Agricultural Meteorology, 19(1978) 349—362

CLASSIFYING ENVIRONMENTS FOR SAMPLING PURPOSES USING A PRINCIPAL COMPONENT ANALYSIS OF CLIMATIC DATA

J. G. PATERSON[1,*], N. A. GOODCHILD[2] and W. J. R. BOYD[1]

[1] *Department of Agronomy, Institute of Agriculture, University of Western Australia, Nedlands, W.A. (Australia)*
[2] *Biometrics Unit, Institute of Agriculture, University of Western Australia, Nedlands, W.A. (Australia)*

(Received November 20, 1976; accepted November 23, 1977)

ABSTRACT

Paterson, J. G., Goodchild, N. A. and Boyd, W. J. R., 1978. Classifying environments for sampling purposes using a principal component analysis of climatic data. Agric. Meteorol., 19: 349—362.

A method is described for specifying climatic variability using a principal component analysis of real and derived meteorological variables. Two vectors derived from the analysis delimited areas in Western Australia between which climatic differences were large relative to those within such areas. These vectors were used to construct a grid map of the region. This map was then used in studies on the ecology of wild oats in Western Australia which required the location of a number of complex experiments at sites chosen to sample regional climatic variation.

A note of caution is given emphasising that variables should not be selected carelessly and stressing the undesirability of necessarily attempting to explain individual vectors.

INTRODUCTION

Studies on the ecology of plant species need quantitative assessment of the environment in which they are located. The specification of climatically homogeneous areas is also necessary so that ecologists can minimise the complications of genotype-environment interaction by more reliably sampling environmental variability.

Fisher (1967) divided the Western Australian arable area into climatically homogeneous regions by a grid with one axis approximately NNW/SSE demarcating mean annual rainfall and the other nearly E/W specifying mean temperature regions. His objective was to derive a rational basis for the siting of cereal variety testing centres. Such an approach can be used in an area like the Western Australian wheatbelt where major orographic irregularities do

*Present address: Department of Agriculture, Jarrah Road, South Perth, W.A. 6151 Australia.

not exist and where there are relatively even graduations of climatic factors such as rainfall (Gentilli, 1971). However, for more detailed studies of the behaviour of species or varieties in different environments more refinement is necessary. For this purpose Goodchild and Boyd (1975) used long-term wheat yields in a principal component analysis to study the variability of yield in the wheat belt.

The objective of the analysis discussed here was to provide a basis upon which an extensive experimental programme could be carried out to test the effects of seasonal conditions and soil fertility on the two major species of wild oats (*Avena* spp.) found in Western Australia, using ecodemes which were specifically adapted to different regions. Ecodemes are defined as composite groups of accessions representing those occurring within a particular habitat (Gilmour and Gregor, 1939).

To some extent the selection of climatic variables was necessarily subjective, but since it was known that the wild oat species differed in their response to length of growing season, to the quantity and distribution of rainfall in the growing season, to vernalization and to photoperiod (Paterson, 1976; Paterson et al., 1976a) it was possible to select logical variables.

METHODS

Climatic variables

Although the duration of the growing season varies widely in Western Australia (Waterhouse, 1969), Boyd et al. (1976) demonstrated that little regional variation in individual climatic factors occurs for those periods of the growing season before, or after, the winter months of June, July and August.

Data for this period of all the climatic variables found to be relevant, except for vernalization, were therefore used in the calculations; the vernalization function, which was the exception, is operative during seedling growth, which often occurs as early as May. As all the climatic information used in this study was obtained from mean monthly data prepared by Waterhouse (1969), from data provided by the Bureau of Meteorology (1977) or derived from the same data by Boyd et al. (in Waterhouse, 1969), only the general details of each derived variable is discussed below.

