LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

no. 66 - 99

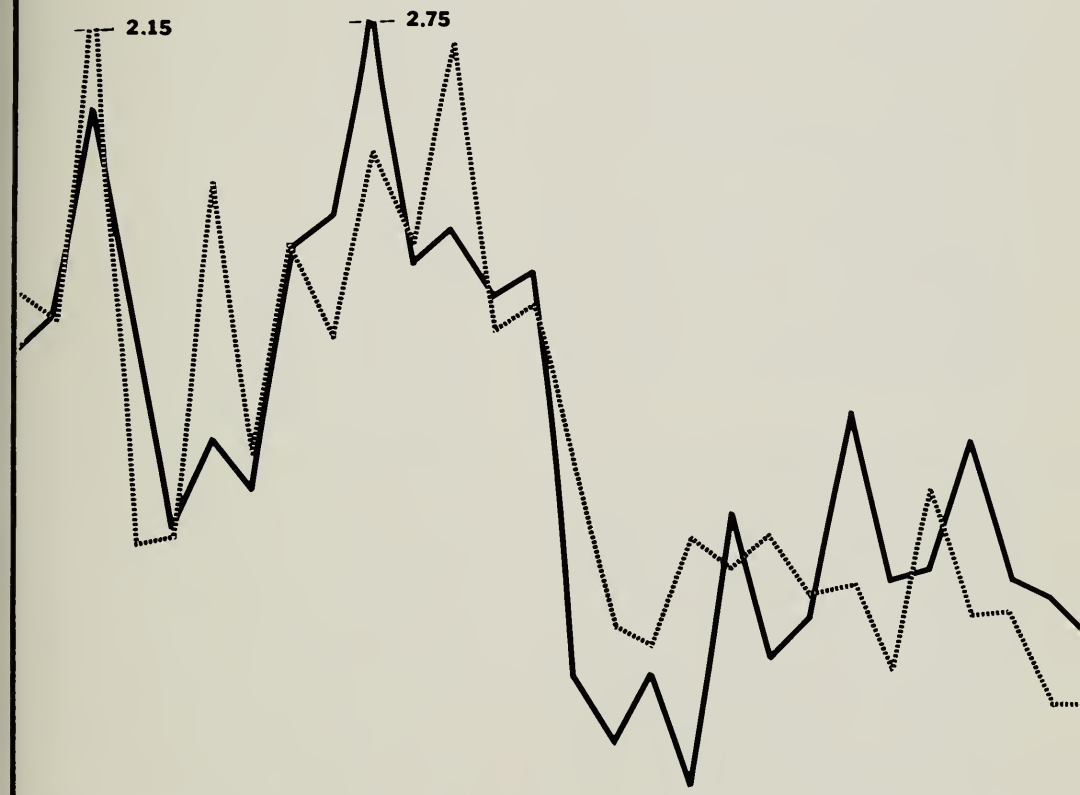
NATURAL HISTORY
SURVEY

(25338—5M—1-71)

STATE OF ILLINOIS
DEPARTMENT OF REGISTRATION AND EDUCATION
NATURAL HISTORY SURVEY DIVISION

The Use of Factor Analysis in Modeling Natural Communities of Plants and Animals

Robert W. Poole



Illinois Natural History Survey
Biological Notes No. 72
Urbana, Illinois February, 1971

The Use of Factor Analysis in Modeling Natural Communities of Plants and Animals

Robert W. Poole

THE PROBLEM OF MODELING COMMUNITIES of plants or animals can be studied either by observing the characteristics of the community as a whole or by determining the interactions among and within individual species. At the community level most attention has been focused on descriptive community analysis, species diversity, and energy flow. At the one- and two-species levels some aspects of the problem that have been, and are being, intensively studied are population demography, predation, competition, parasitism, and spatial distribution. These basic interactions have been reasonably well described, and they have been integrated in the modeling of spruce budworm populations in Canada (Morris 1963). However, even this one-species model is very complex and requires the determination of a large number of parameters.

It is just as conceivable to go from the community level to the individual species as from the species to the community. The purpose of this paper is to explore this approach using a statistical technique known as factor analysis. Factor analysis is a statistical technique for picking out the underlying factors causing the variance in a set of variables.

Factor analysis originated in the psychological sciences but is now also being used in the biological sciences. Its first uses in biology were by Goodall (1954) and Sokal & Hunter (1955), and it has since been used extensively in numerical taxonomy (e.g. Sokal & Sneath 1963; Schnell 1970) in the delimiting of the natural associations of plants (e.g. Dagnelie 1965) and in palaeoecology by Reyment (1963).

Factor analysis, primarily the form known as principal components analysis, has been used in biology for the most part as a classification technique, although there have been some attempts to make associations between environmental variables and species using correlation coefficients in a factor-analysis framework (e.g. Dagnelie 1965). Factor analysis was originally developed to estimate and define the factors causing the observed responses in a series of variables and is here used in this sense rather than as a classification technique.

This paper is divided into three parts. The first gives a brief review of basic ecological principles necessary for the following two sections. The second section describes the statistical procedures considered and the analysis of a specific example. The third section considers the assumptions of the factor analytic model and compares them to the initial ecological generalities to see if the model really does mirror the workings of the community or if it only produces a set of mathematically correct but ecologically meaningless numbers. I have tried to emphasize the implications of the assumptions underlying the factor analysis and deemphasize the mathematics. Many university computing centers have the programs used in this paper, and interested persons can find the mathematics underlying the technique in Harman (1967).

I wish to express my appreciation to those persons who have either read the manuscript or helped with the analysis of the example used in this paper: Dr. George F. Kawash of the Department of Psychology, University of Illinois; Dr. K. W. Dickman of the SOUPAC office of the Department of Computer Science, University of Illinois; Dr. Robert H. Whittaker of the Division of Biological Sciences, Cornell University; Dr. Richard B. Root of the Department of Entomolgy, Cornell University; Mrs. Kathleen Eickwort of the Department of Entomology, Cornell University; and my wife Beverly. I also wish to thank Dr. Philip W. Smith and O. F. Glissendorf of the Illinois Natural History Survey for their editorial contributions to the paper.

FACTORS AND SPECIES POPULATIONS

A population of an animal rarely stays at a constant level; usually it is either increasing or decreasing. Whether and how much a population increases or decreases depends on the environmental factors controlling the limits of that population. If conditions are favorable, the population increases; if they are not favorable, it decreases. A species population can be affected by several factors, and the factors may be interacting among themselves. This basic relationship is diagrammed in Fig. 1. Not all of the factors are of equal importance to the species population, one factor usually being more important than the others. If the effect of a factor on a population depends on the density of the population, it is referred to as a density-dependent factor, and if it does not depend on density, it is referred to as a density-independent factor.

