Abstracta Botanica 17(1-2): 141-146, 1993

The correspondence between geographic position and climatic ordinations of the major vegetation zones of China: a Procrustes analysis approach

E. Feoli[1], J. Podani[2] & C. Y. Sun[3]

International Institute of Earth, Environmental and Marine Sciences and Technologies, Trieste, I-34100, Italy. Permanent addresses:
1 Department of Biology, University of Trieste, Italy.
2 Department of Plant Taxonomy and Ecology, L. Eötvös University, Ludovika tér 2, H-1083, Budapest, Hungary.
3 Institute of Botany, Academia Sinica, Beijing, China.

Keywords: Correlation, Multiple regression, Phytoclimatic classes, Phytoclimatic zones.

Abstract: The relationship between climatic ordinations and the geographic position of meteorological stations has been studied via Procrustes analysis in order to measure the effects of some geographic features on climatic variation within phytoclimatic zones of China. The analysis showed that for nearly all phytoclimatic zones there is a high correspondence between the position of the metereological stations in multidimensional climatic space and in the territory. This means that the climatic variation follows a regular pattern in the territories of phytoclimatic zones. However, multiple regression confirmed that Procrustes measures of dissimilarity within each zone are significantly correlated with the standard deviation of altitude and the number of rivers of that zone. This indicates that these two geographic features affect the pattern of climatic variation.

Introduction

Climatic data collected worldwide by meteorological stations are often used to create phytoclimatic classifications (Holdrige 1947, Walter & Lieth 1967). These classifications are based on the assumption that climate is the dominating factor in controlling vegetation variation. This looks true at large-scale: in an extensive multivariate study Sun & Feoli (1991) have found that climate by itself is responsible for more than 60% of the differences between the major vegetation types of China. Phytoclimatic classes obtained by classification of vegetation types based on climatic data may be used to subdivide a territory into phytoclimatic zones. The climatic variation within a given geographical area and within each of its phytoclimatic zones may be affected by latitude, longitude and altitude as well as by other geographic features such as presence of mountains and rivers. It is expected that the high correspondence between the geographic position (latitude, longitude and altitude) and the position of the meteorological stations in multidimensional climatic space should be very high if the landscape would vary very regularly. On the other hand, low correspondence would mean that the landscape is heterogeneous and that climatic predictions would be very difficult if based merely on latitudinal - longitudinal - altitudinal data. Procrustes analysis provides measures of correspondence between the geographic position of the stations and their position in multidimensional climatic space, while multiple regression analysis provides measures quantifying how much landscape heterogeneity may influence the climatic pattern and therefore vegetation pattern within phytoclimatic zones. The present paper investigates within the Chinese territory the correspondence between the climatic ordination of meteo-

rological stations and their geographic position in order to measure the predictability of climate with respect to vegetation within phytoclimatic zones.

Material and methods

Data and previous analyses

Although a more detailed account is given in Sun and Feoli (1991), it is necessary here to give a brief summary of data and some earlier results, because they provide the starting point for this study.

The basic data set comprises meteorological observations for 106 climatic variables measured from 1951 to 1980 at 644 stations all around China. Based on these observations, a 644 by 106 data matrix of average monthly values was created. The application of sum of squares ranking (Orlóci 1973) reduced the number of variables to 10, with 94% efficiency. The 644 by 10 data matrix is used in this paper to generate ordinations of stations.

The 644 sites were assigned to each of the 23 main vegetation types currently recognized in China (Chang 1988). By selecting a random seed station from each vegetation type, iterative nonhierarchical clustering (Anderberg 1973) was performed based on the 106 original climatic variables, in order to define the 23-climatic clusters. The resulting clusters were further subjected to incremental sum of squares clustering, and 8 major groups were identified from the resulting dendrogram. These groups largely correspond to the eight vegetation zones suggested by Chang (1988) for China. In the sequel, these eight groups will be referred to as *phytoclimatic zones* (see Table 1, for names and the numbers of stations included (N) and Figure 1 for the geographic distribution of these zones).

The landscape heterogeneity of each phytoclimatic zone was expressed in terms of the standard deviation of elevation (SDE) and the number of river branches (NR). The first variable clearly shows if there are excessive elevation differences among stations, whereas the other reflects geomorphological diversity in a rough way.