Potential soil moisture storage

On average, in excess of 50% of the mean annual rainfall in South-Western Australia occurs during the three winter months of June, July and August (Table I). Waterhouse (1969) states that throughout most of this period the amount of moisture received exceeds that required to sustain cereal crop growth, thereby enhancing the possibility of transient waterlogging in mid-

TABLE I

The mean annual rainfall and the proportion falling in winter (June, July and August) at selected sites in South-Western Australia

Site	Number of years	Annual rainfall (mm)	Winter rainfall (%)
Albany	95	945 ± 156	41
Bridgetown	87	855 ± 152	51
Bunbury	100	881 ± 167	55
Carnamah	85	396 ± 103	53
Chapman	68	462 ± 118	58
Collie	76	983 ± 187	54
Esperance	6	700 ± 89	45
Geraldton	33	476 ± 134	59
Katanning	83	488 ± 99	45
Kellerberrin	82	338 ± 93	46
Merredin	71	332 ± 83	44
Moora	76	467 ± 110	54
Morawa	59	349 ± 92	48
Mullewa	79	347 ± 99	52
Northam	93	438 ± 108	54
Perth	102	879 ± 164	57
Wandering	87	630 ± 161	53
York	97	455 ± 110	54

winter whilst ensuring recharge of the soil profile before the rainfall decreases sharply in September. Potential soil moisture storage was therefore estimated for the winter period from rainfall received less potential plant use. This latter value was calculated from the modified Meyer ratio $I \times E_p^{0.75}$ (Butler and Prescott, 1955) where E_p is open pan evaporation, and I is a factor calculated to relate rainfall requirement to E_p. Fischer and Kohn (1966) proposed $I = 1{\cdot}65$ for a growing cereal crop and this figure was used. Isograms of potential soil moisture storage, given in Fig.1, reflect the combined effects of the contrasting possibilities of growth limitation in the winter (due to excess moisture) and spring (due to deficient moisture).

Vernalization function

Paterson et al. (1976a) showed that *Avena fatua* required little or no cold treatment for floral initiation whereas *A. barbata* selections differed widely in their cold requirement. Since *Avena* species germinate with the first rains, usually in May, the vernalization function (day-degrees) was calculated for the months of May and June. The formula $\Sigma(10 - \bar{x}_{\min})$, where n is the number of days in each month and $x_{\min}$ — the mean monthly minimum temperature (°C) — is less than 10, was used (Fig.2). The upper level at which vernalization is likely to occur is 10°C (Calder, 1966).

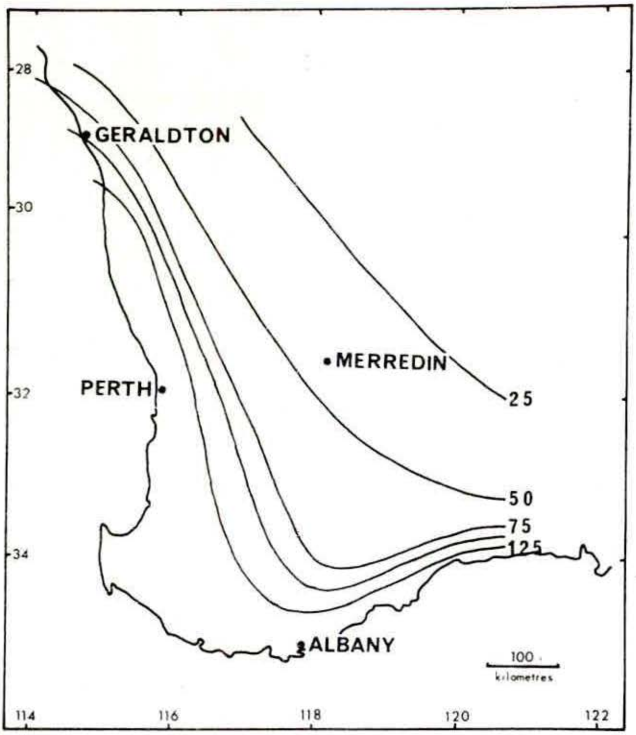

Fig.1. The distribution of isograms of potential soil moisture storage (mm) in South-Western Australia, calculated as the difference between winter rainfall and an estimate of potential evapotranspiration.