In a community of two or more species, a factor influencing one species may also influence other species in the community. The effect of this common factor may vary from species to species, being more important

This paper is published by authority of the State of Illinois, IRS Ch. 127, Par. 58.12. Dr. Robert W. Poole is Assistant Taxonomist, Section of Faunistic Surveys and Insect Identification, Illinois Natural History Survey, Urbana.

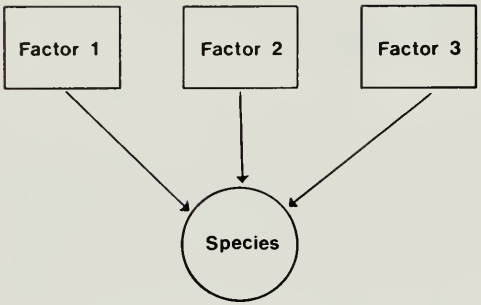


Fig. 1.—Diagrammatic representation of the influence of three factors on one species.

for one than for another. At the same time a species may be influenced by a factor or set of factors which affect it only. These will be referred to as specific factors. This relationship is diagrammed in Fig. 2. Even in this relatively simple community with uncorrelated factors, the complexity is evident.

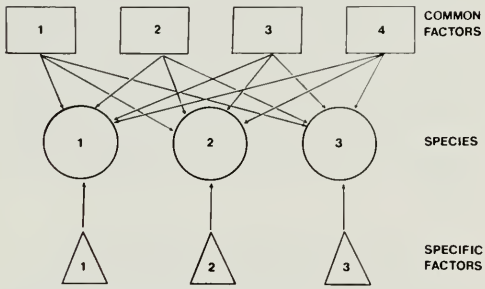


Fig. 2.—Diagrammatic representation of the influence of three specific and four common factors on three species.

If two species share a common factor or factors, the changes in their populations will be correlated. For example, if two species are both limited by rainfall and rainfall is increased, both populations will increase. However, if one species is only slightly dependent on rainfall and the other strongly so, the changes will be disproportionate and the correlation less. As a simple principal rule, the correlations among a group of species making up a community are determined by the species' mutual association with a group of common factors.

In essence, factor analysis takes a matrix of correlation coefficients among a set of variables and reduces it to a series of mathematical common factors that account for the correlations among the variables.

TECHNIQUES

The procedures carried out in this paper are calculation of the correlation matrix, estimation of communalities (the amount of variance caused by factors common to other species), factoring of the matrix using the principal axis method, rotation to a specified hypothesis (transformation of the numbers to other biologically meaningful numbers), calculation of factor scores, and the formulation of the so-called specification equations for each species to serve as a model of the community. If all of the above steps except the factoring of the calculated correlation matrix are skipped, the result is a form of factor analysis normally referred to as principal components analysis.

Principal Components Analysis

Mathematically, factor analysis resolves a correlation matrix (a covariance matrix can also be used in some cases) into a $n \times k$ factor matrix where the number of factors, k, is usually smaller than n, the number of variables (in this case species). This factor matrix has the characteristic that when multiplied by its transpose (rows and columns interchanged) it restores the original correlation matrix. In matrix notation

$$R = V_0 V_0{}'$$

where R is the correlation matrix, V_0 the factor matrix, and $V_0{}'$ the transposed factor matrix. Basically the problem is to resolve the correlation matrix into its latent roots and vectors (also referred to as eigenvalues and eigenvectors).

Principal components analysis assumes that all of the variance of each species can be accounted for by a set of factors common to all of the other species in the community and lumps variance due to specific factors and error factors in with the common factors. In the actual computation, the loadings (weights) of the first factor on each species (the latent vector or eigenvector) are calculated in such a way as to remove the maximum amount of variance from the matrix as can be explained by one factor. The effects of this factor are then subtracted from the correlation matrix. A second factor is then calculated from this reduced correlation matrix, and so forth until the reduced correlation matrix consists of essentially all zeros.

These calculations have been carried out on data given by Hunter (1966). Hunter measured the species populations of *Drosophila* at several sites in Colombia, principally near Bogotá. I have analyzed her data for "Pine Woods," a government-protected pine forest near Bogotá. The census was carried out from September, 1961 to December, 1963 (28 months) by sweeping a net over bait. In Hunter's table for Pine Woods, the figures for each month are lumped, and the abundance of each species expressed as a percentage of the total *Drosophila* community. In cases where a species frequency was less than 1 percent, it is listed only as present. In my analysis when a species is listed as "present," I have considered it to be absent because individuals of that species made up less than 1 percent of the total. Of the 11 species listed by Hunter, I have analyzed only 10 because the 11th, "*dreyfusi* 22," was very rare.

TABLE 1. — Correlations among the frequencies of 10 species of *Drosophila* over a period of 28 months at Pines Woods (near Bogotá, Colombia).

	melanogaster	*pseudoobscura*	*bandeirantorum*	*"tripunctata 20"*	*hydei*	*immigrans*	*viracochi*	*mesophragmatica*	*brncici*	*gasici*
melanogaster	1.00									
pseudoobscura	.37	1.00								
bandeirantorum	.34	.85	1.00							
"tripunctata 20"	−.01	.46	.52	1.00						
hydei	−.08	−.09	.02	−.00	1.00					
immigrans	.09	.24	.31	.57	−.18	1.00				
viracochi	.11	−.13	−.00	−.32	.38	−.02	1.00			
mesophragmatica	−.42	−.79	−.80	−.55	−.09	−.56	−.21	1.00		
brncici	.43	.45	.61	.28	−.09	.05	−.15	−.41	1.00	
gasici	−.13	−.20	−.18	.37	−.01	.58	.01	−.26	−.25	1.00

The correlation matrix was calculated (Table 1) using the Pearson product-moment correlation coefficient and factored using the principal axis method. The resulting factors are shown in Table 2, which also shows

TABLE 2. — Calculated factors from the principal components analysis.

Factor	Variance	Percent Variance	Cumulative Percent
1	3.7921	37.9214	37.9214
2	1.9637	19.6365	57.5579
3	1.5131	15.1310	72.6889
4	0.9162	9.1621	81.8510
5	0.7103	7.1034	88.8544
6	0.5365	5.3646	94.3189
7	0.2941	2.9414	97.2603
8	0.1728	1.7282	98.9885
9	0.0971	0.9712	99.9596
10	0.0040	0.0403	100.0000

that the first three factors account for about 73 percent of the variance in the correlation matrix. The total number of factors extracted by the principal axis method cannot exceed the number of variables. Each of the calculated factors is affected in part by the inclusion of error variance and variance due to specific rather than common factors. Therefore the factors become more and more trivial and unreliable as the factoring proceeds, so that the factors calculated after the first few have no real meaning. A commonly used breaking point in factoring is when the eigenvalue of the factor falls below 1.00 (listed under variance in Table 2). Using this criterion, the first three factors are significant. The factor loadings of each factor on the 10 species are given in Table 3. Factor loadings are a type of correlation between a factor and a variable, or more specifically, the weight of each factor in accounting for the variance of a given variable. In other words if factor 1 had a loading of .47 on a given variable,

TABLE 3. — Computed factor loadings from the principal components analysis on the 10 species of *Drosophila*.