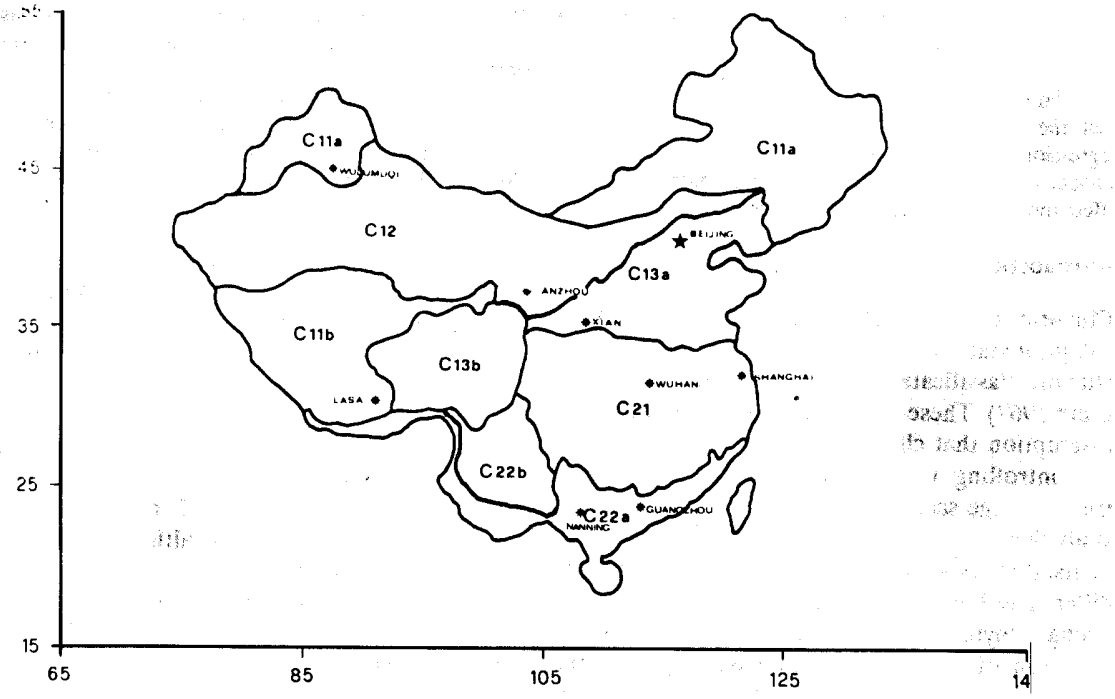

Figure 1. Geographic distributions of phytoclimatic zones of China. C11a: Cold-temperate and temperate coniferous forest and steppe. C11b: Alpine grassland. C12: Temperate desert and steppe. C13a: Warm temperate deciduous broadleaved forest. C13b: Montane temperate steppe and high mountain subtropical evergreen broadleaved forest. C21: Subtropical evergreen broadleaved forest. C22a: Subtropical montane evergreen broadleaved forest. C22b: Tropical rain and monsoon rain forest.

Ordinations

Ordination of the meteorological stations was obtained by standardized *principal component analysis* (PCA, see e.g., Jolliffe 1986) for the complete data set and for each phytoclimatic zone, leading to 9 different ordinations.

Procrustes analyses

This method introduced by Schönemann & Carroll (1970) and Gower (1971) allows the comparison of configurations of m points in n-dimensional spaces represented by n x m matrices. Two such matrices, **X** and **Y** could be compared by the sum of squared positional differences given by

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-y_{ij})^2 = \mathrm{tr}\,(\mathbf{X}-\mathbf{Y})'(\mathbf{X}-\mathbf{Y})$$

Since there is no preferred position and orientation of the configurations of points in ordination scattergrams, ordination scores can be considered arbitrary. We can move the points in **Y** relative to **X** through rotation, reflection and translation to optimize the goodness of fit of **Y** to **X**. The first step is the centring of **X** and **Y** so that both sets of points have their centroid at the orgin (for convenience, **X** and **Y** will denote centered matrices in the sequel). The required rotation of **Y** is obtained by its multiplication with the n x n orthogonal transformation matrix

$$\mathbf{H} = \mathbf{VU}'$$

where **U** and **V** result from the singular value decomposition of

$$\mathbf{X}'\mathbf{Y} = \mathbf{USV}'$$

After finding these matrices, the goodness of fit statistics will be computed as

$$R_E^2 = \mathrm{tr}(\mathbf{X}-\mathbf{YH})'(\mathbf{X}-\mathbf{YH}) = \mathrm{tr}\,(\mathbf{XX}') + \mathrm{tr}\,(\mathbf{YY}') - 2\,\mathrm{tr}\,(\mathbf{YX}'\mathbf{XY}')^{1/2}$$

