Published November, 1990

Statistical Comparison of Soil Map-Unit Boundaries

M. H. Nash and L. A. Daugherty*

ABSTRACT

Locating the exact boundaries of soil map units is one of the primary objectives for soil surveyors. Statistical methods were used to assure the most accurate location. Soil spatial variability, autocorrelation function, and soil boundary locations were examined along a 2700-m transect in southern New Mexico. Eighty-nine observation points were equally spaced along the transect. Selected physical and chemical characteristics through the transect were determined. A multivariate method of principal-component analysis was used to produce one set of data. These data were first inspected for stationary manner, i.e., that the mean and variance of each property remain fairly constant for each data set. Log-normal transformation was used to detrend the data. The stationary manner of autocorrelations was tested with semivariograms. The range of dependence obtained from the autocorrelations and semivariograms was used in a squared-Euclidean-distance procedure to locate the soil boundaries. These boundaries were compared with those obtained by conventional soil-survey methods. Some of the calculated boundaries agreed with those obtained by conventional soil survey. The latter method is more economical and more productive than the statistical method.

A LONG-TERM ECOLOGICAL RESEARCH (LTER) program of the National Science Foundation was established in 1982 to evaluate the effects of human perturbations on the stability and productivity of major ecosystems in the USA. At one LTER site in southern New Mexico on the Jornada Experimental Range, data are collected on numerous biotic factors that are affected by soil variability. A knowledge of soil variability, therefore, is essential to properly monitor and understand much of the LTER data.

One of the major difficulties in devising a system of soil classification for an area is to create classes that are spatially coherent and allow local mapping to be done easily and sensibly. In conventional soil survey, observations are taken at intervals related to the rate of change of the soil; knowledge of the boundaries is essential to objectively determine soil change. Boundary location by the field soil surveyor is subject to interpretation (Webster, 1973).

The soils along a transect at the LTER site were surveyed using conventional techniques of the USDA Soil Conservation Service (Soil Survey Staff, 1972; Nash and Daugherty, 1990). In addition, detailed soil data were collected along the transect to aid in the location of soil boundaries through numerical processes. The objectives of the study were to (i) determine the range of dependence of selected soil variables and apply the information to soil boundary location, and (ii) compare the boundaries obtained by geostatistical methods with those obtained by conventional soil survey.

MATERIALS AND METHODS

The study was conducted at the LTER site on the New Mexico State University College Ranch, about 40 km north of Las Cruces. The data were collected along a transect extending from an ephemeral dry lake (playa) to the adjacent piedmont slope. The transect was a chronosequence that traversed three geomorphic surfaces (Gile and Grossman, 1979). The lower (youngest) end of the transect is on the basin floor (Lake Tank surface) or playa, which dates from late Pleistocene. The middle part is on the Jornada II geomorphic surface, while the upper part is on the Organ and Isaack's complex geomorphic surface. The middle and upper parts are of Pleistocene age.

The 3-km-long transect was divided into 89 observation sites spaced 30 m apart. At each site, samples were taken to a depth of 120 cm, divided into four intervals (0–30, 30–60, 60–90, and 90–120 cm). There were 356 samples (89 sites $\times$ 4 depths) taken as bulk samples with a 7-cm-diam. bulk auger. Particle-size analysis was determined by the pipette method, while $CaCO_3$ and organic C were determined by the titration method (Soil Survey Staff, 1972). Soil pH was determined in a 1:1 soil/water suspension (Peech, 1965). The variables used in this study are among those most used in soil survey, namely, clay, very coarse sand (VCS), coarse sand (CS), medium sand (MS), fine sand (FS), very fine sand

Dep. of Agronomy and Horticulture, New Mexico State Univ., Las Cruces, NM 88003-0003. Journal article JA1449 of the Agric. Exp. Stn., New Mexico State Univ., Las Cruces. Received 17 Apr. 1989. *Corresponding author.

Published in Soil Sci. Soc. Am. J. 54:1677–1681 (1990).

(VFS), silt (SI), coarse fragments (CF), $CaCO_3$, organic matter, and soil pH.

For the purpose of this study, soil boundary locations along the transect were established by using the procedure outlined by the Soil Survey Staff (1972), with a little modification, using the following procedure (Nash and Daugherty, 1990):

1. Samples from each depth along the transect were examined in the field.
2. The information concerning the change in soil texture, coarse fragments, soil color, and $CaCO_3$ obtained from each site was evaluated and recorded.
3. The change in landscape, and macro- or microtopography were considered along with the information obtained from each site to locate the soil boundaries.
4. A site most representative of each map unit was chosen for detailed sampling and was described according to Soil Survey Staff (1972) procedures.

