NOTES

Potential Errors in the Application of Principal Component (Eigenvector) Analysis to Geophysical Data

THOMAS R. KARL, ALBERT J. KOSCIELNY AND HENRY F. DIAZ

National Climatic Center, Asheville, NC 28801

10 January 1982 and 29 April 1982

ABSTRACT

Principal component (PC) analysis performed on irregularly spaced data can produce distorted loading patterns. We provide an example to demonstrate some distorted patterns which can result from the direct application of PC analysis (or eigenvector analysis, factor analysis, or asymptotic singular decomposition) on irregularly spaced data. The PCs overestimate loadings in areas of dense data. The problem can be avoided by interpolating the irregularly spaced data to a grid which closely approximates equal-area.

1. Introduction

In recent years meteorologists and climatologists have increasingly used PCs (or their facsimile, eigenvectors) to delineate patterns of temperature, pressure, precipitation, drought, etc. (Gilman, 1957; Kutzbach, 1967; Sellers, 1968; Craddock and Flood, 1969; Dyer, 1975; Klugman, 1978; Walsh and Mostek, 1980; Diaz and Fulbright, 1981; Diaz, 1981; Walsh and Richman, 1981; Brinkman, 1981; Rasmusson *et al.*, 1981). The resulting patterns produced by the PC analyses are often given a physical interpretation and/or are used in subsequent analyses. When irregularly spaced data with systematic biases in the location of the data points are used to produce the PC patterns, then the relation to the actual patterns is likely to be distorted to some extent.

Nonuniformly spaced data are analyzed by Sellers (1968), Klugman (1978), Diaz and Fulbright (1981), Diaz (1981), Walsh and Mostek (1980), Walsh and Richman (1981) and Rasmusson *et al.* (1981). Of these data sets only the Diaz and Fulbright (1981), Diaz (1981) and Rasmusson *et al.* (1981) have obvious systematic biases in the location of the data points analyzed. Because we have worked closely with the Diaz and Fulbright (1981) data set, we have selected it to compare PC patterns derived from irregularly spaced data with regularly spaced gridded data.

2. Data

Patterns are derived from normalized state average winter temperatures across the contiguous United States. These data span 86 winters, 1895–96 through 1980–81. Nearly two-thirds of the data (represented by 31 states) are contained in the eastern half (east of 95°W) of the United States, while only one-third of the data (represented by 17 states) spans the western half (west of 95°W) of the United States. The number of states is more evenly distributed in north–south directions. In order to avoid the overemphasis of the eastern United States (particularly the northeast) the data were interpolated from state centers onto a nearly equal-area grid (Fig. 1). The normalized temperature departures at any grid point were estimated by the weighted average of the temperature from four adjacent states, one in each of the quadrants delineated by north–south and east–west lines through the grid point. The weighting factor for the temperature estimates is the reciprocal of the distance between the state center and grid point. If one or more quadrants contains no data (as is the case for states bordering oceans or other countries) then the estimation involves only the remaining quadrants.

3. Results

Figs. 2–4 depict the analyses derived from both the gridded and non-gridded (state) data for the first three unrotated PCs, orthogonally rotated PCs via the varimax criterion (Kaiser, 1958), as recently described by Richman (1981), and obliquely rotated PCs via the direct oblimin method (Jennrich and Sampson, 1966). The oblimin technique minimizes the function

$$G = \sum_{i \neq j} \left[\sum_{k=1}^{p} a_{ki}^2 a_{kj}^2 - \frac{\Gamma}{p} \left(\sum_{k=1}^{p} a_{ki}^2 \right) \left(\sum_{k=1}^{p} a_{kj}^2 \right) \right], \quad (1)$$

where i and j vary from 1 to m, m is the number of components (factors), p the number of variables, and

