United States
Department of
Agriculture

Forest Service

**Pacific Southwest
Research Station**

P.O. Box 245
Berkeley
California 94701

Research Note
PSW-416
November 1993

Squared Euclidean Distance: A Statistical Test To Evaluate Plant Community Change

Raymond D. Ratliff **Sylvia R. Mori**

Ratliff, Raymond D.; Mori, Sylvia R. 1993. *Squared Euclidean distance: a statistical test to evaluate plant community change.* Res. Note PSW-RN-416. Albany, CA: Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture; 4 p.

The concepts and a procedure for evaluating plant community change using the squared Euclidean distance (SED) resemblance function are described. Analyses are based on the concept that Euclidean distances constitute a sample from a population of distances between sampling units (SUs) for a specific number of times and SUs. With different times, the distances will be intracluster or intercluster. Intercluster distances represent a control treatment. If the communities differ between times, the population will contain clusters (regions of high density with short distances between SUs) that are separated by regions of low density (great distances between SUs). Within- and between-years mean squares from analyses of variance for each species (sp) in the data matrix can be used to compute the intracluster and intercluster mean distances. A multivariate ANOVA gives a test of the hypothesis that the intracluster mean distance is equal to the intercluster mean distance at an overall error rate approximately equal to α. The statistical distribution of SEDs is the distribution of a linear combination of independent Chi-square random variables. Knowledge of this distribution and use of the usual approximations make the estimation of approximate confidence intervals for the mean intercluster and intracluster distances and their difference possible and reliable. The confidence intervals may be examined and a decision made regarding any indicated change in the community. A simple example is provided to permit study of computational methods.

Retrieval Terms: clusters, density, distance, multivariate techniques

Detecting plant community change is an essential component of resource management. Early detection is essential to prevent regression, change away from those attributes desired by the manager. Desired attributes may represent the climax or potential natural plant communities or a seral community critical to survival of a rare or endangered animal species. Managers face the problem of deciding whether there is statistically significant change in a plant community as a whole. Over time, changes in individual species or other attributes or combinations of them may produce conflicting signals, some indicating progression and some indicating regression, and thereby contribute to the problem.

The task of sorting through various signals and integrating information to make decisions (formerly the domain of experienced observers) can be minimized by multivariate techniques used in quantitative plant ecology. Today's managers have the tools to classify and ordinate species, quadrats, and communities; compare different classifications; and isolate the factor or factors which best reflect and account for differences among community types. Multivariate techniques are well documented, researched, and computerized.[1,2] But when is change significant?

Until recently most multivariate methods assumed a data structure that did not incorporate time.[3] As a result, they did not normally serve well as decision-making tools for those monitoring long-term ecological change.

Evolution of the technology, therefore, continues with most of the effort directed to ordination techniques.[4,5] Less effort to incorporate time appears directed to techniques of classification. Nevertheless, methods using discriminant analysis with classification have been proposed.[6,7,8] Also, multivariate analysis of variance has been used to assess change in species composition over time.[9]

This note presents the concepts for detecting plant community change over time, using squared Euclidean distance (SED) as the resemblance function, a mathematical formula for measuring the dissimilarity of objects.[1,3] The concepts derive directly from a procedure for deciding final clusters of a dendrogram.[10] That procedure analyzed distances between the objects classified in order to decide whether two or only one cluster was actually present.

BASICS

Individual objects of classification often are quadrats or transects from which data on species abundance or other characteristics are obtained. As used here, they correspond to arbitrary sampling units (SUs).[1]

Clustering strategies define how the SUs are to be sorted, based on their resemblances (hereafter called distances). Considering plant community change, the strategy used to cluster SUs is simply their sampling periods (season or year).

In order to have statistical validity for detecting change, using the squared Euclidean distance model (equation 1), at least two independent data sets (usually obtained at different times) must be available. Also, the SUs need to be randomly selected each time. Because each SU can belong to one sampling period only, their classification is exclusive and extrinsic.[11]

In multidimensional space, clusters are regions of high density separated by regions of low density.[12] As related to numbers of SUs, a cluster requires an area relationship; but as related to compactness or variation, a cluster requires closely similar SUs.[10]

Given samples that are exclusive and extrinsic, change over time in a plant community will be indicated by clustering or lack of clustering. When two or more clusters are separated, density in the intercluster (between-group) space is low compared to that of the intracluster (within-group) space. In terms of distances, the intercluster distances (those among SUs of different clusters) are greater than the intracluster distances (those among SUs of the same clusters). When two or more clusters cannot be distinguished, intracluster space is no more dense than intercluster space, and intracluster distances equal intercluster distances. Therefore, we fail to reject the hypothesis of no difference between communities.

