MULTICOLLINEARITY WITHIN SELECTED WESTERN NORTH AMERICAN TEMPERATURE AND PRECIPITATION DATA SETS.

John Philip Cropper

ProSight Corporation
4626 E. Ft. Lowell Rd., Suite E
Tucson, Arizona 85712

ABSTRACT

This paper is concerned with examining the degree of correlation between monthly climatic variables (multicollinearity) within data sets selected for their high quality. Various methods of describing the degree of multicollinearity are discussed and subsequently applied to different combinations of climate data within each site.

The results indicate that higher degrees of multicollinearity occur in shorter data sets. Data consisting of 12 monthly variables of a single parameter (temperature or precipitation) have very low degrees of multicollinearity. Data set combinations of two parameters and lagged variables, as commonly used in tree-ring response function analysis, can have significant degrees of multicollinearity. If no preventative or corrective measures are taken when using such multicollinear data, erroneous interpretations of regression results may occur.

Der vorliegende Beitrag befaßt sich mit der Prüfung des Korrelationsgrades (Multikollinearität) zwischen monatlichen Klimawerten innerhalb eines Datensatzes. Es werden unterschiedliche Verfahren zur Beschreibung des Multikollinearitätsgrades diskutiert und anschliessend auf verschiedene Kombinationen von Klimadaten innerhalb eines Standortes angewandt.

Die Ergebnisse zeigen, daß in kürzeren Datenreihen höhere Kollinearitätsgrade vorliegen. Die 12 Monatswerte eines einzigen Parameters (Temperatur oder Niederschlag) weisen eine sehr geringe Multikollinearität auf. Datenkombinationen aus zwei Parametern und Variablen in Phasenverschiebung, wie dies in der Responsefunktionsanalyse allgemein üblich ist, können signifikante Multikollinearitätsgrade besitzen. Falls bei Benutzung multikollinearer Daten keine vorbeugenden oder korrektiven Gegenmaßnahmen getroffen werden, können Regressionsergebnisse falsch interpretiert werden.

Cet article concerne l'étude du degré de corrélation existant entre des variables climatiques mensuelles (multicollinéarité) au sein de séries choisies pour leur qualité élevée. Plusieurs méthodes utilisées pour décrire le degré de multicollinéarité sont soumises à discussion puis appliquées à différentes combinaisons des données climatiques au sein de chaque site.

Les résultats indiquent que des degrés de multicollinéarité plus élevés apparaissent dans des séries de données plus courtes. Les données représentant les 12 variables mensuelles d'un paramètre isolé (température ou précipitation) ont des degrés de multicollinéarité très bas. Les combinaisons de données formées par deux paramètres et des variables obtenues par décalage de séries telles qu'elles sont communément utilisées dans l'analyse des fonctions de réponse dendrochronologiques peuvent avoir des degrés significatifs de multicollinéarité. Si des mesures préventives ou corrections ne sont pas prises lors de l'utilisation de telles données multicollinéaires, on peut aboutir à des interprétations erronées des résultats obtenus par le calcul des régressions.

INTRODUCTION

Dendrochronologists are becoming more aware of possible statistical problems associated with the processes of calibration and prediction of tree-ring data. Potential problems exist in the basic techniques of tree-ring analysis such as indexing of tree-ring series, calculation of response functions, calibration against climatic series and reconstruction (or prediction, in the statistical sense) of past climatic events. The

problems can be divided into three general categories: (a) those within the tree-ring data (b) those within the climatic data and (c) interaction between tree data and climate data. Investigation by Cook and Peters (1981) on the use of cubic-spline curve fitting for standardization, and by Meko (1981) on the use of ARMA modeling techniques are primarily aimed at problems within the tree-ring data. Fritts (1976) has used principal components and cannonical regression techniques for calibration and climatic reconstruction in an attempt to minimize problems associated with interaction between tree and climate data. Little research has been aimed at problems within the climate data alone.