which is a symmetric measure. R_E^2 depends on the actual scales of **X** and **Y**, therefore this statistic is not appropriate for arbitrarily scaled ordinations. Inclusion of a scaling parameter c in the transformation of **Y** to c**YH** resolves this problem

$$c = \mathrm{tr}(\mathbf{YHX}')/\mathrm{tr}(\mathbf{YY}')$$

leading to

$$R_S^2 = \mathrm{tr}\,(\mathbf{XX}') - 2\,(\mathrm{tr}\,(\mathbf{YX}'\mathbf{XY}')^{1/2})^2 / \mathrm{tr}(\mathbf{YY}')$$

This coefficient is unsymmetric, therefore Gower (1971) suggested to scale each configuration after centering, so that

$$\mathrm{tr}(\mathbf{XX}') = \mathrm{tr}(\mathbf{YY}') = 1$$

to yield a coefficient denoted by d^2. As Sibson (1978) pointed out, R_S^2 may be standardized directly

$$\gamma_S = R_s^2 / \mathrm{tr}\,(\mathbf{XX}')$$

resulting a range between 0 and 1, and there is a simple relationship between this and Gower's measure

$$d^2 = 2\{1 - (1-\gamma_s)^{1/2}\}$$

Then it becomes clear that d^2 ranges from zero to 2, but this simple fact is not mentioned in most texts.

Procrustes analysis was used here for the comparison of the PCA ordinations with the actual geographical positions of the stations. All subsets of the PCA scores and the geographical coordinates were normalized to unit sum of squares from the respective centroids in order to obtain a symmetric measure of distance between the configurations (Digby & Kempton 1987). The distance is expressed as the sum of squared distances (denoted by d^2) between the corresponding points of the two ordinations such that the point configurations are fitted to each other as perfectly as possible. Two series of analyses were made. In the first, the scores on PCA axes 1 and 2 were contrasted with the first two geographical variables, i.e., latitude and longitude. The second series of analyses used three PCA dimensions and three geographical variables, latitude, longitude and elevation above sea level (altitude).

The assessment of how closely PCA ordinations reflect the geographical positions requires comparison of resulting scores. Even though the theoretical upper limit of d^2 is 2, regardless the number of dimensions and points, the absolute d^2 measures derived from Procrustes analyses involving different numbers of points and dimensions are not comparable directly. As Podani (1991a) pointed out, a value of d^2=.99, for example, for 10 points does not indicate closer relationships between the respective configurations than, say, a score of 1.00 for 100 points. The reason is that these two values are not equally likely for randomly generated ordinations under different circumstances. A more meaningful approach is thus to consider random expectations and probability

distributions, estimated by Monte Carlo simulation (Podani 1991a). As a measure of ordination dissimilarity, the actual d^2 value divided by the random expectation is used,

$$\delta^2 = d^2/(E(d^2))$$

so that measures derived from situations with different numbers of points and/or dimensions will become directly comparable. The simulated frequency distributions facilitate assessment of the significance of ordination dissimilarity. In this study, for each combination of the number of points and dimensions, 1000 pairs of random ordinations were simulated to estimate the expected values and probabilities.

Regression and correlation

In order to obtain a rough estimate on the relationships between the δ^2 values and landscape heterogeneity, multiple regression analysis was performed between δ^2 as dependent variable and SDE, NR and AREA as independent variables. Product moment correlation coefficient was also calculated between δ^2 and such variables.

Computer programs

Principal components analyses were performed using program PRINCOMP, whereas Procrustes comparisons were done by a modified version of program PRANA, both programs are included in the SYN-TAX IV package (Podani 1991b). The random number generator routine used in the simulations was taken from Park and Miller (1988). The number of river branches was calculated from the vegetation map of China and the estimation of surface area of phytoclimatic zones was performed by the Map II geographical data processing system (Kirby et al. 1989). Multiple regression and the correlations were calculated using program SPSS (Nie et al. 1975). Computations were carried out on a Macintosh IIfx computer of IIEM and the Vax mainframe of the computer center of University of Trieste.

Results and discussion

The 9 ordination scattergrams are not shown here. Instead, we present the percentages explained by components in Table 1. Note that the 10 variables retained explain 94% variability of the original 106. Thus, the efficiencies in this table refer to 94%.