Growth function

The growth function (day-degrees) for the June to August winter period was calculated using the formula $\Sigma n(\bar{x} - 4.4)$, as suggested by Nuttonson (1958) where n is the number of days in each month and $\bar{x}$ is the mean monthly maximum temperature (°C) (Fig.3). Plant growth was assumed to cease at temperatures lower than 4.4 °C.

Radiation

Radiation data was available from only a few stations and the values used were estimated from these. The function was calculated as the summation of total radiation (M J m^{-2}) over the winter period; variation during this period was small (Fig.4).

Growing season

Growing season duration was included since it was recognised that late flowering ecodemes of *Avena* predominate in areas of long growing season and vice versa (Paterson, 1974). The opening data of the season was based on the occurrence of precipitation relative to a cereal seed's moisture requirement for germination, which was based on the formula $0.54 \times E_p^{0.75}$ (Prescott, 1949). The closing data was predicted on the long-term averages of the dates on which cereal grain depots nearest to each data site were opened (Co-operative Bulk Handling, Perth, W.A.). These represent the average date by which cereal grain moisture has fallen to 12.5%. From these figures the closing date was obtained by deducting 21 days, which defined the approximate cessation of photosynthesis. Isograms are presented in Fig.5.

Due to the wide range of soil types and profiles found in Western Australia (Mulcahy, 1967) and the extremely limited information on soil moisture characteristics it was not possible to consider water balances or to use such considerations to define the closure of the growing season.

Analysis

The five variables described above are not completely independent of each other, therefore each describes variation common to one or more of the others as well as providing additional information. It is therefore desirable to com-

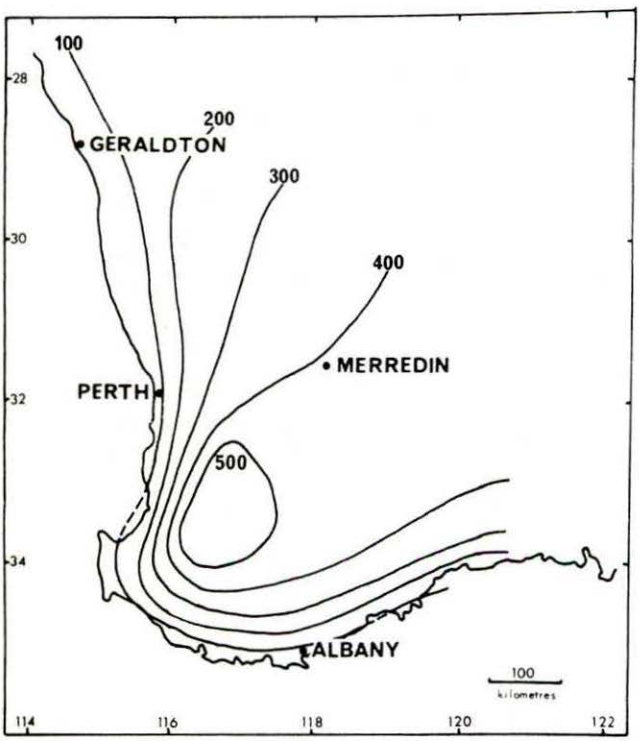

Fig.2. The distribution of isograms of the vernalization function (day-degrees for May and June) in South-Western Australia.