When data are used in a regression analysis, the presence of multicollinearity (correlations among the independent variables) can have many consequences. Johnston (1972, p160) lists the main consequences of multicollinearity in a regression analysis as: (1) loss of precision of the estimates (2) incorrect rejection of variables (3) overly sensitive estimates to particular data sets. This can result in incorrect regression weights (and subsequent interpretations), and unstable regression equations leading to poor quality reconstructions.

The purpose of this study is to investigate to what extent there is any correlation among climatic variables from western North America meteorological stations and hypothesize how this may affect the results of any subsequent multivariate analysis.

CHOICE OF CLIMATE DATA SET.

The data set used in this study is a subset of quality-tested data available from North America. Originally 167 meteorological station records were evaluated for the coherency of their temperature and precipitation records (DeWitt 1978). Data grids of temperature and precipitation stations were developed as a result of the above analysis. From two of these quality-tested data grids a subset of 32 stations from western North America consisting of stations that had both temperature (46-grid) and precipitation (52-grid) records (DeWitt 1978) was selected (Table 1, Figure 1). In this way, each of the available 32 climatic data sets contained quality-tested data for both the temperature and precipitation variables. In addition to the selected stations from western North American, data from a meteorological station used for an unpublished comparison of response function programs (Wigley and Lough 1981) were also included for analysis (these data have been called "Norwich" data).

MEASURES OF MULTICOLLINEARITY

The use and interpretation of a multiple regression model often depend explicitly or implicitly on the estimates of the individual regression coefficients. If there is no linear relationship between the regressors, they are said to be orthogonal. When the regressors are orthogonal, it is relatively easy to make inferences that identify the relative effects of the regressor variables, or to select an appropriate set of variables for the model (Montgomery and Peck 1982, p287). In some situations the regressors are nearly perfectly linearly related, and in such cases the inferences based on the multiple regression model can be misleading or erroneous. How misleading the results are depends to a large degree on the severity of the multicollinearity.

Several techniques have been proposed for detecting multicollinearity, but little or no work has been done concerning critical levels for each technique. The desirable characteristics of a diagnostic procedure are that it directly reflect the degree of the

multicollinearity problem and provides information helpful in determining which regressors are involved.

Table 1. Names of Climatic Data Stations Analysed.

Map No.	Name	ID
1	FLAGSTAFF, Arizona	00223010
2	PHOENIX, Arizona	00266481
3	TUCSON, Arizona	00278815
4	YUMA WSO AP, Arizona	00259660
5	EUREKA, California	00412910
6	NEEDLES FAA AIRPORT, California	00476118
7	RED BLUFF, California	00427292
8	RIVERSIDE, California	00467470
9	SACRAMENTO, California	00427633
10	SAN DIEGO, California	00467740
11	VISALIA, California	00459367
12	CANON CITY, Colorado	00511294
13	GRAND JUNCTION, Colorado	00523488B
14	TRINIDAD, Colorado	00518429
15	BOISE, Idaho	01051002
16	HAVRE, Montana	02433996
17	HELENA, Montana	02444055B
18	KALISPELL, Montana	02414558
19	ELKO FAA AP, Nevada	02622573
20	RENO, Nevada	02616779
21	ROSWELL, New Mexico	02977610
22	BAKER KBKR, Oregon	03580417
23	ROSEBURG, Oregon	03537331
24	RAPID CITY, S. Dakota	03956937
25	AMARILLO, Texas	04110211
26	EL PASO, Texas	04152797
27	PAROWAN, Utah	04246686
28	ABERDEEN, Washington	04510000B
29	YELLOWSTONE PARK, Wyoming	04819905B
30	BANFF, Alberta	06150520
31	CALGARY, Alberta	06131101
32	KAMLOOPS, British Columbia	06263779
	NORWICH data, England	Not available

A very simple measure of multicollinearity is obtained by inspection of the off-diagonal elements in the $\tilde{X}X$ correlation matrix (a matrix used in deriving regression coefficients) (see Draper and Smith 1981 p74). If two regressors are nearly linearly dependent then the degree of correlation between them will be near unity. Examining the simple correlation between the regressors is helpful in determining near linear dependency between pairs of regressors only. When more than two regressors are involved in a near linear dependency, there is no assurance that any of the pairwise

correlations will be large. Thus, although this is a very simple technique, inspection of the pairwise correlations is generally not sufficient for detecting anything more complex than pairwise multicollinearity.