Statistical Approach

Soil surveyors observe numerous soil parameters in order to group similar soils. In statistics, there are several multivariate methods that can be used to produce sets of data that are most similar or dissimilar. In this study, two multivariate methods were chosen: (i) principal-component analysis was used to find the range of dependence of the multivariate data by using geostatistical procedures; and (ii) squared Euclidean distance was used to produce sets of data that have the most differences among their properties. In the squared-Euclidean-distance procedure, the charted peaks of the differences are used to determine soil boundaries. The procedure requires the choice of a window that includes several sample sites (Fig. 1). One problem with the squared-Euclidean-distance method is the choice of the correct window size. If the window is too narrow, the resulting figures will be noisy and many peaks will appear. If, on the other hand, the window is too wide, boundaries may be hidden because the window includes two or more boundaries (Burgess et al., 1981; Webster, 1978). To overcome this problem, a geostatistical method was used to obtain the range of dependence of the data set. Therefore, semivariance and autocorrelations were calculated.

Principal-Component Analysis

Principal-component analysis is one of the multivariate methods used in this study. Principal-component analysis is a method that finds a set of orthogonal axes in the direction of greatest variance among individuals. These axes are linear combinations of the original variant of the linear formula

$$C_{i1} = A_{i1}X_1 + \ldots\ldots + A_{in}X_n \quad i = 1,2,3, \ldots, n \quad [1]$$

where there are n variables, $X_1, X_2, \ldots, X_n$.

The coefficients, A, are chosen in such a way that the first component, C_{i1}, has as large a variance as possible, and has maximum contribution to the total variance (Hair et al., 1979, p. 85–112). The components are ranked in order according to the proportion of the total variation. If the original variates are highly correlated, a single principal component may express most of the variation and may be adequate as a measure of the individuals (Webster and Wong, 1969). A Statistical Analysis System program (SAS Institute, 1982) was used to calculate the principal components in this study.

Squared Euclidean Distance

Squared Euclidean distance (SED) is the other multivariate-analysis technique used in this study. When there are two variables to be considered simultaneously, likeness between any two individuals can be measured as the distance between them (Webster, 1973). The distance can be calculated by Pythagora's theorem. If the coordinates of two points i and j, are Z_{i1} and X_{j1}, and Z_{i2} and X_{j2}, then the distance D_{ij} between them is given by the formula

$$D_{ij} = [(Z_{i1} - X_{j1})^2 + (Z_{i2} - X_{j2})^2]^{1/2} \quad [2]$$

If there are n values, then Eq. [2] can be rewritten

$$D_{ij} = \sum_{i=1}^{n} [(Z_{ij} - X_{ij})^2]^{1/2} \quad [3]$$

The distance D_{ij} is often known as pythagorean distance, Euclidean distance, or taxonomic distance between the individuals. Therefore, adjacent samples along the transect can be compared for similarity by using a simple SED. A plot of the squared Euclidean distance with position along the transect will show strong peaks if the midpoint of the window is at or near a boundary.

Semivariance

Specialized statistical methods, known as regionalized variable theory, have been developed for the study of local variation (Matheron, 1971). A regionalized variable is a variable with a definite value at each point in space, i.e., it has a value at every point X in three-dimensional space. If regionalized variables $Z(X)$ and $Z(X + h)$ had a set of pairs of samples for specific h intervals apart called the lag, then variability between these regionalized variables can be estimated by the semivariance $\gamma(h)$ as calculated in the following equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2 \quad [4]$$

where $Z(x_i)$ is the value of the variable measured at point x, and $N(h)$ is the number of pairs of points separated by a distance h (Jornel and Huijbregts, 1978).

A semivariogram represents the similarity that exists between the variable measured at one point and the variable measured at another point some distance away. This variation may be called spatial similarity or spatial correlation.