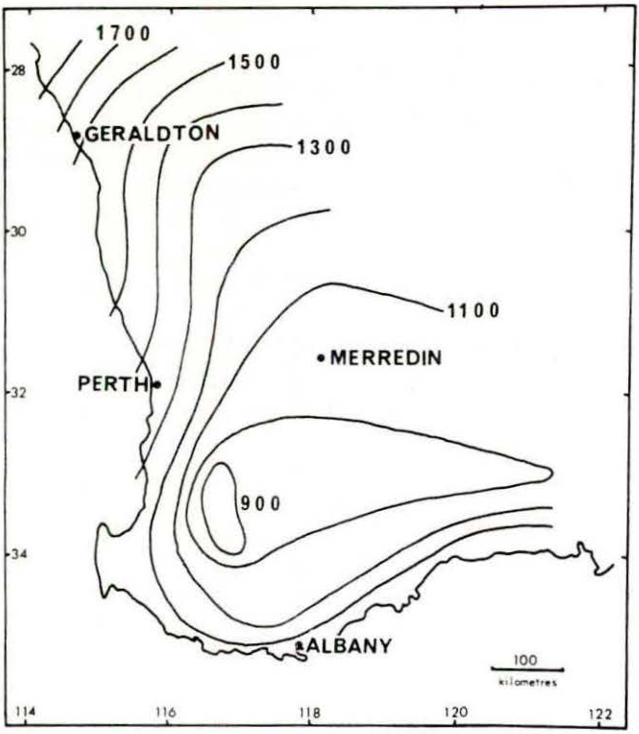

Fig.3. The distribution of isograms of the winter growth function (day-degrees) in South-Western Australia.

bine these variables in such a way that each new variable will account for information independent of that accounted for by any other new one; this transformation must be done in such a way that the sum of the variances of the new variables is equal to the total variance of the original ones. Such a transformation is provided by a set of principal components of a variance covariance matrix.

Details of the method of principal component analysis are in Seal (1964) and in many general texts on matrix algebra. It can be shown that principal components are merely a rotation of the original variate axes. Hence the data points, and any other hypothetical points expressible in terms of the original variates, can be fixed in relation to the principal components. A simple explanation of the principal component analysis is given in the Appendix. An additional advantage of the principal component analysis is that it frequently enables the dimensionality of the data to be reduced, with little loss of information, by dropping one or more of the smallest sized components; the small components or vectors usually provide no interpretable information.

This form of analysis was used to extract the five possible principal components from the standardised variance covariance matrix, that is the correlation matrix, of the five original climatic variables. The correlation matrix was used because the original variates were measured in different units and their absolute variances bear no relation to each other. For individual vectors the value for each recorded site was calculated and these values were inserted at the site position on a map. It was then possible to construct isograms by joining points of equal value. Linear interpolation using existing site values was used, where necessary, to enable contours to be drawn at the required intervals.

RESULTS

The sum of the latent roots of the two largest principal components accounted for 90.2% of the total variance of the climatic data (Table II). The remaining three components were all small and the sum of their variances (latent roots) accounted for less than 10% of the total variance; they were therefore not considered further. The two large components, being orthogonal to each other, were convenient geometrical axes upon which to base geographically located isograms describing variation of the climate. The plots were completed on a map of the agricultural area of Western Australia using the standard

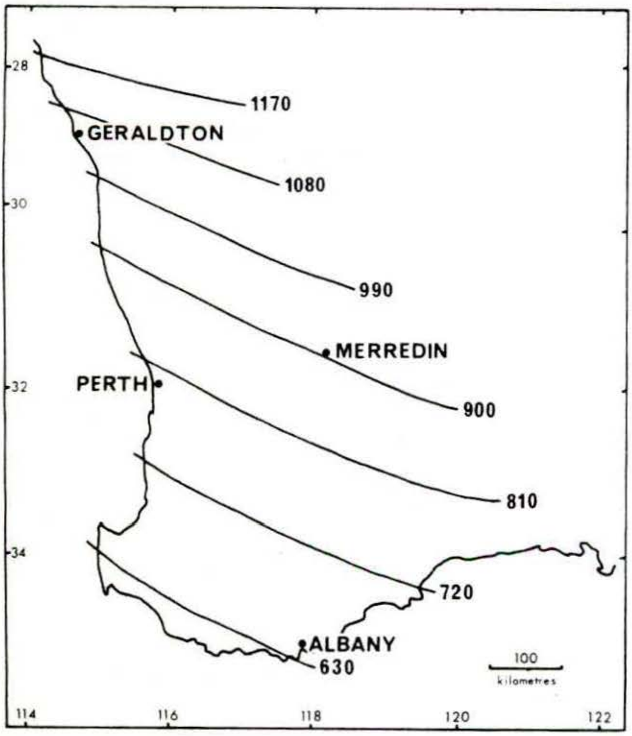

Fig.4. The distribution of isograms of total winter radiation ($M\ J\ cm^{-2}$) in South-Western Australia.