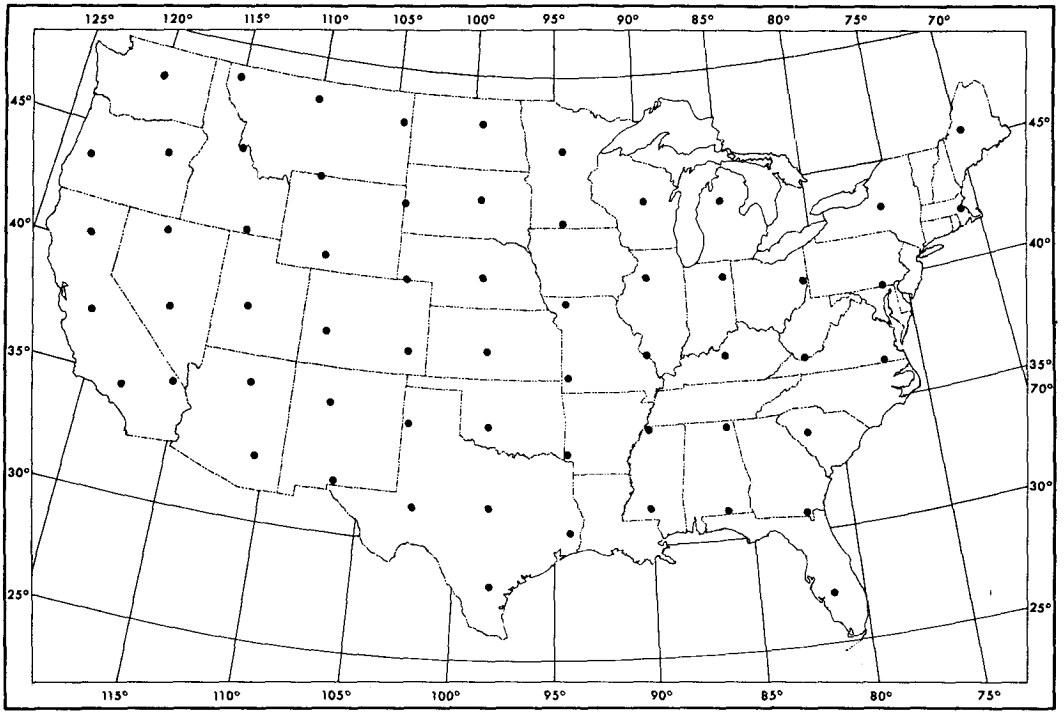

FIG. 1. Grid points used for the PC analysis.

a_{kj} the matrix of component (factor) loadings. The results in Fig. 4 are derived for $\Gamma = 0.25$ based upon an intensive Monte Carlo study to test goodness-of-fit between rotated factors and known input structure (Richman, 1982).[1] For the obliquely rotated PCs the patterns are derived from the structure matrix (**Q**), which depicts the loadings between components and the grid points. As such the variance associated with each component is not unique. This arises because the components are no longer orthogonal. The structure matrix is derived using the intercomponent correlation matrix Φ and the pattern matrix **B** such that

$$\mathbf{Q} = \mathbf{B}\Phi. \tag{2}$$

The pattern matrix identifies the loading (or correlation in our example) between a grid point and a specific component independent of the other components (Rummel, 1970; Johnston, 1978).

Regions are shaded in Figs. 2–4 where approximately half (49%) of the variance is explained by each component. The variance associated with each of the PCs is depicted in the lower left corner of the figures. Since the correlation matrix was used in the PC analysis, the isopleths in Figs. 2–4 can be regarded as the correlation coefficients (loadings) of the grid points with a specific PC. As such the positive (negative) loadings do not mean that above (below) average temperatures are associated with a specific PC. Instead they should be viewed as the weighting on a specific PC or pattern. Positive loadings or correlations (in proportion to their square) are the result of similar responses of standardized temperature departures to the given PC or pattern, while negative loadings are indicative of just the opposite response.

Notice in Fig. 2 that the east–west patterns, $T_s - 1$ (T_s = state temperature, -1 = first component) versus $T_g - 1$ (T_g = gridded temperature, -1 = first component) and $T_s - 2$ versus $T_g - 2$ change more than the north–south pattern, $T_s - 3$ versus $T_g - 3$. The sign reversal in the third PC is not very significant as the relation between the northern and southern United States changes little. However, in the first PC the negative loadings have disappeared in the gridded data set and the center of very high positive loadings has shifted to the central portion of the country. The variance associated with the first PC has also been significantly reduced. In $T_g - 2$, east–west differences have been accentuated as compared to $T_s - 2$, and the variance associated with $T_g - 2$ has been substantially increased.