Autocorrelation

Initially, changes in soil properties along a transect were analyzed using an autocorrelation method (Webster, 1977; Webster and Cuanalo de Lac, 1975). With this method, dependence of a soil property with distance is expressed by autocorrelation. The plot of the autocorrelation function with respect to the distance is called an autocorrelogram. The autocorrelogram represents the relationship between the autocorrelation coefficient (a measure of the linear correla-

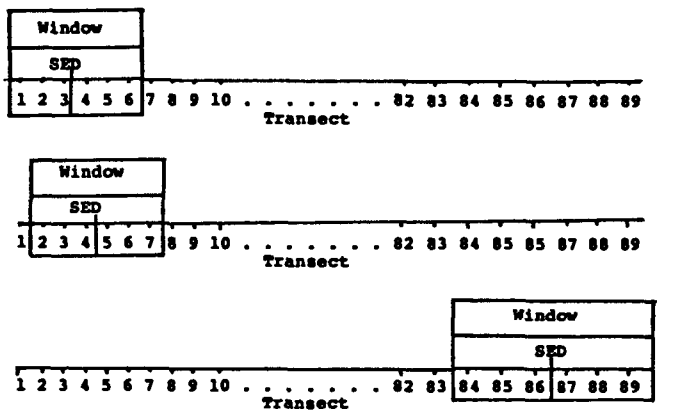

Fig. 1. Procedure for calculating sliding window distance by squared Euclidean distance (SED) along a transect with 89 observations. Window width set equal to six.

tion between a spatial series and the same series at lag h) and the values of h.

An equation used to calculate the autocorrelation function (Davis 1973, p. 232) is:

$$r(h) = \frac{[(n-h)(\Sigma X_{i+h}X_i) - \Sigma(X_{i+h})\Sigma(X_i)]/(n-h)(n-h-1)}{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2]/n(n-1)} \quad [5]$$

where

$r(h)$ = autocorrelation function value, and
n = number of observations

If the correlogram shows a high correlation between (x_i) and (x_{i+h}), the observations are dependent. As the distance (h) increases, the covariance decreases gradually until it becomes zero, and no correlation or spatial dependence exists between the two series at that lag.

Removing the Drift

If data exhibits drift, the trend should be removed (reference). A logarithmic transformation model was used to remove the drift (trend) from these data. The model used is represented in the following equation:

$$y' = \beta_0 + \beta_1 x' + e \quad [6]$$

where

y' = $\log_{10} y$;
x' = $\log_{10} x$;
e = error term;
β_0 = the intercept; and
β_1 = the slope.

The residuals of the fitted model were used in subsequent calculations of the semivariograms and correlograms (Olea, 1975). Therefore, the covariance and the variance of the residual were plotted vs. h. If the semivariogram and the covariance overlap, the data are stationary (Olea, 1975). The range then can be determined from the intercept of the semivariogram with the total variance or the sill. This should be at the same position that the covariance intercepts the sill and the correlogram reaches zero or a negative value.

RESULTS AND DISCUSSION

The principal-component-analysis method is a way to arrange the multivariate data into individuals of principal-component variables along one or more axes (Webster, 1977). The first principal component (Table 1) is the most significant and was used to generate a new set of data to be used in geostatistical techniques (Webster and Wong, 1969; Webster, 1973; Webster and Cuanalo de Lac, 1975; Nash, 1985).

The variograms and the correlograms for the first principal components of the 10 variables were overlaid to test if these data are stationary, i.e., if no trends are present (Olea, 1975) (Fig. 2). Note that the autocovariance scale is inverted because of a reciprocal relation with semivariance for the stationary process. None of the sample depths in Fig. 2 meet the station-

Table 1. Eigenvalues and percent total variation represented by the first principal component for each sample depth.

Depth	Eigenvalues	Total variance
cm		%
0–30	4.07	40.73
30–60	4.35	43.50
60–90	4.66	46.59
90–120	4.66	46.59