If any multicollinearity exists within the $\acute{X}X$ correlation matrix it results in large values along the diagonal of the inverse $\acute{X}X$ matrix. Since the diagonal of this inverse is used to determine the variance of individual regression coefficients we can view the diagonal as a series of factors by which the variances of the estimated regression coefficients are increased due to near linear dependencies among the regressors. Marquardt (1970) has called these diagonal elements the "Variance Inflation Factors" (VIF). The VIF for each term in the model measures the combined effect of the dependencies among the regressors on the variance of that term. One or more large VIF's indicate multicollinearity. Practical experience indicates that if any of the VIF's exceed 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery and Peck 1982, p300). The largest VIF can be used as an indication of the degree of multicollinearity in a particular data set.

The characteristic roots (eigenvalues) of $X\acute{X}$ can be used to measure the extent of multicollinearity in the data. If there are one or more near linear dependencies in the data, then one or more of the characteristic roots will be small.

Some analysts (Casella 1977, Faden 1978) prefer to use the condition number of $\acute{X}X$, defined as the ratio of the maximum to minimum eigenvalue, as a measure of the degree of multicollinearity. Faden (1978) considered a condition number of 40 to indicate low multicollinearity and any value above 1000 to indicate serious multicollinearity. Montgomery and Peck (1982, p301) indicate that if the condition number is less than 100 there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity and if the number exceed 1000 then severe multicollinearity is indicated.

Another measure of the degree of multicollinearity, used by Hoerl and Kennard (1970, p70), is the sum of the reciprocals of the eigenvalues of the $\acute{X}X$ correlation matrix. In a 10 variable example they considered a sum of 33.8 to indicate a large number of significant interfactorial correlations. When this value is divided by the number of variables in the problem it indicates how many more times larger the variances of the regression coefficients are than they would be for an orthogonal system. This measure has many similarities to the VIF discussed earlier, in fact the sum of all the VIF's for a problem equals the sum of the reciprocals of the eigenvalues.

In addition to the above statistics, which describe the amount of inflation caused by linear associations within the data set, another measure of multicollinearity is the determinant of the $\acute{X}X$ correlation matrix. The determinant of a matrix is computed during the process of matrix inversion. If the matrix were "singular" then there would be high multicollinearity, the determinant would equal zero and no inverse would exist. Compared to the above mentioned statistics, the determinant is a poor measure because it gives little information that can be interpreted directly in respect to the estimates of the regressors or their variances.

DEGREE OF MULTICOLLINEARITY IN DATA

In evaluating the set of climatic data, the measures of Hoerl (sum of reciprocal eigenvalues), Marquardt (VIF) and Faden (condition number) were used in addition to the matrix determinant. For although the first three statistics are closely related they do measure slightly different aspects of the data.