ANALYSIS

We know of no statistical packages for using the distance matrix directly to test for community change.

One analysis procedure viewed intracluster (within-group) distances and intercluster (between-group) distances as responses to a two-level treatment (within-group and between-group levels, respectively) in the completely randomized design analysis of variance.[10] Rightly, the authors of that procedure cautioned against making statements of statistical significance because the assumption of independence of the responses required by that statistical model is violated.

No two distances are computed using the same two sets of data, and the calculation of one distance does not require that other distances be calculated first. Nevertheless, some distances are not independent from each other because some of them share the same SU. For example: in year 1 the distance between SU_1 and SU_2 and the distance between SU_1 and SU_3 are not independent random variables; they have SU_1 in common. But the first distance and the distance between SU_3 of year 1 and any one SU of year 2 are independent random variables.

Definition of Squared Euclidean Distance (SED)

Let $w_{i,j,k}$ be the measured characteristic of species k in SU i of year j, where $i = 1,2,...,n$, $j = 1,2,...,y$ and $k = 1,2,..., sp$. The $w_{i,j,k}$'s of species k are assumed normal with equal variance that is the sum of two variance components, one measuring the within-year variability and the other measuring the between-year variability. Then the squared Euclidean distance is defined as follows:

$$\mathrm{SED}_{i_1,i_2}^{j_1,j_2} = \Sigma_{k=1}^{sp} \left(w_{i_1,j_1,k} - w_{i_2,j_2,k}\right)^2. \quad (1)$$

If $j_1 = j_2$, we have an Intracluster SED, and if $j_1 \neq j_2$, we have an Intercluster SED.

SED, like standard Euclidean distance, does not measure on a relative scale, and spurious SU classifications may occur as a result.[1,3] Nevertheless, using it allows us to avoid the problem of the lack of independence. Also, the statistical distributions of most distance random variables are unknown, with the result that confidence intervals with known probabilities cannot be computed. If the species' measurable characteristics (such as biomass, cover, abundance, etc.) are normally distributed, then the statistical distribution of SEDs is the distribution of a linear combination of independent Chi-square random variables that can be approximated to a Gamma distributed random variable[13] or to a Normal random variable through the Central Limit Theorem[14] or estimated by computer routines.[15] Approximation to known distributions makes it possible to compute confidence intervals for the mean intercluster and intracluster distance and for their difference. A level of significance can be stated approximately or almost exactly.

Distances between SUs are not all independent, but the measured characteristics of individual species (cover percentages, for example) are independent random variables. With the multivariate analysis of variance (MANOVA) approach, one can test whether the between-year variance component vector (vector of the species' variance components) is zero.

Using MANOVA amounts to testing the hypothesis that the intracluster mean SED is equal to the intercluster mean SED because the null hypothesis that the year variance component is zero for each species is equivalent to the null hypothesis that the sum of the year variance components is zero. But twice this sum is the difference between the intracluster and intercluster means.

MANOVA generally outputs the univariate analysis of variance (ANOVA) for each species. The ANOVA table gives the values of the between-year and within-year mean squares, $\mathrm{MS}_{B,k}$ and $\mathrm{MS}_{W,k}$ for each species k, $k = 1,2,..., sp$. Using these mean square values eliminates the need to compute each SED individually. The intracluster mean

$$\mathrm{SED} = 2\Sigma_{k=1}^{sp}\mathrm{MS}_{W,k}, \text{ and} \quad (2)$$

intercluster mean

$$\mathrm{SED} = 2\{\Sigma_{k=1}^{sp}\mathrm{MS}_{B,k} + (n\text{-}1)\Sigma_{k=1}^{sp}\mathrm{MS}_{W,k}\}/\mathrm{n}. \quad 3)$$

Also, the difference mean SED =

$$2\{\Sigma_{k=1}^{sp}\mathrm{MS}_{B,k} - \Sigma_{k=1}^{sp}\mathrm{MS}_{W,k}\}/\mathrm{n}. \quad (4)$$

EXAMPLE WITH CALCULATIONS

In order to demonstrate the procedure, we use a simulated data set comprised of three species on three SUs at two times *(table 1)*.