	Factor 1	Factor 2	Factor 3
melanogaster	−.4750	−.3843	−.0687
pseudoobscura	−.8581	−.2522	.0020
bandeirantorum	−.9002	−.2406	.1279
"tripunctata 20"	−.6886	.4540	−.1654
hydei	.0789	−.0303	.7648
immigrans	−.5498	.6897	−.0449
viracochi	.1171	.0108	.8662
mesophragmatica	.9091	−.1325	−.3291
brncici	−.6233	−.4481	−.1364
gasici	−.0914	.8906	.0226

and factor 2 a loading of .03, factor 1 would be more important to the variable than would factor 2.

Rotation

The set of factors arrived at in the preceding section and the loadings of the factors on the variables are only one of an essentially infinite number of possibilities. In other words there is an infinite number of factor matrices that when multiplied by their transpose will restore the original correlation matrix. The factors as they come out of the principal axis method are orthogonal to each other (uncorrelated). These calculated factors do not necessarily correspond in any way with the real attributes of the environment controlling the fluctuations of the species populations. One of this infinite array of answers is the correct one, however, and the problem is to find it. The variables can be plotted on each factor as has been done in Fig. 3 for factors 1 and 2. Factor 3 could be included and the variables would then be in a three-dimensional space. The addition of a fourth factor would be in hyperspace. Any of the possible solutions to the problem can be arrived at by rotating these axes (factors) and reading off the new factor loadings on each variable. This is an oversimplified explanation of rotation and a more complete account can be found in Cattell (1965) and Harman (1967).

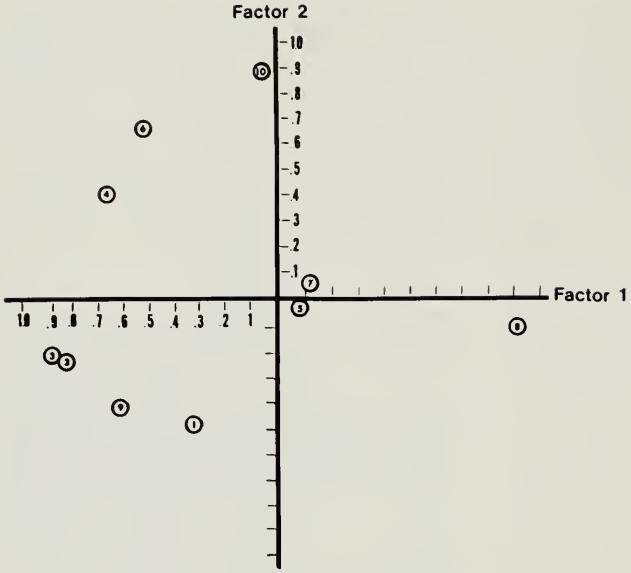


Fig. 3.—Loadings of the 10 species of *Drosophila* on factors 1 and 2.

In the principal axis method the first factor is calculated to account for as much of the variance in the correlation matrix as possible. The method attempts to have this factor loaded as heavily as possible with all of the variables. It is possible that a factor such as temperature would influence all of the species strongly, and in this case the calculated factor as it comes from the analysis would accurately reflect the actual environmental factor. However, it is also possible that a factor may be important to only two or three species and relatively unimportant to the others in a community. In this second case the factors coming from the principal axis analysis would not fit the real situation and must be rotated to a position where they do. The above situation is satisfied by rotation to what is known as simple structure. The factors coming from the principal axis analysis are orthogonal to each other, but very often, probably usually, the factors operating on the species are correlated with each other. By rotating to simple structure, the factors are allowed to be correlated with each other. Mathematically, rotation to simple structure attempts to correlate a factor with the smallest number of variables possible. In other words each factor should affect only a few variables.

In rotation, the original factor matrix (V_0) is multiplied by a transformation matrix (T) giving a new matrix referred to as the reference vector matrix (V_{rs})

$$V_{rs} = V_0 T$$

The reference vector matrix does not give the new loadings of the factors on the variables for reasons discussed by Cattell (1965). To calculate the new factor loadings, a new matrix termed the factor-pattern matrix is calculated as

$$V_{rp} = V_{rs} D^{-1}$$

where D is the diagonal matrix of the reciprocal square roots of the diagonal elements of the inverted matrix of the reference-vector correlations. The reference correlations are computed by multiplying the transformation matrix by its transpose

$$C_{rs} = T' T$$

where C_{rs} is the matrix of correlations between the reference vectors, T the transformation matrix, and T′ the transpose of the transformation matrix.

Several mechanical programs are available for rotation to simple structure. The program Oblimax (Pinzka & Saunders 1954) was found to give the most reasonable answers in this case and has been used in this

Table 4. — New factor loadings (the factor-pattern matrix) after rotation to simple structure using the Oblimax program on the 10 species of *Drosophila*.

	Factor 1	Factor 2	Factor 3
melanogaster	−.5791	.2648	−.0889
pseudoobscura	−.8942	.0453	−.0339
bandeirantorum	−.9589	.0280	.0925
"tripunctata 20"	−.4056	−.6187	−.2004
hydei	−.1262	.0744	.7833
immigrans	−.2098	−.8148	−.0725
viracochi	−.0988	.0462	.8881
mesophragmatica	.8618	.3365	−.2956
brncici	−.7255	.2903	−.1641
gasici	.2786	−.9021	.0152

analysis. The factor-pattern matrix after rotation to simple structure is shown in Table 4. A comparison of Tables 3 and 4 shows few significant changes because of rotation to simple structure, using the Oblimax program (the signs have been changed in factor 2).

The new factors produced by rotation to simple structure are not necessarily orthogonal and may be correlated (oblique). The correlation matrix of these three factors is given in Table 5.

TABLE 5. — Correlations among the factors after rotation to simple structure using the Oblimax program.

	Factor 1	Factor 2	Factor 3
Factor 1	1.0000	−.1875	−.1987
Factor 2	−.1875	1.0000	.0469
Factor 3	−.1987	.0469	1.0000

Communalities

In the principal components analysis 1's are entered in the diagonal of the correlation matrix because the correlation of a variable with itself is 1. In factoring the matrix this presumes that all of the variance of a species can be accounted for by factors common to other species. However, a species is normally affected not only by common factors, but by factors specific to it, and also error factors.

