3	99.2	99.7	100.0
B. All except SAND, SILT and CLAY:													
Eigenvalue	4.56	2.86	2.14	1.49	0.67	0.44	0.32	0.20	0.16	0.11	0.04	0.01	0
Contribution (%)	35.1	21.9	16.5	11.5	5.2	3.3	2.5	1.6	1.2	0.8	0.3	0.1	0
Cumulative (%)	35.1	57.0	73.5	85.0	90.2	93.5	96.0	97.6	98.8	99.6	99.9	100.0	100.0

Table 5. Total and mean prediction errors using 13-variable multiple regression prediction models derived from stratified populations

Vegetation Unit	Total prediction error				Mean prediction error				
	+ errors	- errors	absolute sum	actual sum	+ errors	- errors	Absolute sum	Actual sum	
I. Grassland									
Valley bottoms	289.3	0	289.3	289.3	18.1	0	18.1	18.1	Model derived from vegetation unit #3
Slopes	127.1	1084.8	1211.9	-957.7	7.9	67.8	75.7	-59.9	
II. Forest									
Black poplar	100.8	90.5	191.3	10.3	6.3	5.7	12.0	0.6	Model derived from vegetation unit #8
Aspen (rose understory)	29.1	367.7	396.8	-338.6	1.8	23.0	24.8	-21.2	
Aspen (pine grass understory)	16.0	339.2	355.2	-323.2	1.0	21.2	22.2	-20.2	
Mean	112.5	376.4	488.9	-263.9	7.0	23.5	30.5	-16.5	

Table 6. Total and mean prediction errors using a 16-variable multiple regression prediction model.

Vegetation Unit	Cumulative prediction error for the vegetation unit				Mean prediction error per observation for the vegetation unit			
	+ errors	- errors	Absolute sum	Actual sum	+ errors	- errors	Absolute sum	Actual sum
I. Grassland								
Valley bottoms (Phpr)	161.0	4.1	165.1	+ 156.9	10.1	0.2	10.3	+ 9.9
Predominantly Forb	295.9	0	295.9	+ 295.9	18.5	0	18.5	+ 18.5
II. Forest								
Black poplar	43.3	38.9	82.2	+ 4.4	2.7	2.4	5.1	+ 0.3
Aspen with rose understory	21.3	65.0	86.3	- 43.7	1.3	4.1	5.4	- 2.8
Aspen with pine grass understory	6.4	97.7	104.1	- 91.3	0.4	6.1	6.5	- 5.7
Mean	105.6	41.1	146.7	+ 64.5	6.6	2.6	9.2	+ 4.0

Table 7. Total and mean prediction errors using a principal component model derived from regression on 16 variables and 12 components.

	Total prediction error				Mean prediction error			
	+ errors	- errors	Absolute sum	Actual sum	+ errors	- errors	Absolute sum	Actual sum
I. Grassland								
Valley bottoms (Phpr)	106.8	18.4	125.2	+ 88.4	6.7	1.1	7.8	+ 5.6
Predominantly Forb	62.5	13.7	76.2	+ 48.8	3.9	0.9	4.8	+ 3.0
II. Forest								
Black poplar	12.2	96.8	109.0	- 84.6	0.8	6.0	6.8	- 5.2
Aspen with rose understory	7.8	122.7	130.5	-114.9	0.5	7.7	8.2	- 7.2
Aspen with pine grass understory	0	171.1	171.1	-171.1	0	10.7	10.7	- 10.7
Mean	37.9	84.5	122.4	- 46.6	2.4	5.3	7.7	- 2.9

Table 8. Total and mean prediction errors using 13-variable multiple regression models derived from stratified populations when sand, silt, and clay are excluded.

Vegetation Unit	Total prediction error				Mean prediction error				
	+ errors	- errors	Absolute sum	Actual sum	+ errors	- errors	Absolute sum	Actual sum	
I. Grassland									
Valley bottoms	274.4	0	274.4	274.4	17.1	0	17.1	17.1	Model derived from vegetation unit #3
Slopes	138.9	6.0	144.9	132.9	8.7	0.4	9.1	8.3	
II. Forest									
Black poplar	27.0	118.1	145.1	-91.1	1.7	7.4	9.1	-5.7	Model derived from vegetation unit #8
Aspen (rose understory)	21.4	321.7	343.1	-300.3	1.3	20.1	21.4	-18.8	
Aspen (pine grass understory)	1.6	380.0	381.6	-378.4	0.1	23.7	23.8	-23.6	
Mean	92.7	165.2	257.9	- 72.5	5.8	10.3	16.1	- 4.5	

Table 9. Total and mean prediction errors using a 13-variable multiple regression model when sand, silt and clay are excluded

Vegetation Unit	Total prediction error				Mean prediction error			
	+ errors	- errors	Absolute sum	Actual sum	+ errors	- errors	Absolute sum	Actual sum
I. Grassland								
Valley bottoms	110.5	11.4	121.9	+ 99.1	6.9	0.7	7.6	+ 6.2
Slopes	72.9	6.6	79.5	+ 66.3	4.5	0.4	5.0	+ 4.2
II. Forest								
Black poplar	12.4	99.8	112.2	- 87.4	0.8	6.2	7.0	- 5.5
Aspen (rose understory)	8.1	121.6	129.7	-113.5	0.5	7.6	8.1	7.1
Aspen (pine grass understory)	0	173.8	-173.8	-173.8	0	10.9	10.9	-10.9
Mean:	40.8	82.6	123.4	- 41.9	2.5	5.2	7.7	- 2.7

File: Singh

INTERNATIONAL GEOLOGICAL CONGRESS

24th SESSION

Montreal, Canada, August 21 - 30, 1972

Section 11: HYDROGEOLOGY

A COMPARATIVE STUDY OF MODELS TO PREDICT HYDROGEOLOGICAL RESPONSE

by

Teja Singh

Part of a 24 Vol. report in the library

DEPARTMENT OF THE ENVIRONMENT
CANADIAN FORESTRY SERVICE
NORTHERN FOREST RESEARCH CENTRE
EDMONTON 70, ALBERTA