Mean Distances

Consider n SUs used to randomly sample a site in each of y years. Where N = yn, there will be N!/2!(N-2)! distances ("md") in multidimensional space. Alternately, the number of distances would be md = $yn(yn-1)/2$. For example, let n = 3 and y = 2, then md = 15 *(table 2)*. A total of six (three from each cluster) are intracluster distances. The remaining nine are intercluster distances.

Even for small size data sets, the number of SEDs to be calculated is large. For example, with 10 SUs and $y = 2$, md = 190. Therefore, calculating and manipulating the distance matrix to obtain the intracluster and intercluster means can become cumbersome, but fortunately we do not have to calculate the distances and means in this way.

The means can be calculated from the within- and between-group mean squares given by analyses of variance *(table 3)*. The sample means of the SEDs are:

intracluster SED = 2(4.279) = 8.56 (equation 2),

intercluster SED = 2(83.213 + 2[4.279])/3 = 61.18 (equation 3), and

difference SED = 2(83.213 − 4.279)/3 = 52.62 (equation 4).

Hypothesis Testing

If $\alpha = 0.10$ and $sp = 3$, the hypothesis that the intracluster and intercluster mean SEDs are equal would be rejected with several approximate multivariate tests (such as Wilks' Lambda F-statistic, Pillai's Trace F-statistic, etc.) that give p-values less than 0.10.[16] For the example we reject the hypothesis and declare that the communities at time 1 and time 2 are different. The difference is reflected by the individual ANOVA's for species 1 and 2 *(table 3)*.

Confidence Intervals

The test of the null hypothesis does not tell how precise the estimations of the mean intracluster and intercluster SEDs are, but the confidence intervals for the mean intracluster and intercluster distances and for their difference provide the needed precision. Confidence intervals provide information about the size and precision of the estimated means; by studying the confidence intervals, whether there is an important difference among communities may be decided. To compute these confidence intervals it is necessary to know the statistical distribution of the SED random variables. If the intracluster and intercluster means are known, then using the fact that the SEDs are sums of multiples of 1-degree-of-freedom Chi-square random variables, one could obtain the statistical distribution of these SEDs via computer algorithms. Also, estimates of these distributions could be generated from the sample SED means.

Table 4 illustrates the above statements by the use of the simulated data of *table 1*.

Table 4 shows the estimated means, their estimated standard errors (SE), and approximate 90 percent confidence intervals. The confidence intervals for the intra- and intercluster means were calculated from the estimated true distributions[15] and from the Gamma approximation.[13] The confidence intervals for the mean difference were calculated from the estimated true distribution and from the Normal approximation.[14] The Normal approximation should be used if n (number of plots/year) is large; the one shown on *table 4* should be taken as a reference only because $n = 3$ is too small.

Note that (using the estimated true distribution) the confidence interval on *table 4* for the difference does not include zero. This is usually interpreted as a significant difference at approximately the 0.10 α-level between the intracluster and

Table 1—***Data[1] for three species on three sampling units (SU) at two times***

	Year 1			Year 2		
	Sampling unit			Sampling unit		
Species	1	2	3	4	5	6
1	5.75	7.79	6.63	2.02	1.10	2.93
2	4.79	3.21	4.44	7.72	8.31	9.49
3	5.38	2.97	5.57	9.99	6.53	8.75

[1]Values are weights (gm/0.09 m^2), simulated for purpose of example.