The variance of a species (σ^2_s) is equal to the variance explicable by common factors (σ^2_{hs}) plus the variance of the species due to specific factors (σ^2_{us}) plus an error term (σ^2_{es}),

$$\sigma^2_s = \sigma^2_{hs} + \sigma^2_{us} + \sigma^2_{es}$$

The term σ^2_{hs} is usually referred to as a variable's communality.

To remove the variance of a species due to specific factors and error terms, communalities for each species must be calculated and substituted for the diagonal elements of the correlation matrix. Unfortunately there are many different techniques used to estimate communalities and none of them is "the best." Also, the subject of communalities is a controversial one.

In a practical sense, with large initial matrices the effect of not calculating communalities on the estimates of the factors is minimal and becomes less and less important for larger and larger matrices. The calculated communalities are important, however, in estimating the reliability of the predictive equations presented later.

Communalities in the factor analysis carried out in the following pages were calculated by replacing the diagonal entry for each row by

$$(r^*_{ik})\ (S_i - r^*_{ik})\ /\ (S_k - r^*_{ik})$$

where

r^*_{ik} = maximum absolute r_{ij}

S_i = absolute r_{ij}

S_k = absolute r_{kj}

The calculated communalities are listed in Table 6. Other techniques of communality estimation were tried: (1) replacing the diagonal entry of a variable by the square of the multiple R of each variable with all other variables, and (2) replacing the diagonal entry of a row by the square root of the average r^2 across the row. The estimated communalities using these two methods are also given in Table 6. The Varimax-rotation program (Kaiser 1958) also gives iterative solutions for the communalities. The calculated communalities using this iterative technique are also given in Table 6.

TABLE 6. — Estimated communalities of the 10 species of *Drosophila* using the following methods: 1) $(r^*_{ik})\ (S_i - r^*_{ik})/(S_k - r^*_{ik})$, 2) square of multiple R, 3) square root of average r^2, 4) iterative.

	1	2	3	4
melanogaster	.2919	.8454	.4075	.3780
pseudoobscura	.8338	.9825	.5499	.7999
bandeirantorum	.8638	.9097	.5708	.8846
"tripunctata 20"	.7150	.7567	.4940	.7077
hydei	.2333	.5736	.3484	.5921
immigrans	.8314	.7779	.4669	.7801
viracochi	.6267	.9399	.3671	.7641
mesophragmatica	.9345	.9932	.5795	.9524
brncici	.4282	.6633	.4592	.6079
gasici	.4036	.9522	.4113	.8021

Factor Identification

The purpose of the analysis is to arrive, mathematically, at a set of factors corresponding to the real factors in the environment that cause changes in populations of the species in the community. This problem has been partially discussed under rotation. There it was shown that factors calculated by the principal axis analysis do not necessarily correspond to any real factors. To make these factors useful, the factor vectors must be rotated in hyperspace to a position where they do correspond to real parts of the environment. The problem of identification can be broken into two stages: (1) rotation of computed factors to where they correlate heavily with real factors of the environment, and (2) the identification of the environmental factors. I will discuss the second stage first.

A set of factors has been calculated that explain part of the variation in the population of a species. However, to be useful these factors must correspond to real parts of the environment that can be identified. Basically we want to know that factor 1 is so highly correlated with rainfall that rainfall, for practical purposes, can be taken as factor 1. Often a person knows *a priori,* or suspects, that species "a" is heavily influenced by some factor such as maximum temperature. Therefore if this species has a heavy loading on one of the factors derived from the factor analysis, it is a good indication that this factor is either maximum temperature or is, in some way, closely correlated with maximum temperature. It is also possible, if measurements of maximum

temperature are available, to include maximum temperature in the data matrices as another variable. If maximum temperature as a variable loads heavily with one factor and little with other factors, it is likely that this factor is in some way related to maximum temperature. Determining the identity of every significant factor is not easy and depends on extensive field work. However, factor analysis indicates how many significant factors to look for, and the weightings of these factors on every species in the community. Even if a factor is interpreted incorrectly, as maximum temperature, the use of maximum temperature measurements for that factor may still give correct predictive answers, a procedure not very scientific but pragmatically important. It must be emphasized that the mathematical factors never exactly correspond to the environmental factors, but they may be so heavily loaded on the environmental factors that measurements of the environmental factors can be used as approximations to the mathematical factors.

The other problem in identification is the rotation of the factors derived from the principal axis method analysis to some position where they correspond to real parts of the environment. If the factors are not rotated, the hypothesis is that the factors tend to influence all of the variables; however, if rotated to simple structure, it is assumed that the real factors tend to influence significantly only a few of the variables. In a real situation neither hypothesis may be the correct one. For example, if in a community of insects rainfall was important to all of the species, but at the same time each species was restricted in its choice of food plants, there would be one factor influencing all of the species, and several other factors that influenced only a few variables each. This situation clearly does not fit the hypothesis behind the factors as they come from the principal axis analysis or after rotation of the factors to simple structure.

It is also possible to rotate the factors to fit a specific hypothesis, but because it is not possible to formulate a specific hypothesis for the example used in this paper, this rotation has not been done. The most difficult problem connected with this type of community analysis should now be apparent. To rotate the calculated factors to a position where they represent real factors of the environment, a correct hypothesis of the type of factors involved and the relative numbers of each (such as two factors influencing all of the species and three factors influencing only two or three species) is needed. The problem is what stage in the identification of factors is to be carried out first—the identification of factors or the rotation of the calculated factors to fit actual factors in the environment. Each is partially dependent on the other. As a working technique it should be possible, by extensive field work and experimentation, to formulate a rough hypothesis as to the percentage of significant factors that will influence a limited number of the species. For example, it might be known that rainfall influences a certain number of species, and there is reason to believe that it is important to virtually all the species in the community. On the other hand, it might be known that most species in the community tend to be limited in their selection of food plants. Given four significant factors, a rough hypothesis might be that one factor influences all of the species, and three others influence only a few of them. From the set of calculated factors, the first (the one accounting for the most variance) is likely to be factor 1 of the hypothesis, with the other three factors being fitted to the groups of species that they load most heavily with. The factors could then be rotated to fit the rough hypothesis, and the hypothesis could possibly be reformulated as a result of the rotation.

Every possible rotation of the factor vectors is, of course, an approximation to the real situation. Some of the approximations will be good, others not so good. A question of practical importance is whether or not the answers derived from each rotation are much different from each other. The answer to this question will only come through use of the factor-analysis technique. In the example used in this paper, the differences between the factor loadings of the orthogonal factors and the factors rotated to simple structure are slight. It has usually been found in psychology that the changes in factor loadings by rotation to simple structure are slight (Kawash, personal communication). Simple structure rotation tends to rotate out small error factors and is used more for that reason than for the hypothesis it represents (Cattell 1965). Even if the factors are not correctly rotated, the appoximation may still be close.