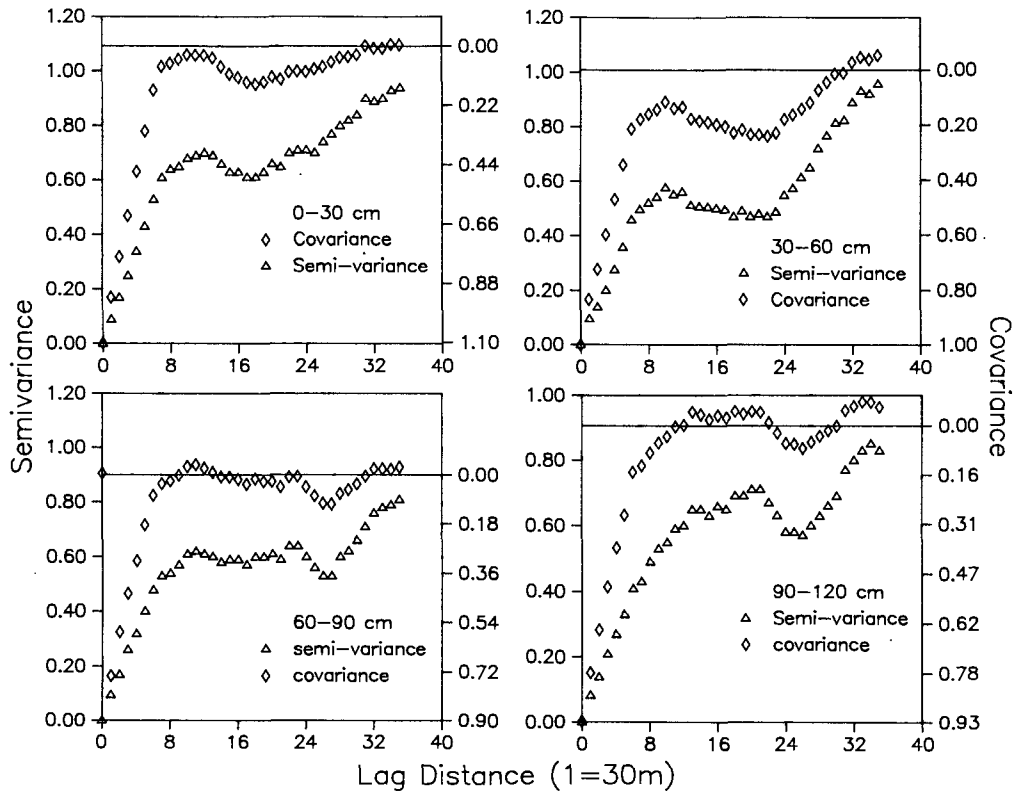

Fig. 2. Semivariogram for the original data for the first principal component at depths of 0 to 30, 30 to 60, 60 to 90, and 90 to 120 cm along the transect.

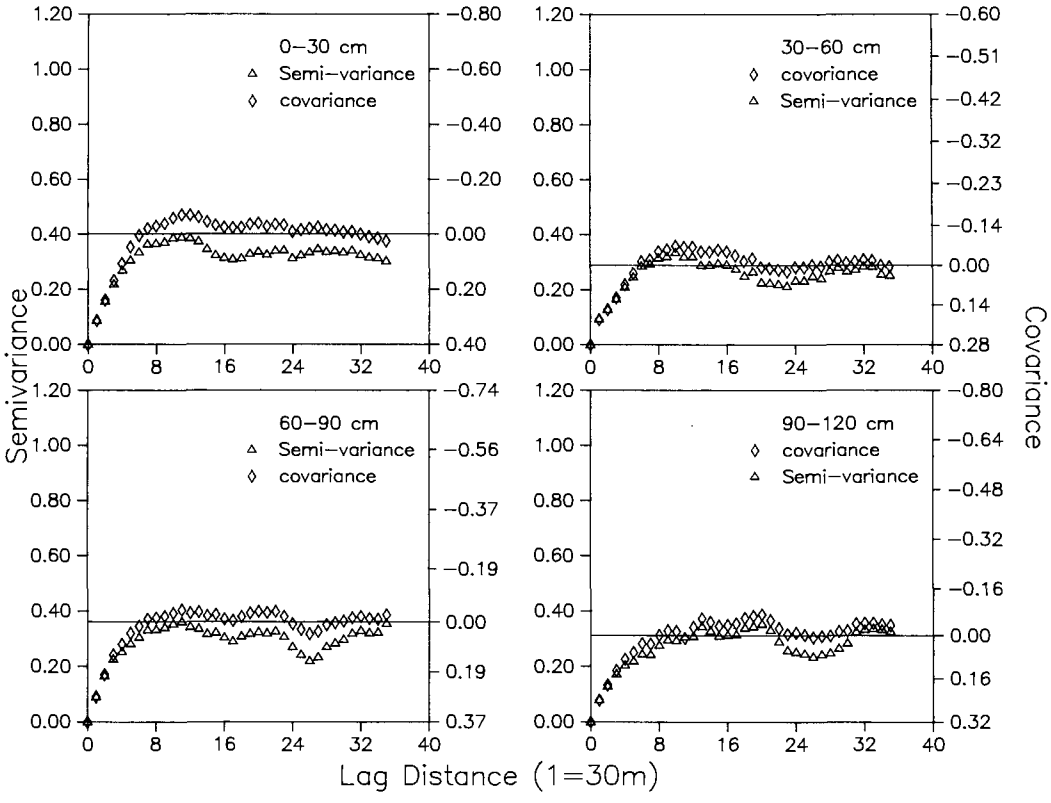

Fig. 3. Semivariogram for the residual data for the first principal component at depths of 0 to 30, 30 to 60, 60 to 90, and 90 to 120 cm along the transect.