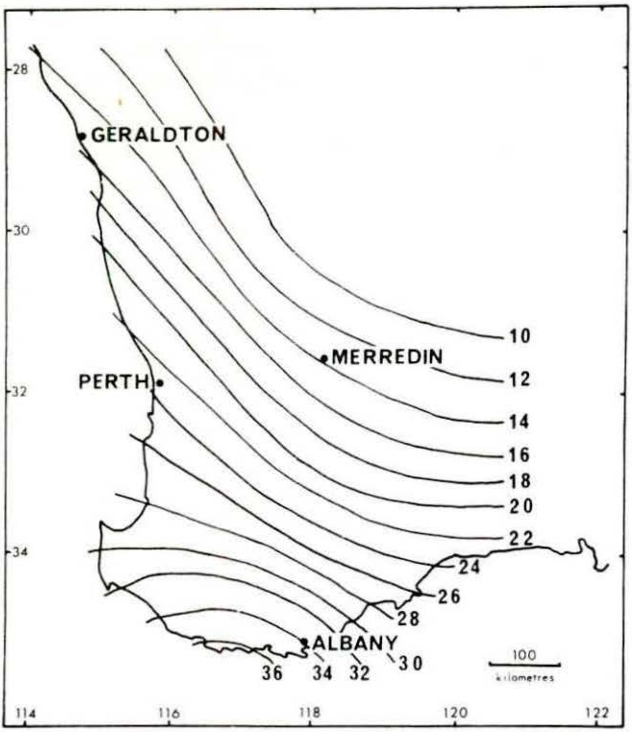

Fig.5. The length of the growing season (weeks) in South-Western Australia based on the moisture requirement for seedling establishment and a derived estimate of the cessation of photosynthetic activity of a cereal-type grass.

procedure for expressing site values in terms of the vector axes instead of the original climatic variates (cf. Seal, 1964). Isograms were formed on the map (Fig.6) for each vector.

The spacing of the isograms of each vector was arranged so that the agricultural area was divided into 52 regions each of which was relatively homogeneous with respect to the two vectors, since the interval between the contours was small.

The composition of the first vector indicated that about 60% of the influence of the five climatic variables is associated with a tendency for moisture, the vernalization function and growing season duration to decrease when the growth function and radiation increase (Table II). The second vector accounted for 30% of the total climatic variance when the effect of the first vector had been removed. This vector represents the growth function, growing season duration and moisture reserve variables operating in a similar direction and opposing that of the vernalization function. Radiation made a negligible contribution to this vector.

TABLE II

The climatic coefficients of the first two principal components (vectors) formed in the analysis of climate

Vector:	1	2
Latent root:	2.97	1.54
Percentage of variance accounted for:	59.4	30.8
Variates:	*Coefficients*	
Potential moisture storage	−.32	+.64
Vernalization function	−.44	−.45
Growth function	+.47	+.45
Radiation	+.53	−.02
Growing season	−.45	−.43