Computational Procedure

Having carried out a principal components analysis of the data and having partially explained the problems of rotation and communalities, the complete factor analysis will now be carried out. In the following section the predictive equations are formulated and the possible usefulness of the technique is discussed.

As discussed in the preceding section, a possible aid in the identification of the factors is to place measurements of presumed factors into the correlation matrix as variables and then note if any of the calculated factors load heavily on them. Hunter (1966) gave data for rainfall, mean maximum temperature, and mean minimum temperature for each month of her study. She assumed that rainfall was one of the most important factors, pointing out that its effect probably acted upon the larvae, or perhaps initiated egg laying in the adults. Hunter stated that the average time for development from egg to adult is about 2 months. Because these three environmental measurements are more likely to be important to the larvae that later give rise to the adults than to the adults directly, the three measurements have been entered as variables with the species with a 2-month lead.

The correlation matrix of the 13 variables was com-

TABLE 7.—Correlations among the 10 species of *Drosophila* and 3 environmental variables, with calculated communalities in the diagonal.

	melanogaster	*pseudoobscura*	*bandeirantorum*	*"tripunctata* 20"	*hydei*	*immigrans*	*viracochi*	*mesophragmatica*	*brncici*	*gasici*	min. temperature	max. temperature	rainfall
melanogaster	.35												
pseudoobscura	.37	.90											
bandeirantorum	.35	.85	.80										
"tripunctata 20"	−.01	.46	.52	.63									
hydei	−.08	−.09	.02	−.00	.17								
immigrans	.09	.24	.31	.57	−.18	.71							
viracochi	−.11	−.13	−.00	−.32	.38	−.02	.58						
mesophragmatica	−.42	−.79	−.80	−.56	−.09	−.56	−.21	.97					
brncici	.43	.45	.61	.28	−.09	.05	−.15	−.41	.44				
gasici	−.13	−.20	−.18	.37	−.01	.58	.01	−.26	−.24	.66			
min. temperature	−.13	−.16	−.11	.41	.03	.47	−.11	−.11	−.22	.62	.60		
max. temperature	.21	−.12	−.12	.05	−.09	.29	−.54	.12	.05	.21	.29	.52	
rainfall	−.19	−.28	−.09	.09	−.10	.01	−.21	.29	−.01	−.00	−.06	.05	.07

puted. Rainfall was not significantly correlated with any of the other 12 variables. Communalities were estimated, as described earlier in this paper, and the estimated communalities were entered in the diagonal of the correlation matrix. The new correlation matrix, with all 13 variables and the estimated communalities in the diagonal, is given in Table 7.

This correlation matrix with estimated communalities was then factored by the principal axis method. The eigenvalues of the first three factors were 3.60, 2.28, and 1.38, about the same as in the principal components analysis. The second eigenvalue is slightly higher, and the first and third eigenvalues are slightly smaller, than in the principal components analysis (Table 2). There is, however, a significant drop from the eigenvalue of the third factor (1.38) to the fourth (.61). Using an associated eigenvalue of 1.00 as a criterion of significance, the loadings of the first three factors were computed and are listed in Table 8. Comparison of Tables 8 and 3 shows few major changes in the first two factors but several in factor 3. None of the three environmental

TABLE 8. — Factor loadings of the first three factors on the 10 species of *Drosophila* and 3 environmental variables.

	Factor 1	Factor 2	Factor 3
melanogaster	.4001	−.2122	.2205
pseudoobscura	.8586	−.3278	.0661
bandeirantorum	.8675	−.2927	.0106
"tripunctata 20"	.6489	.3999	.0612
hydei	−.0495	−.0746	−.3590
immigrans	.5451	.6116	−.0763
viracochi	−.0852	−.2172	−.7811
mesophragmatica	−.9466	−.0052	.3344
brncici	.5231	−.3158	.2746
gasici	.1217	.7909	−.2287
min. temperature	.1066	.7497	−.0947
max. temperature	.0181	.4004	.5398
rainfall	−.1726	.1044	.2062

TABLE 9. — Calculated reference-vector structure matrix after rotation to simple structure using the Oblimax program.

	Reference Vector 1	Reference Vector 2	Reference Vector 3
melanogaster	.3869	.1672	.1778
pseudoobscura	.8813	.1214	.0121
bandeirantorum	.8917	.0705	−.0342
"tripunctata 20"	.4832	−.5052	.1608
hydei	.0509	−.0243	−.3676
immigrans	.3563	−.7173	.0703
viracochi	.1471	−.0075	−.8113
mesophragmatica	−.9562	.3252	.3045
brncici	.5194	.2502	.2109
gasici	−.0577	−.8286	−.0488
min. temperature	−.0889	−.7476	.0727
max. temperature	−.2096	−.2173	.6141
rainfall	−.2346	.0042	.2201

TABLE 10. — Calculated factor-pattern matrix after rotation to simple structure using the Oblimax program.

	Factor 1	Factor 2	Factor 3
melanogaster	.3994	.1678	.1841
pseudoobscura	.9097	.1218	.0125
bandeirantorum	.9205	.0708	−.0354
"tripunctata 20"	.4987	−.5068	.1664
hydei	.0525	−.0244	−.3806
immigrans	.3677	−.7196	.0727
viracochi	.1519	−.0075	−.8398
mesophragmatica	−.9870	.3263	.3153
brncici	.5361	.2510	.2183
gasici	−.0596	−.8312	−.0505
min. temperature	−.0918	−.7499	.0753
max. temperature	−.2164	−.2180	.6357
rainfall	−.2422	.0042	.2279

variables is heavily loaded on any of the three factors, although minimum temperature has a moderately heavy loading on factor 2, suggesting that factor 2 may in some way be associated with minimum temperature. Rainfall is lightly loaded on all three factors, and it is thus unlikely that it has any association with the three factors.

The factor matrix was then rotated to simple structure using the Oblimax method, and the resulting reference-vector structure matrix is given in Table 9. The calculated factor-pattern matrix is given in Table 10.

Predictive Equations

The next stage is the formulation of what are known as specification equations. These equations specify the weights to be given to each factor in accounting for the score (observed measurement of some kind) of each variable. The specification equation can be written in a general form as

$$V_{ji} = s_{ji}F_{1i} + s_{j2}F_{2i} + \text{-----}s_{jk}F_{ki} + s_jF_{ji} + s_{je}F_{ei}$$

as given by Cattell (1965). If there are k observations, the score on a variable on one of these observations is equal to the sums of the scores of the factors ($F_{j's}$) influencing the variable as modified by the significance or weight of each factor to the variable (the $s_{j's}$). These factors include a series of common factors, any specific factors there may be, and an error factor. The specification equations will be the basic predictive equations. In the example analyzed in this paper there are 10 species measured at 28 observations, giving a total of 280 specification equations. To formulate the set of equations for all species in the community, it is necessary to calculate first the factor-score matrix (F_p) and secondly the factor-pattern matrix (V_{fp}) which gives the necessary values of the $s_{j's}$.