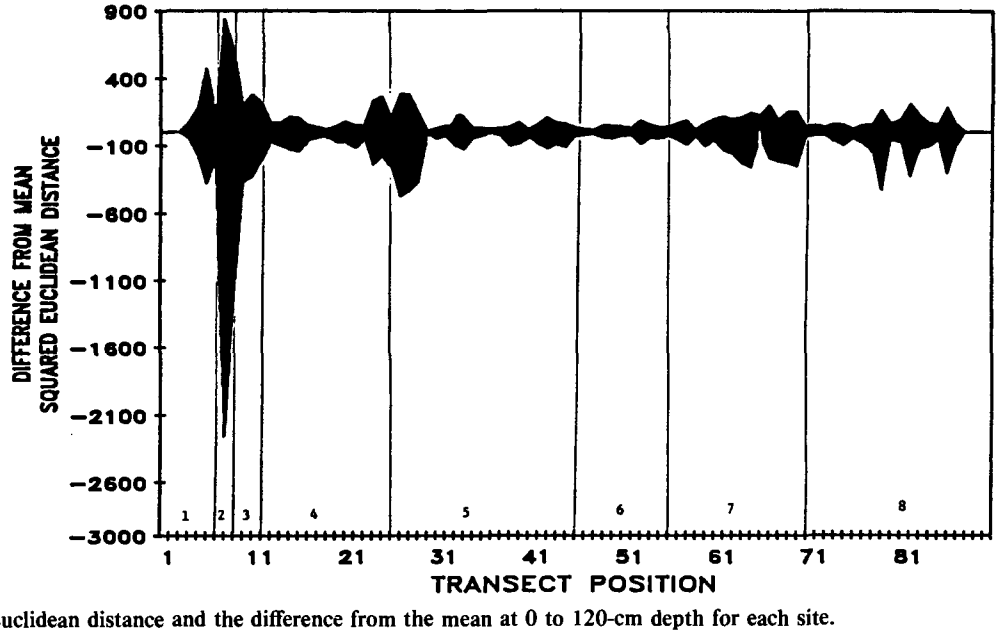

Fig. 4. Squared Euclidean distance and the difference from the mean at 0 to 120-cm depth for each site.

ary condition. The residuals from the logarithmic transformation to detrend the first principal-component data were used to find the rank of dependence. Figure 3 shows the semivariogram and autocorrelogram for the residual data for the first principal component. The autocovariance scale is inverted because of a reciprocal relation with semivariance for the stationary process.

The first principal component, after removing the trend, has the following lags: for the first depth of these variables, 0 to 30 cm, $h = 7$ (210 m), for the second depth, 30 to 60 cm, $h = 6$ (180 m), for the third depth, 60 to 90 cm, $h = 7$ (210 m), and for the fourth 90 to 120 cm, $h = 9$ (270 m) (Fig. 3). The lag distance or the range of dependence was chosen to initiate the choice of the width of the window in the search for boundaries. According to Webster (1978), the smaller the window size, the better the prediction of soil

Table 2. Major soils series and station position and classification along the transect.

Zone		Stations		
No.	Series	Range	No.	Classification
1	Dalby variant	1–6	6	Typic Torrert, fine
2	Headquarters variant	7	1	Ustollic Haplargid, fine-loamy
3	Headquarters	8–10	3	Ustollic Haplargid, fine-loamy
4	Bucklebar	11–25	15	Typic Haplargid, fine-loamy
5	Berino	26–45	20	Typic Haplargid, fine-loamy
6	Onite	46–55	10	Typic Haplargid, coarse-loamy
7	Dona Ana variant	56–70	15	Typic Haplargid, coarse-loamy
8	Aladdin	71–89	19	Torriorthentic Haplustoll, coarse-loamy

† All soils have mixed mineralogy and thermic temperature regimes.

boundary by using the SED method. A window size of six observations was chosen.