The first PC changes little from the state to the gridded data for both the orthogonally (Fig. 3) and the obliquely (Fig. 4) rotated PCs. Notice however, the patterns of $T_g^r - 2$ and $T_g^r - 3$ (Fig. 3) are not properly captured in the patterns depicted by $T_s^r - 2$ and $T_s^r - 3$. For example, the high positive load-

[1] Personal communication.

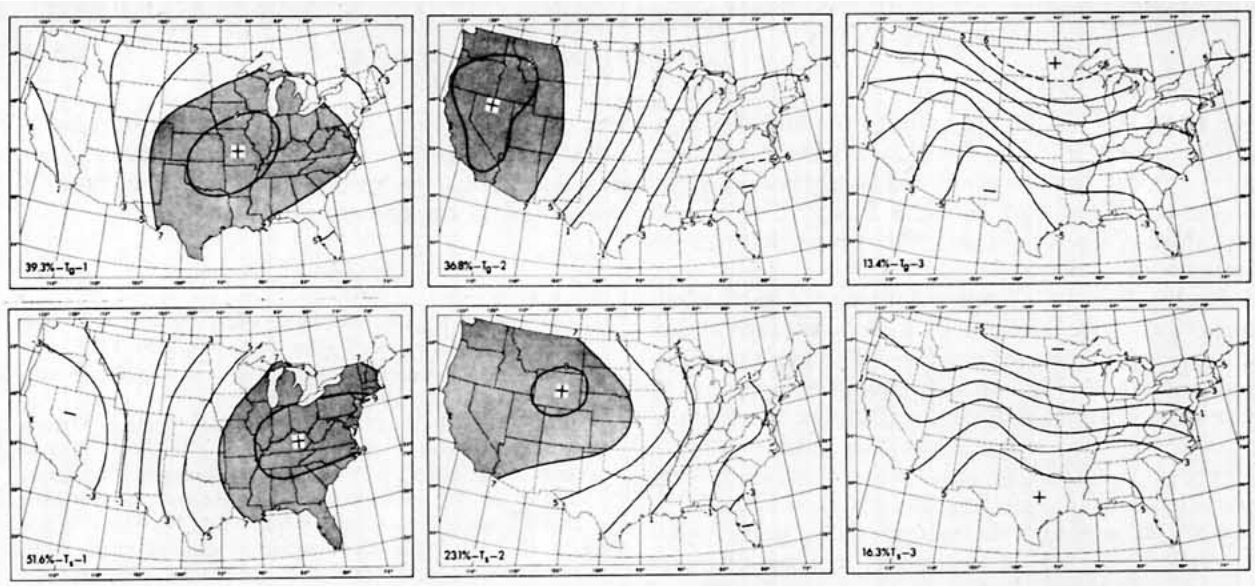

FIG. 2. Unrotated PCs (×100) which were derived from directly analyzing state average winter temperature data (T_s) and gridded data (T_g).