The factor-score matrix is computed by multiplying the reference-vector structure matrix by the basic diagonal of the original correlation matrix. In computation this step was done by inverting the correlation matrix, multiplying that by the matrix of standard scores for the variables standardized by rows, and multiplying the resulting matrix by the reference-vector structure matrix (V_{rs}) or

$$F_p = V_{rs}\,\delta$$

where F_p is the factor score matrix, V_{rs} the reference-vector structure matrix, and δ the basic diagonal of the correlation matrix. The resulting factor-score matrix for the 28 observations is given in Table 11. The factor scores are the standard scores for the factors calculated for a particular rotation. If the factors have been rotated to where they correspond to real parts of the environment, the factor-score matrix gives estimated standard scores for the environmental factors. If the rotation is not the correct one, the numbers are only numbers that will reproduce the scores on the variables. It is, of course, impossible to use them predictively if they are not real.

Having calculated the factor-score matrix and the factor-pattern matrix, it is now possible to estimate the value of a variable on any observation. As an example, the standard score of *Drosophila pseudoobscura* at observation 4 (December, 1961) equals the sums of the factor scores as weighted by the factor loadings for that period plus specific factor scores, plus an error term. In other words

Drosophila pseudoobscura$_{(4)}$ = (.9097)(.7821) + (.1218)(−.1245) + (.0126)(−.7125) + specific factors$_{(4)}$ + error factors$_{(4)}$

Drosophila pseudoobscura$_{(4)}$ = .6873 + specific factors$_{(4)}$ + error factors$_{(4)}$. All scores are in standard form.

Table 11. — Calculated factor score matrix for the 28 observations from the Oblimax rotation to simple structure

Observation	Factor 1	Factor 2	Factor 3
1	.4840	1.0297	−.5129
2	.5836	.9663	−1.0332
3	1.7794	.3722	−.6943
4	.7821	−.1245	−.7125
5	−.4793	−.0247	.8224
6	−.1264	.3034	2.3076
7	−.3645	.8503	1.6247
8	1.1017	.4221	.7145
9	1.2421	.0689	−.0601
10	3.2278	−1.3368	−1.9288
11	1.1410	−1.0098	−.6190
12	1.1349	.3455	.5760
13	.8159	−.0459	−.4046
14	.9906	−.3255	−.8530
15	−1.4451	.9605	1.8509
16	−1.9348	1.2871	1.7579
17	−1.2463	.6319	.7126
18	−2.1194	1.3567	2.2356
19	−.3433	−.2672	.0467
20	−1.2586	.4976	1.0491
21	−.4868	−3.9399	.4878
22	.4718	−1.9417	−1.0491
23	−.6639	−.4351	−1.1375
24	−.8399	1.2284	−.1013
25	.0412	.0449	−2.7617
26	−.5968	−.8071	−.5184
27	−.8115	−.2905	−.5392
28	−1.0795	.0832	−.8220

Theoretically if the scores for the common factors, the specific factors, and the error factors were known, the predicted scores would exactly fit the actual scores of the variables (species population levels). However, in this case nothing is known of the specific factors and the error factors, and the predictions are based only on the variance attributable to common factors. Where common factors account for a large percentage of the variance of a species, the predictions should be fairly accurate. In a species population influenced to a large extent by specific factors and error factors, the predictions will not be as good. To a certain extent, the reliability of the estimates can be judged from the size of the species population's communality, species with large communalities being more predictable than those with small communalties. This procedure, in essence, pretends that specific and error factors do not exist.

Graphs of the predicted and observed abundances (as standard scores) of the 10 species are given in Fig. 4–13. It is clear that for many of the species, particularly the common ones, predicted and actual values agree quite well, although there are still some deviations. Devia-

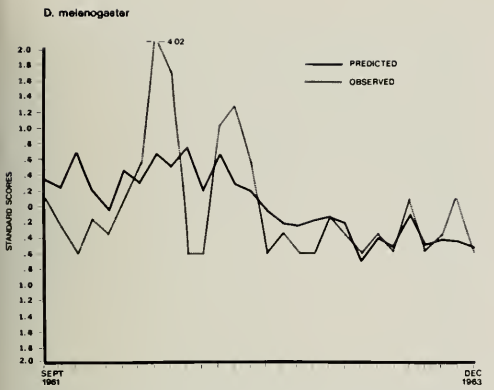


Fig. 4.—Predicted versus observed abundances (standardized) for *Drosophila melanogaster* from September, 1961 to December, 1963.

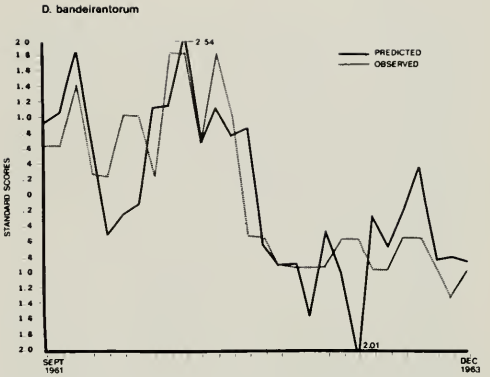


Fig. 7.—Predicted versus observed abundances (standardized) for *Drosophila bandeirantorum* from September, 1961 to December, 1963.

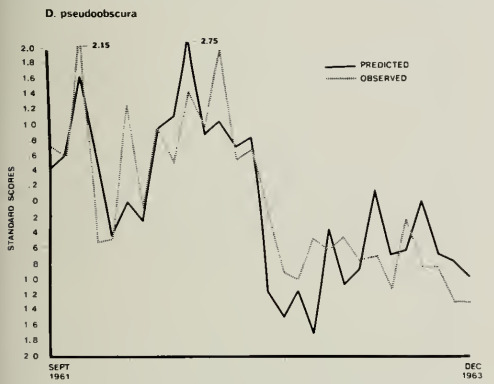


Fig. 5.—Predicted versus observed abundances (standardized) for *Drosophila pseudoobscura* from September, 1961 to December, 1963.

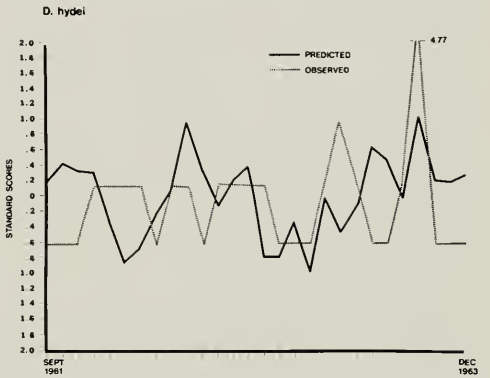


Fig. 8.—Predicted versus observed abundances (standardized) for *Drosophila hydei* from September, 1961 to December, 1963.