Table 2—***Squared Euclidean distances (SEDs) for data in* table 1**

SU[1] pair		SEDs for individual species: 1		2		3		Sum
1 & 2	=	$(5.75-7.79)^2$	+	$(4.79-3.21)^2$	+	$(5.38-2.97)^2$	=	12.466
1 & 3	=	$(5.75-6.63)^2$	+	$(4.79-4.44)^2$	+	$(5.38-5.57)^2$	=	0.933
2 & 3	=	$(7.79-6.63)^2$	+	$(3.21-4.44)^2$	+	$(2.97-5.57)^2$	=	9.619
4 & 5	=	$(2.02-1.10)^2$	+	$(7.72-8.31)^2$	+	$(9.99-6.53)^2$	=	13.166
4 & 6	=	$(2.02-2.93)^2$	+	$(7.72-9.49)^2$	+	$(9.99-8.75)^2$	=	5.499
5 & 6	=	$(1.10-2.93)^2$	+	$(8.31-9.49)^2$	+	$(6.53-8.75)^2$	=	9.670
Intracluster mean	=							**8.558**
1 & 4	=	$(5.75-2.02)^2$	+	$(4.79-7.72)^2$	+	$(5.38-9.99)^2$	=	43.750
1 & 5	=	$(5.75-1.10)^2$	+	$(4.79-8.31)^2$	+	$(5.38-6.53)^2$	=	35.335
1 & 6	=	$(5.75-2.93)^2$	+	$(4.79-9.49)^2$	+	$(5.38-8.75)^2$	=	41.399
2 & 4	=	$(7.79-2.02)^2$	+	$(3.21-7.72)^2$	+	$(2.97-9.99)^2$	=	102.913
2 & 5	=	$(7.79-1.10)^2$	+	$(3.21-8.31)^2$	+	$(2.97-6.53)^2$	=	83.440
2 & 6	=	$(7.79-2.93)^2$	+	$(3.21-9.49)^2$	+	$(2.97-8.75)^2$	=	96.466
3 & 4	=	$(6.63-2.02)^2$	+	$(4.44-7.72)^2$	+	$(5.57-9.99)^2$	=	51.547
3 & 5	=	$(6.63-1.10)^2$	+	$(4.44-8.31)^2$	+	$(5.57-6.53)^2$	=	46.479
3 & 6	=	$(6.63-2.93)^2$	+	$(4.44-9.49)^2$	+	$(5.57-8.75)^2$	=	49.305
Intercluster mean	=							**61.181**

[1]Sampling Unit

Table 3—***Analysis of variance for individual species using data in* table 1**

	Species			
Values from ANOVA	1	2	3	Total
Between-year mean square	33.229	28.514	21.470	83.213
Within-year mean square	0.942	0.750	2.587	4.279
F-statistic	35.275	38.019	8.300	
P-value	0.004	0.004	0.045	

Table 4—*Means and standard error estimates (SE) and approximate 90 percent confidence intervals for the intracluster, intercluster squared Euclidean distances (SED) and their difference based on data in table 1*

SED	Estimated		Approximate central 90 pct confidence intervals		
	Mean[1]	SE	Estimated true distribution[2]	Gamma approximation[3]	Normal approximation[4]
intra	8.56	4.06	[3.3 , 16.1]	[3.1 , 16.1]	
inter	61.18	46.06	[9.4 , 140.0]	[9.2 , 151.1]	
diff.	52.62	46.00	[1.3 , 138.1]		[-37.5 , 142.8]

[1] Calculated using equations 2, 3, and 4.
[2] After Wolfram (1988).[15]
[3] After Satterthwaite (1946).[13]
[4] After Feller (1950).[14]

the intercluster distances; we reject the hypothesis that the intracluster mean distance is equal to the intercluster mean distance.

It is possible that the hypothesis that the variance component due to year is zero could be rejected for one species at α/sp-level without rejecting the hypothesis of equality of the intracluster and intercluster mean distances using the confidence intervals. The explanation of this apparent contradiction lies in the increased variation resulting from adding the variation of non-significant differences of the other species; the variance of a sum of independent variables is the sum of their variances. Even though one species could have changed significantly, change of the community as a whole could not be sufficient to indicate a difference.

Such a result may indicate season-to-season or year-to-year fluctuations, especially if species amounts increase and then decrease or vice versa. When true change is occurring, the hypothesis should be rejected at a later observation by analyses of both individual species and the confidence interval for the difference of the intracluster and intercluster means.

ADVANTAGES / DISADVANTAGES

Formal comparisons with other procedures have not been made, and other procedures may serve well for detecting plant community change. For example, procedures such as CANOCO[4] and multiple response permutation[2] can be used by including "time" effects as an ecological vector. They are usually complex, however, and not based on common statistical methods. Good understanding of their limitations at each step in the analyses and careful interpretation of their outputs are required.

A disadvantage of our approach is that the statistics are valid only under the assumption that the distribution of each species response is normal, and because year is assumed to be a random effect, the variance of the response encompasses the within-year and between-year variation. Nevertheless, our approach uses time of observation as the clustering strategy and asks whether SEDs between SUs of the same and different times support the hypothesis of no change in the plant community. Common statistical methods are used and facilitate interpretation of results. Moreover, this approach gives a confidence interval for the expected change.