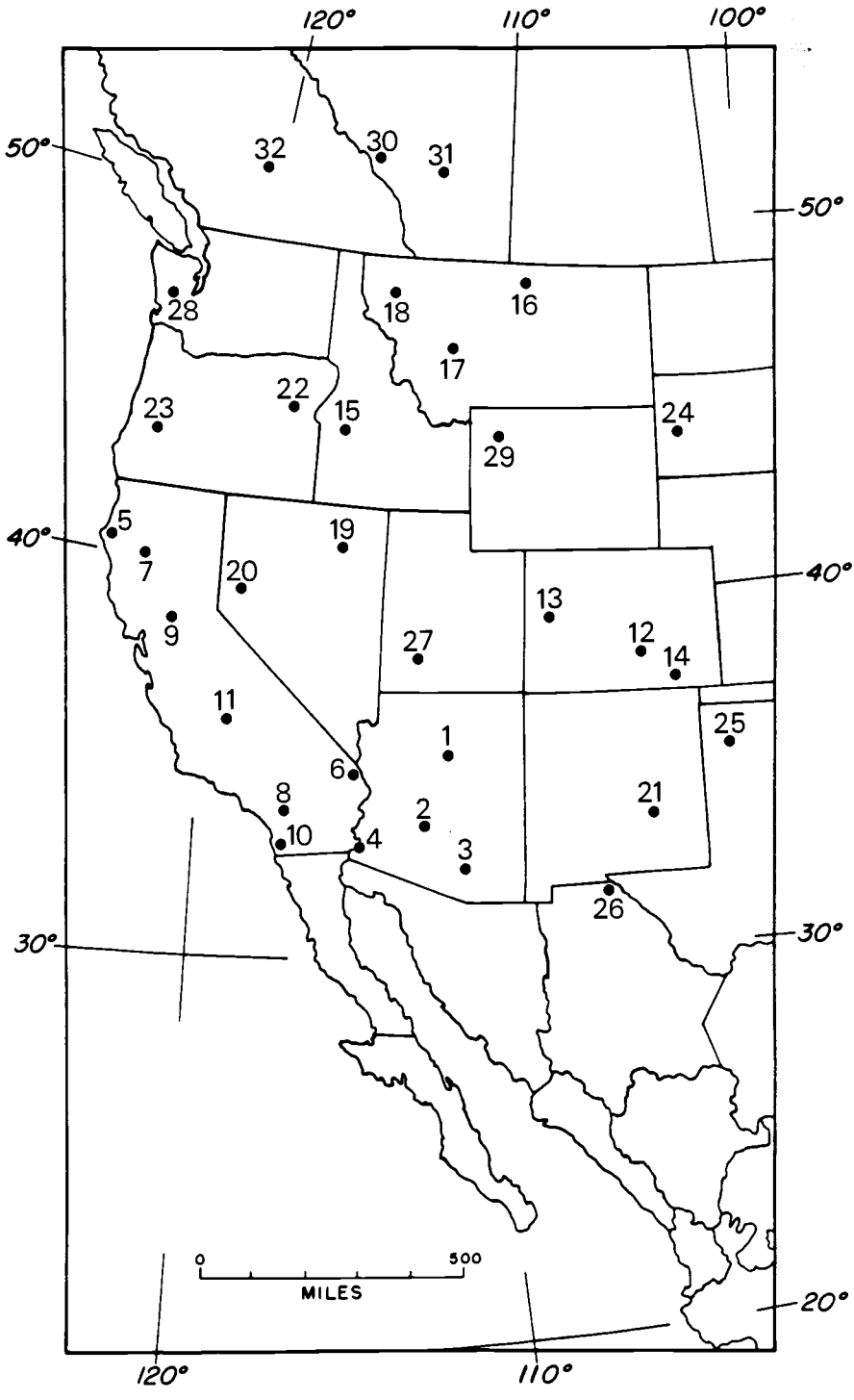

FIGURE 1. Locations of the 32 western North American climate data sites.

Each of the 32 western North American climatic data sets, plus the Norwich data, were arranged into 12 configurations (6 mixtures each for 2 time periods) prior to their evaluation for degree of multicollinearity. The different configurations are:

i 12 months (prior September to August) of temperature
ii 12 months (prior September to August) of precipitation
iii 15 months (prior June to August) of temperature
iv 15 months (prior June to August) of precipitation
v 12 months of temperature PLUS 12 months of precipitation
vi 15 months of temperature PLUS 15 months of precipitation

The data were analysed for the maximum record length available for a site and also for a common record length of the most recent 58 years.

Table 2. Summary of Multicollinearity Results for 58 Year Data Sets.