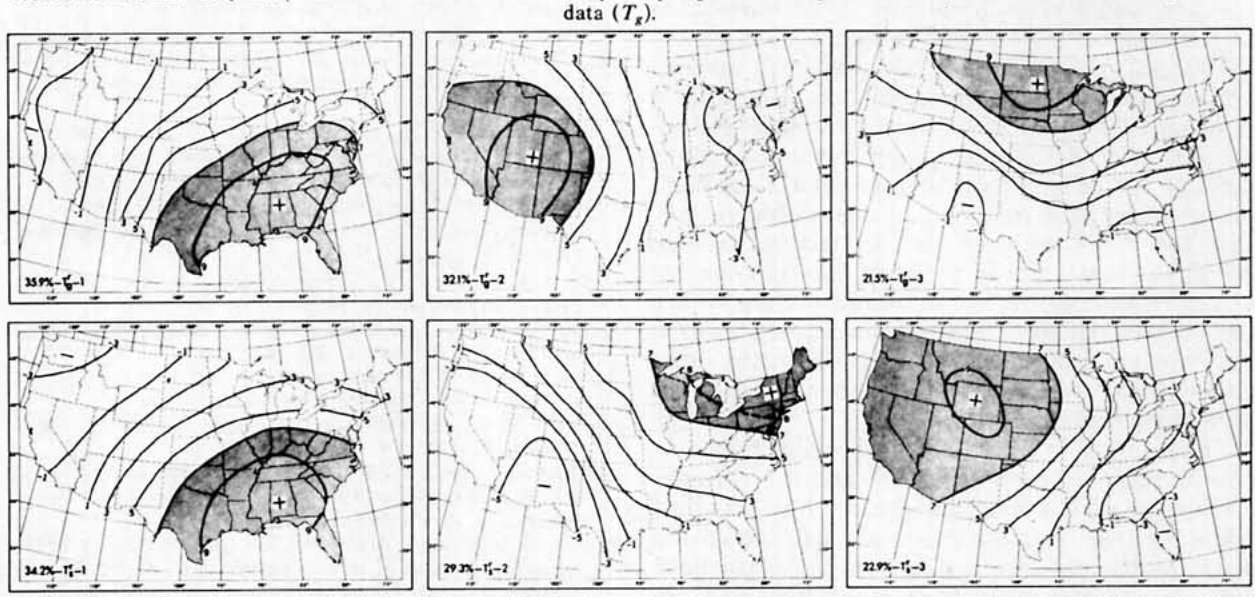

FIG. 3. Orthogonally rotated PCs (×100) which were derived from directly analyzing state average data (T'_s) and gridded data (T'_g).

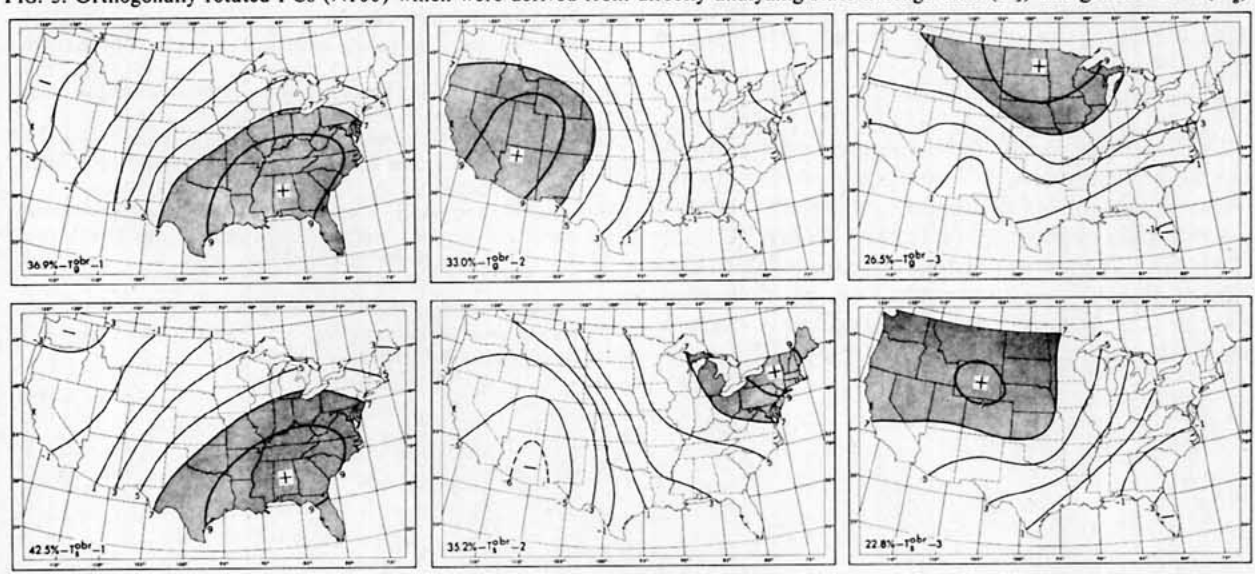

FIG. 4. Obliquely rotated PCs (×100) which were derived from directly analyzing state average data (T_s^{obr}) and gridded data (T_g^{obr}). Structure loadings are depicted.[2]

[2] The superscript "obr" is used to identify obliquely rotated components.

ings of $T_s^r - 3$ are located between the high loadings of $T_g^r - 2$ and $T_g^r - 3$. The patterns in $T_s^r - 2$ and $T_s^r - 3$ fail to capture the broad homogeneous areas of the southwest and northern Plains. The second and third obliquely rotated PCs have similar problems.