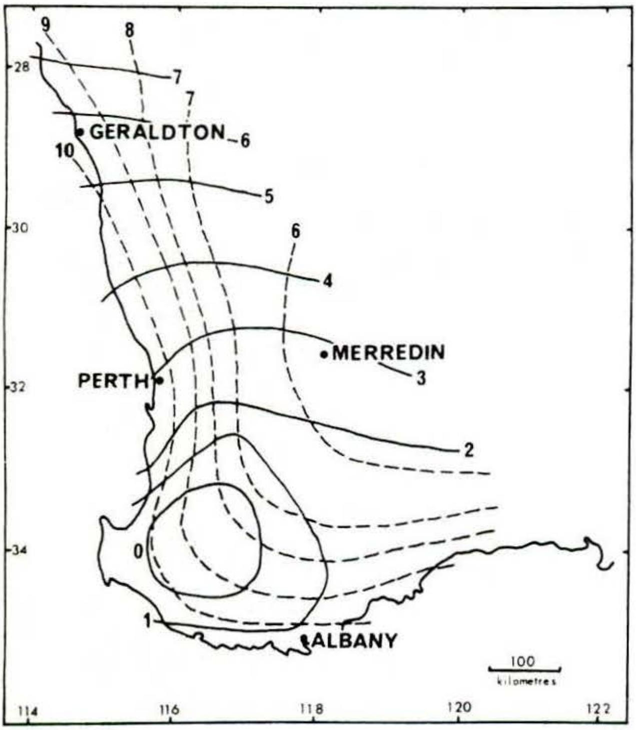

Fig.6. The distribution of isograms drawn from the first (—) and second (- - -) principal components of an analysis of five climatic variables in South-Western Australia.

DISCUSSION

In ecological studies it is often important to define areas which are similar in terms of climate. In many respects one of the best quantitative measures of the environment for a crop is the yield of that crop, the yield of some related species or the yield of some species of known similar requirements. Thus the crop is used as an integrative biological indicator of environmental conditions. This approach led Goodchild and Boyd (1975) to use long-term mean wheat yields to define environmental regions which were then used by Boyd et al. (1976) in a study of the effect of environmental conditions on wheat yields. For wild oats yield data were not available and some other method of differentiating the relative biological value of the environments was required.

Using two vectors of the principal component analysis it was possible to stratify the area analysed in two directions so that each direction was independent and at right angles to the other. Provided the lines of the grid so defined are arranged at relatively small intervals on a vector the area between any pair of grid lines will vary minimally with respect to that vector. Therefore the area enclosed by two sets of grid lines will be relatively homogeneous in terms of the two vectors. The limits to the climatic changes in these areas will depend on the size of the intervals between the isograms. The experimenter has control over this distance and thus can also to some degree determine the homogeneity of the individual sample areas.

Provided the assumption that the original climatic variables are relevant to the growth and development of wild oats is valid, and there is good evidence that it is (Paterson et al., 1976a, b), the areas enclosed by the grid lines delimit regions within which sites for studying the ecological status of wild oats may be selected. Close limits to climatic variability were not particularly important. The main requirement was that the chosen sites should be climatically dissimilar and hence the division shown in Fig.6 was acceptable. Time, labour, costs and the availability of suitable plots limited the number of sites that could be studied. As can be seen from the figure the coverage was not complete but the areas did provide a reasonable sampling of the environments that had been determined by the relatively wide grid. Recognising also the large area covered it seemed that this was a practicable sample method.

The results of the experiments are described by Paterson (1974). He showed how certain environments, defined in terms of the principal component analysis, favoured some plant ecodemes at the expense of others. The experiments also enabled him to demonstrate experimentally the deductions he had made at an earlier date on the distribution preferences of the two species and on their favoured ecological niches (Paterson, 1974, 1976; Paterson et al., 1976a, b, c). Not only did the analysis described in this paper furnish a means of sampling the agricultural area of Western Australia but it also provided some quantitative assessment of the climatic conditions at each selected experimental site.

Consideration now needs to be given to principal component analysis as a tool which can be used in environmental studies. What the method does in effect, after the selection of relevant climatic variables, is to define the directions of progressively decreasing variation so that the first vector is the direction of greatest variability and so on. Some directions may be nearly invariate and may be discarded. It may or may not be reasonable to place a meaning on some or all the vectors; such interpretations, if any, are unimportant at least in the context of the type of work discussed here. The method merely requires the definition of the directions of greatest variation and these are defined by the climatic data alone.