The d^2 values, associated probabilities, expectations from the simulated distribution, and standardized dissimilarities are summarized in Tables 2- 3. The symbol <.001 in the tables means that the actual d^2 was lower than the simulated minimum. This is the most common outcome showing that, in general, ordinations based on climatic information agree very well with the geographical positions of stations. The "worst" case is the warm temperate zone for 2 dimensions, because the value of 1.543 can come from random ordinations at the conventional .05 probability level.

The correspondence between ordinations and geographical data is better in 2 dimensions than in 3 for three zones: cold-temperate, alpine grassland and temperate desert. For the tropical forest

Table 1. Cumulative percentages of efficiency of the first three PCA axes (1, 2, 3) of phytoclimatic zones. N = number of stations.

Cluster	N	1	2	3
1 Cold-Temperate	95	48.99	70.75	84.99
2 Alpine grassland	44	50.88	75.48	86.61
3 Temperate desert	141	43.35	68.07	83.95
4 Warm temperate	75	46.09	69.84	80.59
5 Montane temperate	23	56.21	76.60	89.83
6 Subtropical forest	175	42.61	63.66	77.45
7 Tropical forest	61	53.58	75.71	84.99
8 Subtropical Montane	30	57.49	78.04	88.50
All stations	644	60.10	74.01	84.92

Table 2. Summary of Procrustes analyses and simulations for 2 dimensions. N = number of stations. For d^2, $\hat{p}$, $E(d^2)$ and δ^2 see text

Cluster	N	d^2	$\hat{p}$	E(d2)	δ^2
Cold-Temperate	95	1.343	<.001	1.762	.76
Alpine grassland	44	1.052	<.001	1.648	.64
Temperate desert	141	1.032	<.001	1.807	.57
Warm temperate	75	1.543	.051	1.727	.89
Montane temperate	23	0.852	<.01	1.521	.56
Subtropical	175	1.555	.001	1.829	.85
Tropical	61	0.696	<.001	1.707	.41
Subtropical Montane	30	1.191	.02	1.569	.76
All stations	644	0.747	<.001	1.906	.39

Table 3. Summary of Procrustes analyses and simulations for 3 dimensions. N = number of stations. For d^2, $\hat{p}$, $E(d^2)$ and δ^2 see text

Cluster	N	d^2	$\hat{p}$	E(d2)	δ^2
Cold-Temperate	95	1.413	.001	1.705	.83
Alpine grassland	44	1.180	<.001	1.569	.75
Temperate desert	141	1.239	<.001	1.761	.70
Warm temperate	75	1.176	<.001	1.670	.70
Montane temperate	23	0.561	<.001	1.393	.40
Subtropical	175	1.185	<.001	1.784	.49
Tropical	61	0.799	<.001	1.632	.66
Subtropical Montane	30	0.550	<.001	1.473	.37
All stations	644	1.072	<001	1.890	.57

Table 4. Summary of Procrustes analysis. δ^2 (3 dimensional and 2 dimensional cases), standard deviation of elevation (SDE), coefficient of variation for elevation (CV), number of rivers (NR), and area of phytoclimatic zone (AREA).

$\delta^2(3)$	$\delta^2(2)$	SDE	CV	NR	AREA
.83	.76	638.3	1.02	424	2208.30
.75	.64	576.7	0.15	287	1494.79
.70	.57	697.7	0.60	281	1873.30
.70	.89	619.5	1.96	274	1003.66
.40	.56	504.3	0.15	183	698.47
.66	.85	465.0	1.33	466	1498.39
.49	.41	328.9	1.56	195	812.50
.37	.76	496.1	0.28	107	382.14

Table 5. Product moment correlation matrix based on the data in Table 4.

	$\delta^2(3)$	$\delta^2(2)$	SDE	NR	AREA
$\delta^2(3)$	1.00				
$\delta^2(2)$	.27	1.00			
SDE	.67	.19	1.00		
NR	.76	.38	.26	1.00,	
AREA	.90	.15	.61	.81	1.00

Table 6. Results of multiple regression analysis. NR = number of rivers, SDE = standard deviation of elevation, AREA = area of phytoclimatic zone, D.V. = dependent variable, I.V. = independent variable, MR = multiple correlation coefficient.