Use of the confidence interval for the difference of the intracluster and intercluster means with analyses of species provides for early detection of community change. Despite the tools available for detecting change in communities, it is still the manager's job to decide whether a statistically significant change is biologically significant and whether action to fine-tune management should be taken.

END NOTES AND REFERENCES

[1]Ludwig, John A.; Reynolds, James F. 1988. *Statistical ecology*. New York: John Wiley and Sons; 337 p.

[2]Biondini, Mario E.; Mielke, Paul W., Jr.; Redente, Edward F. 1988. *Permutation techniques on Euclidean analysis spaces: a new and powerful statistical method for ecological research*. Reprinted in the Handbook of Vegetation Science 11, Kluwer Academic Publishers, Dordrecht, The Netherlands. 1991; 221-240.

[3]Orloci, Laszlo. 1978. *Multivariate analysis in vegetation research*. 2nd ed. The Hague: Junk; 451 p.

[4]Ter Braak, Cajo J. F. 1986. *Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis*. Ecology 67:1167-1179.

[5]Foran, B. D.; Bastin, G.; Shaw, K. A. 1986. *Range assessment and monitoring in arid lands: the use of classification and ordination in range survey*. Journal of Environmental Management 22:67-84.

[6]Jameson, Donald A.; Gierisch, Ralph K.; Wallace, Stanton; Robinson, Robert. 1970. *Range condition evaluation by discriminant analysis*. Science Series No. 8. Fort Collins, CO: Range Science Department, Colorado State University; 26 p.

[7]Mosley, Jeffrey C.; Bunting, Stephen C.; Hironaka, M. 1986. *Determining range condition from frequency data in mountain meadows of central Idaho*. Journal of Range Management 39:561-565.

[8]Ratliff, Raymond D.; Westfall, Stanley E. 1989. *Monitoring plant community change: an application of quadrat classification + discriminant analysis*. Vegetatio 80:1-9.

[9]Stroup, Walter W.; Stubbendieck, J. 1983. *Multivariate statistical methods to determine changes in botanical composition*. Journal of Range Management 36:208-212.

[10]Ratliff, Raymond D.; Pieper, Rex D. 1981. *Deciding final clusters: an approach using intra- and intercluster distances*. Vegetatio 48:83-86.

[11]Williams, W. T. 1971. *Principles of clustering*. Annual Review of Ecology and Systematics 2:303-326.

[12]Hartigan, John A. 1975. *Clustering algorithms*. New York: John Wiley and Sons; 351 p.

[13]Satterthwaite, Franklin E. 1946. *An approximate distribution of estimates of variance components*. Biometrics Bulletin 2:110-114.

[14]Feller, William. 1950. *An introduction to probability theory and its applications*. New York: John Wiley and Sons; 461 p.

[15]Wolfram, Stephen. 1988. *Mathematica: a system for doing mathematics by computer*. New York: Addison-Wesley, Inc.; 749 p.

[16]Wilkinson, Leland. 1989. *SYSTAT: The system for statistics*. Evanston, IL: SYSTAT, Inc.; 822 p.

THE AUTHORS:

Raymond D. Ratliff is a range scientist assigned to the Pacific Southwest Research Station's Wildlife Monitoring and Range, Research Unit, headquartered at 2081 East Sierra Avenue, Fresno, CA 93710. **Sylvia R. Mori** is a mathematical statistician assigned to the Station's Statistics and Computer Services Group, headquartered at 800 Buchanan St., Albany, CA 94710.

The Forest Service, U.S. Department of Agriculture, is responsible for Federal leadership in forestry. It carries out this role through four main activities:

- Protection and management of resources on 191 million acres of National Forest System lands
- Cooperation with State and local governments, forest industries, and private landowners to help protect and manage non-Federal forest and associated range and watershed lands
- Participation with other agencies in human resource and community assistance programs to improve living conditions in rural areas
- Research on all aspects of forestry, rangeland management, and forest resources utilization.

The Pacific Southwest Research Station

- Represents the research branch of the Forest Service in California, Hawaii, American Samoa and the western Pacific.

Federal Recycling Program
Printed on Recycled Paper