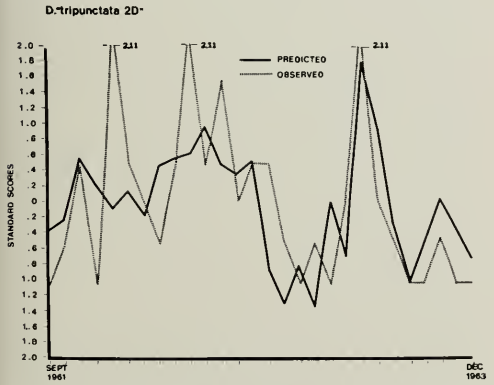


Fig. 6.—Predicted versus observed abundances (standardized) for *Drosophila "tripunctata* 20" from September, 1961 to December, 1963.

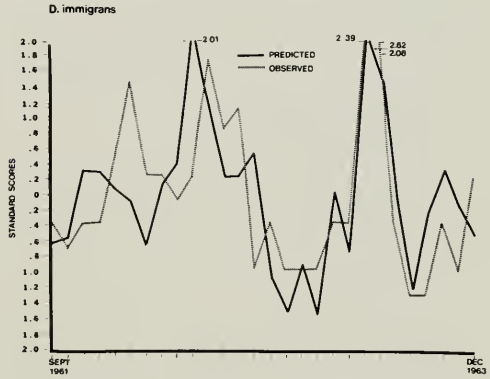


Fig. 9.—Predicted versus observed abundances (standardized) for *Drosophila immigrans* from September, 1961 to December, 1963.

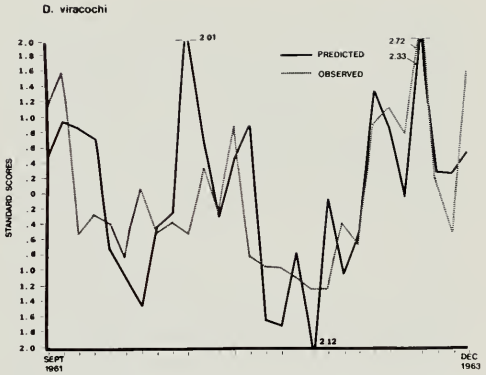


Fig. 10.—Predicted versus observed abundances (standardized) for *Drosophila viracochi* from September, 1961 to December, 1963.

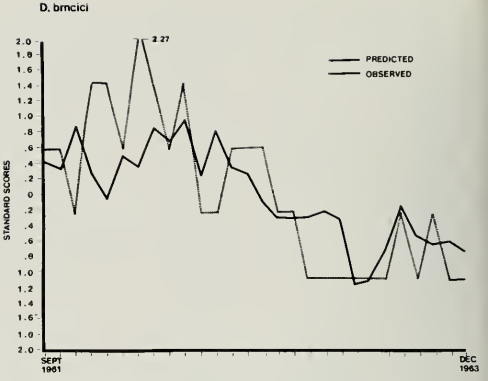


Fig. 12.—Predicted versus observed abundances (standardized) for *Drosophila brncici* from September, 1961 to December, 1963.

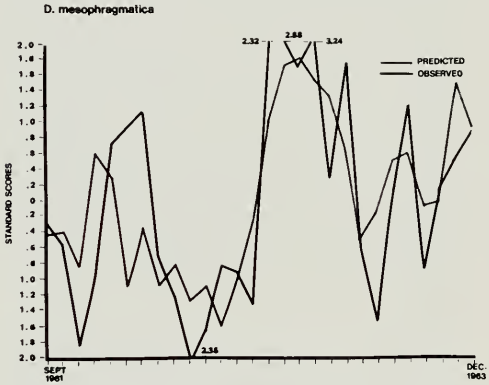


Fig. 11.—Predicted versus observed abundances (standardized) for *Drosophila mesophragmatica* from September, 1961 to December, 1963.

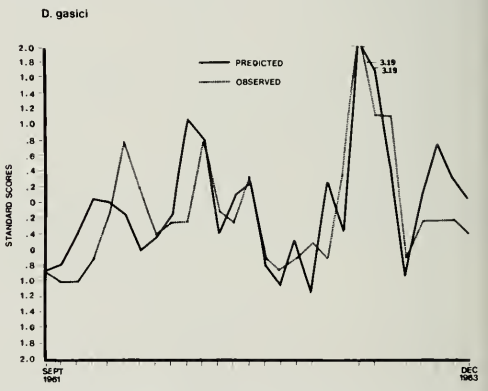


Fig. 13.—Predicted versus observed abundances (standardized) for *Drosophila gasici* from September, 1961 to December, 1963.

tions of the predicted from the actual values (with the predicted values converted to raw scores), as measured by a Chi-square goodness-of-fit test for *Drosophila pseudoobscura* and *mesophragmatica*, are highly significant; the fit is hardly perfect.

Some general observations are:

1. Common species are better modeled than rarer ones.
2. Large, long-term changes can usually be predicted, but short-term, small fluctuations cannot, particularly for the rarer species.
3. The last 14 months present a better fit than the first 14 months.

In a practical sense the predicted population levels from the above analysis are rather trivial because values of a set of common factors were calculated from 28 observations on 10 variables (species); using these calculated factor scores, the population levels of the species for each period were recalculated. However, if the factors influencing the species of a given community have been identified by a previous factor analysis and the rotation properly carried out, it is possible later to make measurements of the factors, standardize them, and then calculate predicted standard scores for all of the species of the community using a set of specification equations as above. I have not been able to do so with these data from the literature, because the factors have not been identified, or if they had been, no measurements are available for them. Also, there is no way to check the validity of the results.

The use of these predictive equations can be illustrated by a possible application. A factor-analysis study carried out on the community of fish in a river had determined that water temperature was one of the important

factors affecting the fish community. It is also known that the establishment of a nuclear reactor on the banks of the river will progressively raise the temperature of the water. The question is: "How will the rise in temperature of the water affect the populations of the fish living in the river?" The expected rises in temperatures with time could be entered into the specification equations. The other factors could be assumed to be constant or estimates of their probable values might be entered, and the predicted population levels of all species of fish in the river estimated for time x. A weakness of the model is that it can never predict a species becoming extinct although it will approach zero frequency as a limit.

DISCUSSION

Like any other statistical technique, factor analysis manipulates data in an attempt to reveal the underlying causes and their importance to the variables measured. Three important assumptions are made about the data when factor analysis is employed (Cattell 1965): (1) individual variables and factors are linearly interrelated, (2) two factors act additively in respect to any given variable, and (3) there are no interaction effects among the variables.