D.V.	I.V.	MR	Equation
$\delta^2(3)$	NR	.76	$\delta^2(3)=30.3+0.11*NR$
$\delta^2(3)$	SDE	.63	$\delta^2(3)=11.6+0.09*SDE$
$\delta^2(3)$	NR SDE	.89	$\delta^2(3)=1.15+0.09*NR+0.06*SDE$
$\delta^2(3)$	NR AREA	.90	$\delta^2(3)=28.7+0.03*NR+0.02*AREA$
$\delta^2(3)$	NR SDE AREA	.92	$d^2(3)=13.6+0.05*NR+0.03*SDE+0.01*AREA$
$\delta^2(2)$	NR	.45	$\delta^2(2)=51.1+0.016*NR$
$\delta^2(2)$	SDE	.38	$\delta^2(2)=39.3+0.053*SDE$
$\delta^2(2)$	NR SDE	.52	$\delta^2(2)=33.6+0.05*NR+0.04*SDE$
$\delta^2(2)$	NR AREA	.59	$\delta^2(2)=52.4+0.13*NR-0.02*AREA$
$\delta^2(2)$	NR SDE AREA	.91	$\delta^2(2)=-7.7+0.2*NR+0.14*SDE-0.05*AREA$

zone, the correspondence is more or less the same. By looking at Table 4, which shows the summary of the Procrustes measurements in 2 and 3 dimensions and geomorphological features, we see that these zones have relatively low density of rivers and relatively low values of the coefficient of variation for elevation (CV). For the other 4 zones, there is a considerable improvement of ordination similarity by adding the third dimension suggesting that for these zones, elevation has important effects on climate.

The correlation between the variables in Table 4 is shown in Table 5. δ^2 (in 3 dimensions) is significantly correlated with AREA and NR. δ^2 (in 2 dimensions) has always lower correlation with the other variables.

The multiple regression results prove that the value of δ^2 is very sensitive to landscape heterogeneity. In the three dimensional case, δ^2 is significantly correlated with NR and if we add either SDE or AREA, the multiple correlation always increases significantly (from 0.76 to 0.90). In two dimensions, the single correlation between the δ^2 value and any other variables is not very

high. However, if we consider NR with SDE and AREA in multiple regression, the multiple correlation coefficient increases to 0.93.

Conclusions

Procrustes analysis proves that there is a strong correlation between the geographical position and climate within the phytoclimatic zones of China and for the whole country. This method can be used to test the effects of landscape heterogeneity with respect to climatic variation. The δ^2 values, as a measurement of ordination dissimilarity, can be used to measure the predictability of geographical position on climate and therefore on vegetation. If the δ^2 value is high, it means that the climatic trends are not very well reflected by geographical position and thus for predicting the vegetation pattern we need to combine the geographical data with climatic data.

References

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York.

Chang, H. S. 1988. The potential evaporation index of vegetation with the relations to the vegetation-climate classification. Acta Phytoecology and Geobotany 13(3):198-206.

Digby, P. G. N. & R.A. Kempton. 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, London.

Gower, J. C. 1971. Statistical methods of comparing different multivariate analyses of the same data. In: F. R. Hodson, D. G. Kendall and P. Tautu (eds), Mathematics in the Archaeological and Historical Sciences, pp. 138- 149. Edinburgh University Press, Edinburgh.

Holdrige, L. R. 1947. Determination of world plant formations from simple climatic data. Science 105:367-368.

Jolliffe, I. T. 1986. Principal Component Analysis. Springer, New York.

Kirby, K. C., M. Pazner and N. Thies. 1989. MAP II version 1.0. Wiley, New York.

Nie, N. H. et al. 1975. SPSS Statistical Package for the Social Sciences. Mc Graw-Hill, New York.

Orlóci, L. 1973. Ranking characters by a dispersion criterion. Nature 244:371-373.

Park, S. K. & K. W. Miller (1988). Random number generators: good ones are hard to find. Communications of the ACM, 31. No.10.

Podani, J. 1991a. On the standardization of Procrustes statistics for the comparison of ordinations. Abstracta Botanica 15:43-46.

Podani, J. 1991b. SYN-TAX IV. Computer programs for data analysis in ecology and systematics. In: E. Feoli and L. Orlóci (eds), Computer Assisted Vegetation Analysis. pp. 437-452, Kluwer, The Hague.

Schönemann, P. H. & R. M. Carroll. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika 35:245-256.

Sibson, R. 1978. Studies in the robustness of multidimensional scaling: Procrustes statistics. J. Royal Statist. Soc. B. 40:234-238.

Sun, Cheng-Yong & E. Feoli. 1991. Numerical phytoclimatic classification of China. International Journal of Biometeorology 35:76-87.

Walter, H. & H. Lieth. 1967. Klimadiagramm-Weltatlas. Gustav Fischer, Jena.