DATA SET	VIF	DET ẊX	HOERL	FADEN
Temperature 12 months				
MAX	3.614	.003522	26.295	39.198
MIN	1.346	.268423	14.972	5.044
RANGE	2.268	.264901	11.323	24.154
Precipitation 12 months				
MAX	2.179	.149778	17.096	9.314
MIN	1.261	.394121	14.123	3.920
RANGE	0.918	.244343	2.973	5.494
Temperature 15 months				
MAX	4.195	.000257	37.985	51.670
MIN	1.438	.128634	19.851	6.820
RANGE	2.757	.128377	18.134	44.850
Precipitation 15 months				
MAX	3.560	.052909	24.181	16.946
MIN	1.429	.195787	18.843	4.695
RANGE	2.131	.142878	5.338	12.251
Temp & Prec 24 months				
MAX	5.093	.000010	61.159	50.310
MIN	2.306	.000953	44.327	17.013
RANGE	2.787	.000943	16.832	33.297
Temp & Prec 30 months				
MAX	7.794	.000000	98.681	119.644
MIN	3.012	.000016	65.313	34.292
RANGE	4.782	.000016	33.368	85.352

The results of the multicollinearity test showed that, in nearly all cases, the full time-period data sets were less nearly collinear than their 58 year abbreviated sets. The full time periods ranged from 59 years for the Norwich data to 129 years for the San Diego record. The range of years involved in the full data analyses does not allow easy valid site-to-site comparisons of the multicollinearity results. For the 58 year data sets a summary analysis (Table 2) indicates that the 12 month precipitation data set is the

Table 3. Multicollinearity Data:
58 Year, 30 Months (15 Temperature & 15 Precipitation)

LOCATION	VIF	DET $\bar{X}X$	HOERL	FADEN
FLAGSTAFF, Arizona	4.3	.000003	77.3	49.1
PHOENIX, Arizona	4.7	.000000*	81.3	54.1
TUCSON, Arizona	4.1	.000000*	84.6	64.8
YUMA WSO AP, Arizona	3.6	.000003	75.6	49.0
EUREKA, California	7.7*	.000001	82.5	66.2
NEEDLES FAA, California	3.3	.000007	74.4	36.9
RED BLUFF, California	3.6	.000002	76.3	39.3
RIVERSIDE FIRE, California	3.4	.000016	65.3	34.3
SACRAMENTO, California	3.4	.000006	70.4	35.2
SAN DIEGO, California	6.5*	.000000*	98.7*	119.6*
VISALIA, California	7.8*	.000001	85.7	64.8
CANON CITY, Colorado	4.1	.000004	72.5	35.7
GRAND JUNCTION, Colorado	3.6	.000005	74.4	42.6
TRINIDAD, Colorado	4.2	.000002	78.5	42.7
BOISE, Idaho	3.5	.000016	67.7	35.4
HAVRE, Montana	3.8	.000002	78.3	40.8
HELENA, Montana	5.0*	.000000*	90.5*	45.7
KALISPELL, Montana	3.0	.000009	68.2	36.4
ELKO FAA AP, Nevada	3.7	.000005	73.8	51.8
RENO, Nevada	6.0*	.000002	77.0	62.1
ROSWELL, New Mexico	4.6	.000000*	83.4	62.1
BAKER KBKR, Oregon	3.9	.000013	71.0	45.2
ROSEBURG, Oregon	3.6	.000009	68.8	35.8
RAPID CITY, South Dakota	3.3	.000010	70.3	43.6
AMARILLO, Texas	4.3	.000000*	86.1	62.8
EL PASO, Texas	5.6*	.000002	77.9	48.0
PAROWAN, Utah	4.7	.000001	80.5	59.2
ABERDEEN, Washington	4.1	.000001	78.4	44.5
YELLOWSTONE PARK, WY	3.8	.000010	73.0	37.4
BANFF, Alberta	4.1	.000002	81.9	45.3
CALGARY, Alberta	5.6*	.000000*	91.3*	70.5
KAMLOOPS, B. C.	4.5	.000001	80.1	37.7
NORWICH DATA, England	6.1*	.000003	78.2	38.6

*indicates value greater than warning level.