No assumptions are made about the distributions of the variables. Various tests for significance of factors do make assumptions regarding the distributions of variables and, for that reason, have been avoided in this paper. It is probable that in any real, relatively large community of organisms all three assumptions will be violated at one time or another. Because of the likelihood of some curvilinear or higher polynomial relationships between factors and variables and because of the existence of nonadditive factors, it is important to know how closely the linear model assumed by the factor analysis approximates the situation where there are some nonlinear relationships between variables and factors.

Cattell & Dickman (1962), using variables and factors between which the relationships were known, showed that if variables are not linearly related to the factors, the factor analysis approximates the determination of the variable by representing a product by a sum. Over a small range this is usually considered to be a good approximation. For example, if a species were determined by two factors acting multiplicatively,

$$\text{Species} = sF_1F_2$$

then the factor-analysis model approximates it by

$$\text{Species} = sF_1 + sF_2$$

After the analysis has been carried out and the number and nature of the factors determined, the linear model can be modified and the predictions improved by experimentally locating nonadditive factors and modifying the series of specification equations. The same can be done with nonlinear relationships between variables and factors. Often the mathematical relationship of a factor to a community of species, if not linear, will be roughly the same for all species (i.e. if the relationship is exponential, it will be exponential for all species).

Two other common situations that modify the relationships between factors and variables are threshold levels and competition for a limited resource. Sometimes a factor influencing a set of variables may operate only above or below a critical value. For example, dispersal in some animals occurs when the population of a species reaches a critical density. The sigmoid curve of population ecology assumes that reaction to increasing density is gradual: the closer the population approaches the carrying capacity of the environment, the slower the rate of growth. It is also possible that there may be a situation where the curve is completely exponential until the carrying capacity has been reached, or surpassed, and a point is reached where density-dependent factors act suddenly. In some predators, search images are formed on abundant species of prey and, when the population of a prey species reaches a critical level, a predator population may begin to attack it to the exclusion of other less common species.

Competition between the members of a community may prove to be more of a problem, and depends on whether populations are controlled by density-dependent or density-independent factors. It is the author's opinion that both types of factors are important in animal communities. One factor influencing a group of species in a community may be a common food resource, such as in a group of insects all feeding on the same species of plant. In the situation of two insect species feeding on one plant species, the feeding of species "a" reduces the amount of factor "X" (the plant) and therefore indirectly influences species "b," the other species feeding on the same species of plant. Factors of this type are referred to as "expendable" and, when they are shown to exist, the specification equations can be modified to take them into account.

The computational steps in the factor analysis technique presented in this paper are outlined in Fig. 14. The assumptions underlying each step of the procedure have been discussed in the Techniques section and will

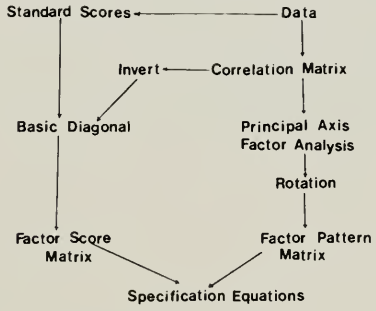


Fig. 14.—Sequence of steps in creating a factor analytical model of a community.

not be repeated here. The experimental steps can be roughly outlined as: (1) definition of the "community" of animals or plants or both to be studied, (2) carrying out of the census, (3) running of the factor analysis, (4) identification of the factors and rotation to a specified hypothesis, (5) formulation of the specification equations (first approximation), (6) discovery and analysis of nonlinear, factor-variable relationships and of nonadditive factors (second approximation), and (7) discovery and measurement of specific factors for each species (third approximation).

As the size of the community studied increases, the number of significant common factors discovered also increases. By increasing the number of species measured, a factor originally specific to one species may now influence a second species and can be picked out by the factor analysis. As more species are considered, more factors must be identified.

Because of the tremendous amount of field work and experimentation needed for this technique, the decision to stop at the first, second, or third approximation will depend on how close the first approximation accurately predicts future changes (or spatial changes) in the species of the community and on how much time and money are available.

A rough approximation is often all that is needed. A farmer usually wants to know only which species, if any, of a set of possible pests will be abundant enough to damage his crops, given a set of conditions that he can predict (e.g., will the application of a certain pesticide in the spring cause an increase in the populations of some potential pest species later in the year?). He is not particularly interested in the exact level of each population.

The factor analysis technique is applicable to modeling communities in both space and time. The factor analysis approach is an improvement over the multiple regression approach (actually a form of factor analysis) in indicating not only how many factors to look for, but also which species are influenced by which factors and the extent of the influences. The psychologists have also found empirically (Kawash, personal communication) that the results of a factor analysis modeling of a siutation using the specification equations tend to be much more useful when applied to similar situations (such as perhaps a model of one river being more applicable to the fishes in an adjacent river), than are the multiple regression equations.

Factor analysis is an extensive and complicated subject. Just how useful this proposed technique will prove can only be known after it has been more extensively used and studied.

LITERATURE CITED

Cattell, R. B. 1965. Factor Analysis: An introduction to essentials. Biometrics 21:190–215, 405–435.

———, and K. Dickman. 1962. A dynamic model of physical influences demonstrating the necessity of oblique simple structure. Psychology Bulletin 59:389–400.

Dagnelie, P. 1965. L'étude des communautés végétales par L'analyse statistique des liasons entre les espéces et les variables écologiques: un exemple. Biometrics 21:890–907.

Goodall, D. W. 1954. Objective methods for the classification of vegetation: III. An essay in the use of factor analysis. Australian Journal of Botany 2:304–324.

Harman, H. H. 1967. Modern Factor Analysis. 2nd ed. University of Chicago Press, Chicago. 474 p.

Hunter, A. S. 1966. High-altitude *Drosophila* of Colombia (Diptera: Drosophilidae). Annals of the Entomological Society of America 59:413–423.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187–200.

Morris, R. F. [ed.] 1963. The dynamics of epidemic spruce budworm populations. Memoirs of the Entomological Society of Canada No. 31.

Pinzka, C., and D. R. Saunders. 1954. Analytic rotation to simple structure, II. Extension to an oblique solution. Research Bulletin RB-54-31. Princeton: Educational Testing Service.

Reyment, R. A. 1963. Multivariate analytical treatment of quantitative species associations: An example from palaeoecology. Journal of Animal Ecology 32:535–547.

Schnell, G. D. 1970. A phenetic study of the suborder Lari (Aves). I. Methods and results of principal components analyses. Systematic Zoology 19:35–57.

Sokal, R. R., and P. E. Hunter. 1955. A morphometric analysis of DDT-resistant and non-resistant house fly strains. Annals of the Entomological Society of America 48:499–507.

———, and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco. 359 p.