Having defined the direction of greatest magnitude care must be taken when attempting to probe further. In the principal component method as used here there is no direct attempt to associate variates, either original or newly derived with crop growth; the analysis is concerned only with variation in the climatic data. If the original variates have no effect on or relationship with the crop variables, then the principal component vectors will only be measures of crop performance by chance. It is therefore essential, before embarking on a principal component analysis for the purpose described here, that the original variates be chosen from those which may be expected to vary in a manner similar to the variability of the crop.

A further word of warning is necessary; the conditions under which these analyses were used were such that changes were part of extensive trends. Thus rainfall varies largely in a west to east direction and the incidence and intensity of the temperature functions also vary in a regular manner. In the example we have given, and because of the absence of major orographic features, sudden and distinct changes over a short geographic distance do not occur. In a situation where there are marked, sudden and unpredictable changes the principal component analysis is less likely to be effective. Indeed, it would be expected to be totally ineffective, unless a large number of recorded sites provided long-term data for the analysis.

ACKNOWLEDGEMENTS

The work described is part of a Ph.D. project of one of us (J. G. P.) who was supported by a University of Western Australia studentship. The authors are also grateful to the Wheat Industry Research Committee of Western Australia for financial support, to Professor W. R. Stern for his continued interest and advise on the manuscript and to Dr. J. J. Landsberg and Mr. L. P. Smith for helpful discussion on a draft of this paper.

APPENDIX

Principal component analysis provides a means of summarising data and reducing its dimensionality with little loss of information. It also enables data to be expressed in terms of a set of wholly independent derived variates defined as linear combinations of the original variates. For an analysis of this type the data need to be set out as a matrix. Mathematically, there are many possible types but the only one of concern here is a symmetrical non-singular matrix, the variance covariance matrix, or its standardised form, the correlation matrix. The matrix is non-singular since it can be represented for this purpose by a single number, a scalar, which is not zero and is called the 'determinant'. Conversely, if the determinant is zero the matrix is singular.

Consider two symmetrical matrices, A and B, one of which is a variance covariance matrix. Theory states that if these matrices have m values in the rows and in the columns there will be m values of λ such that the matrix $(\mathbf{A} - \lambda\mathbf{B})$ will be singular. The values of λ satisfying this requirement are known as the latent roots or eigen values.

The singularity condition of a matrix implies that if the elements of any row or column, treated as a column vector, are premultiplied by a special set of multipliers, treated as a row vector, the result will be zero. Thus taking a matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

with a vector of multipliers:

$$v_1 \quad v_2 \quad v_3 = v$$

Then:

$$v_1 \quad v_2 \quad v_3 \quad \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix}$$

$$= v_1 a_{11} + v_2 a_{12} + v_3 a_{13} = 0$$

There is one set of special multipliers corresponding to each value of λ derived from the matrix **A**. These sets are the latent vectors or eigen vectors of **A**.

If the matrix **B** is of the form:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ i.e., it is the Unit Matrix I}$$

then the statement given above becomes:

"$(\mathbf{A} - \lambda\mathbf{I})$ is singular"

The latent vectors derived from **A** in this case are also known as the principal components and their latent roots (λ) represent that part of the total variance accounted for by the principal component.

Mathematically these results can be expressed as:

$$(\mathbf{A} - \lambda\mathbf{I})\,v = 0$$

or as the determinental equation:

$$\mathbf{A} - \lambda \mathbf{I} = 0$$

where $|Z|$ is the determinant of a matrix $\mathbf{Z}$. The result of the extraction of the latent roots and vectors of a variance covariance matrix is known as a principal component analysis. The m vectors or principal components ν will usually be found to account for unequal proportions of the total variance. Those components accounting for a small part of the total variance may be neglected as being of minor significance in the data being studied.