most stable (values for all statistics closest to 1.0) and has the smallest range of values for most of the statistics. This stability is not unexpected due to the observed lack of coherency among different months of precipitation data. There is little difference among the remaining single-parameter data sets, with only slight increases in multicollinearity from precipitation data to temperature data and from 12 to 15 variable sets (Cropper 1984). The most seriously multicollinear data are to be found in the 30-variable mixed temperature and precipitation data sets. In this configuration there are 8 sites, out of the 33 tested, that have VIF scores of 5.0 or greater, 7 sites with matrix determinants less than 0.0000005. Three sites have values for the Hoerl statistic greater than 90.0 (3.0 * 30 variables) and one site had a value for the Faden statistic greater than 100.0 (Table 3). Thus, of the 33 data sets tested 12 yielded multicollinearity warnings from at least one evaluation criterion and three yielded multiple warnings.

The most commonly used data mixes in tree-ring response function analysis include from 28 to 32 variables distributed equally between the two variables of temperature and precipitation. Therefore these response function analyses would be most susceptible to the problems of multicollinearity.

CONCLUSIONS

The tests described above indicate that, depending upon the amount and type of mixing of the data, the degree of multicollinearity can increase many fold. In all cases when a calendar year of a single variable type (temperature or precipitation) is used, the degree of multicollinearity is minimal and should have little effect on subsequent analyses. When two variable types are used together and lagging within the variables is also included there is a potential of larger, possibly serious, degrees of multicollinearity.

Depending upon the purpose of the multivariate analysis that is using the climatic data, the effects of multicollinearity have varying impacts. Montgomery and Peck (1982) indicate that predictive models are little affected, but that the estimates of the individual regression weights (as in response functions) are much affected by multicollinearity. Thus if the aim (or sideline) of this analysis is to indicate which of the input variables is most important in affecting tree growth and the degree of multicollinearity within the independent data is not taken into account, then there is a possibility that the interpretation could be in error.

ACKNOWLEDGEMENTS

These analyses were performed while being supported as a Graduate Associate by funds to the Laboratory of Tree-Ring Research from the State of Arizona and by the National Science Foundation, Climate Dynamics Section, under Grants ATM-8303192 and ATM-8115754 (H. C. Fritts principal investigator).

REFERENCES

Casella, G.
1977 Minimax Ridge Regression Estimation. PhD dissertation, Purdue University.
Cook, E. R. and Peters, K.
1981 The smoothing spline: a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. *Tree-Ring Bulletin* 41: 45-53.

Cropper, J. P.
1984 Multicollinearity within western North American climatic data. Northern Hemisphere Climatic Reconstruction Group. Laboratory of Tree-Ring Research, Technical Note 32. University of Arizona, Tucson.

DeWitt, E.
1978 Temperature and Precipitation Station Selection. Northern Hemisphere Climatic Reconstruction Group. Laboratory of Tree-Ring Research, Technical Note 3. University of Arizona, Tucson.

Draper, N. and Smith, H.
1981 *Applied Regression Analysis* (Second edition). John Wiley and Sons.

Faden, V. B.
1978 Shrinkage in Ridge Regression and Ordinary Least Squares Multiple Regression Estimators. PhD Dissertation, University of Maryland.

Fritts, H. C.
1976 *Tree Rings and Climate.* Academic Press, London.

Hoerl, A. E. and Kennard, R. W.
1970 Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55-67.

Johnston, J.
1972 *Econometric Methods* (Second Edition). McGraw-Hill, New York.

Marquardt, D. W.
1970 Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* 12: 591-612.

Meko, D. M.
1981 Applications of Box-Jenkins Methods of Time Series Analysis to the Reconstruction of Drought from Tree-Rings. PhD Dissertation, University of Arizona, Tucson.

Montgomery, D. C., and L. A. Peck
1982 *Introduction to Linear Regression Analysis.* John Wiley and Sons, New York.

Vinod, H. D.
1976 Application of new ridge regression methods to a study of Bell System scale economies. *Journal of the American Statistical Association* 71(356): 835-841.

Wigley, T. M. L. and J. M. Lough
1981 Tree-Ring Response Functions: a Review and Program Intercomparison. Climatic Research Unit, University of East Anglia, Norwich, England.