The SED method produced prominent peaks on the SED graphs (Fig. 4). Soil positions located in the field were tested with this boundary (Table 2). This method found similar boundaries to those located by the field survey technique, but the comparison was not exact. Most of the sharp boundaries recognized in the field are represented by prominent peaks. The gradual boundaries separating the Bucklebar, Berino, and Onite map units are also evident by using the SED method. The boundaries obtained by this method are comparable for all depths (Fig. 4).

This procedure, however, suggested an additional boundary near Position 80 (2400 m). This new boundary is related to change in coarse-fragment and sand content. At this position, therefore, the coarse fragments become a dominant feature. This boundary was missed by the conventional soil survey because of the apparent similarity of topsoil after Site 70.

SUMMARY AND CONCLUSION

For data of the type commonly obtained from soil transects, correlograms or semivariograms can be interpreted to determine the average spacing between soil observations (Webster, 1978). This spacing can then be used to set the width of a window in a systematic search for individual boundaries on a transect. In this study, the range of dependence was found to be 180 m. This range was used to design the window width in the SED method (Webster, 1978). This method showed boundaries close to those found by field survey techniques, but not at the exact location, and also suggested an additional boundary near Site 80 (2400 m). The SED method found the new boundary near Site 80 due to the change in coarse fragments and sand content of these sites.

The analysis used in this study produced a good agreement between the boundaries obtained by statistical methods and those obtained by conventional methods. The correlation between boundaries observed in this study with boundaries located by conventional soil-survey methods supports the use of conventional soil-survey procedures for locating soil boundaries for most soil-survey uses. The analysis used in this study shows the variation vertically and horizontally between the soil sites. It is a good tool for transect evaluation in detailed soil survey, and is a good indicator of variation.

REFERENCES

Burgess, T.M., R. Webster, and A.B. McBratney. 1981. Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. J. Soil Sci. 32:643–659.

Davis, J.C. 1973. Statistics and data analysis in geology. John Wiley & Sons, New York.

Gile, L.H., and R.B. Grossman. 1979. The desert project soil monograph. USDA-SCS.

Hair, J.F., Jr., R.E. Anderson, R.L. Thatem, and B.J. Grjablowsky. 1979. Multivariate data analysis with readings. Pennwell Publ., Tulsa, OK.

Journel, A.G., and C.H.J. Huijbregts. 1978. Mining geostatistics. Academic Press, New York.

Matheron, G. 1971. The theory of regionalized variables and its applications. Les Cahiers du Centre Morphologie Mathematique de Fontainebleau no. 5. Edit par l'Ecole Nationale Superieur des Mines de Paris.

Nash, M.H. 1985. Numerical classification, spatial dependence, and vertical kriging of soil sites in southern New Mexico. M.S. thesis. New Mexico State Univ., Las Cruces.

Nash, M.H., and L.A. Daugherty. 1990. Soil–landscape relationships in alluvium sediments in southern New Mexico. N.M. Agric. Exp. Stn. Bull. no. 746. New Mexico State Univ., Las Cruces.

Olea, R.A. 1975. Measuring statistical dependence with semivariograms. Series on spatial analysis, no. 2. Kansas Geol. Surv., Lawrence.

Peech, M. 1965. Hydrogen-ion activity. p. 914–926. *In* C.A. Black (ed.) Methods of soil analysis. Part 2. ASA, Madison, WI.

SAS Institute. 1982. Econometric and time-series library. SAS Inst., Cary, NC.

Soil Survey Staff. 1972. Soil survey laboratory methods and procedures for collecting soil samples. USDA Soil Surv. Invest. Rep. no. 1. Revised ed. U.S. Gov. Print. Office, Washington, DC.

Webster, R. 1973. Automatic soil-boundary location from transect data. Math. Geol. 5:27–37.

Webster, R. 1977. Spectral analysis of gilgai soil. Aust. J. Soil Res. 15:191–204.

Webster, R. 1978. Optimally partitioning soil transect. J. Soil Sci. 29:388–402.

Webster, R., and H.E. Cuanalo de Lac. 1975. Soil transect correlograms of North Oxfordshire and their interpretation. J. Soil Sci. 26:176–401.

Webster, R., and I.F.T. Wong. 1969. A numerical procedure for testing soil boundaries interpreted from air photographs. Photogrammetria 24:59–72.