The information used in a principal component analysis is that contained in the original data; each principal component and latent root is obtained separately and independently so that the total of the latent roots equals the total variance. The information contained in each principal component is clearly a part of the whole and this information is not included by any other component. Thus the principal components are independent in the statistical sense. They represent the data as do the original variates, the latter being represented in matrix A by the rows or the columns. The principal components may be considered as new variates which, like the originals, define axes through the plot of the data. The principal component axes are at right angles to each other and are so oriented that the largest lies along the longest axis of the data points, the next largest along the next longest axis at right angles to the first and so on. It is not difficult to see that these axes can be used as co-ordinates in a multidimensional space.

REFERENCES

Boyd, W. J. R., Goodchild, N. A., Waterhouse, W. K. and Singh, B. B., 1976. An analysis of climatic environments for plant-breeding purposes. Aust. J. Agric. Res., 27: 19—33.

Butler, P. F. and Prescott, J. A., 1955. Evapotranspiration from wheat and pasture in relation to available moisture. Aust. J. Agric. Res., 6: 52—61.

Calder, D. M., 1966. Inflorescence induction and initiation in the *graminae*. In: F. L. Milthorpe and J. D. Ivins (Editors), The Growth of Cereals and Grasses. Butterworths, London, pp. 59—73.

Fischer, R. A. and Kohn, G. D., 1966. The relationship between evaporation and growth in the wheat crop. Aust. J. Agric. Res. 17: 255—267.

Fisher, H. M., 1967. Cereal yield tests in 1966. J. Dep. Agric. W.A., 8: 372—378.

Gentilli, J., 1971. Australian Climatic Patterns. Nelson, Melbourne.

Gilmour, J. S. L. and Gregor, J. W., 1939. Demes: a suggested new terminology. Nature (Lond.), 3642: 333.

Goodchild, N. A. and Boyd, W. J. R., 1975. Regional and temporal variations in wheat yield in Western Australia and their implications in plant breeding. Aust. J. Agric. Res., 27: 209—217.

Mulcahy, M. J., 1967. Landscapes, laterites and soils in southwestern Australia. In: Landform studies from Australia and New Guinea. Aust. Nat. Univ. Press, Canberra, A.C.T.

Nuttonson, M. Y., 1958. The Physical Environment and Agriculture of Australia. Am. Inst. of Crop Ecology, Washington, D.C.

Paterson, J. G., 1974. The Distribution and Ecology of Wild Oats (*Avena* spp.) in the

Agricultural Enironment of Western Australia. Unpublished Ph.D. thesis of the University of Western Australia.

Paterson, J. G., 1976. The distribution of *Avena* species naturalised in Western Australia. J. Appl. Ecol., 13: 257–264.

Paterson, J. G., Boyd, W. J. R. and Goodchild, N. A., 1976a. Vernalization and photoperiod requirement of naturalised *Avena fatua* L. and *Avena barbata* Pott ex Link in Western Australia. J. Appl. Ecol., 13:265–272.

Paterson, J. G., Boyd, W. J. R. and Goodchild, N. A., 1976b. Effect of temperature and depth of burial on the persistence of seed of *Avena fatua* L. in Western Australia. J. Appl. Ecol., 13: 841–847.

Paterson, J. G., Goodchild, N. A. and Boyd, W. J. R., 1976c. Effect of storage temperature, storage duration and germination temperature on the dormancy of seed of *Avena fatua* L. and *Avena barbata* Pott ex Link. Aust. J. Agric. Res., 27: 373–379.

Prescott, J. A., 1949. Climatic expressions and generalised climatic zones in relation to soils and vegetation. Proc. Spec. Conf. Agric., Adelaide and Canberra, pp. 27–33.

Seal, H., 1964. Multivariate Statistical Analysis for Biologists. Methuen, London.

Waterhouse, W. K., 1969. An Agro-Ecological Survey of Cereal Production in the Agricultural Areas of Western Australia. Unpublished M.Sc. thesis of the University of Western Australia.