Buell (1971, 1978) points out that for application to meteorological analysis, PCs are better formulated in terms of an integral equation. The numerical approximation to the integral equation, appropriate when the covariance function is known empirically at discrete points in space, contains quadrature factors proportional to the areas represented by the observation points. Applying PC analysis directly to irregularly spaced data incorrectly assumes equal quadrature factors. Hence, direct analysis of irregularly spaced data leads to improper weighting of the significance of each variable in the correlation (or covariance) matrix used in the PC analysis. In PC analysis (or eigenvector analysis) each variable is given equal weight. In factor analysis each variable is weighted in terms of its shared variance. No account is made of the geographical distribution of the data. If the geographical area of interest contains one or more localized areas of dense observations, then the PCs (or eigenvectors or factors) tend to reflect the variability of the local data-rich area(s) at the expense of the larger overall pattern. When the irregular data is interpolated to a equal-area grid, the importance of all the variables across the entire geographical area is properly characterized in the correlation (covariance) matrix.

The decision to use normalized winter temperatures in this example is somewhat arbitrary. Higher loadings would be obtained in areas of high standard deviations (the Northern Plains in particular) if the derived patterns were obtained from non-normalized data (the covariance matrix). The differences are comparable in magnitude to the differences between the irregularly and regularly spaced data used in Figs. 2–4.

4. Conclusion

The intent of many meteorological and climatological applications of PC analysis is to identify the predominant modes of spatial variability across a given domain of geographical extent. When PC analysis is directly applied to irregularly spaced data then the variability (or homogeneity) of the element in question for localized data-rich regions is overemphasized at the expense of other regions. This problem can be overcome by interpolating the data to an equal-area grid. PC analysis performed on the gridded data will then properly delineate the patterns within the spatial domain of interest.

REFERENCES

Brinkmann, W. A. R., 1981: Sea level pressure patterns over eastern North America. *Mon. Wea. Rev.*, **109,** 1305–1317.

Buell, C. E., 1971: Integral equation representation for factor analysis. *J. Atmos. Sci.*, **28,** 1502–1505.

——, 1978: The number of significant proper functions of two-dimensional fields. *J. Appl. Meteor.*, **17,** 717–722.

Craddock, J. M., and C. R. Flood, 1969: Eigenvectors for representing the 500 mb geopotential surface over the Northern Hemisphere. *Quart. J. Roy. Meteor. Soc.*, **95,** 576–593.

Diaz, H. F., 1981: Eigenvector analysis of seasonal temperature, precipitation and synoptic scale system frequency over the contiguous United States. Part II: Spring, summer, fall and annual. *Mon. Wea. Rev.*, **109,** 1285–1304.

——, and D. C. Fulbright, 1981: Eigenvector analysis of seasonal temperature, precipitation and synoptic-scale frequency over the contiguous United States. Part I: Winter. *Mon. Wea. Rev.*, **109,** 1267–1284.

Dyer, T. G. J., 1975: The assignment of rainfall stations into homogeneous groups: an application of principal component analysis. *Quart. J. Roy. Meteor. Soc.*, **101,** 1005–1013.

Gilman, D. L., 1957: Empirical orthogonal functions applied to thirty-day forecasting. Sci. Rep. No. 1, Contract AF19(604)-1283, MIT, 129 pp.

Jennrich, R. I., and P. F. Sampson, 1966: Rotation for simple loading. *Psychometrika,* **31,** 313–323.

Johnston, R. J., 1978: *Multivariate Analysis in Geography.* Longman Group Ltd., London, 127–183.

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

Klugman, M. R., 1978: Drought in the upper Midwest. *J. Appl. Meteor.*, **17,** 1425–1431.

Kutzbach, J. E., 1967: Empirical eigenvectors of sea level pressure, surface temperature and precipitation complexes over North America. *J. Appl. Meteor.*, **6,** 791–802.

Rasmusson, E. M., P. A. Arkin, W. Chen and J. B. Jalickee, 1981: Biennial variations in surface temperature over the United States as revealed by singular decomposition. *Mon. Wea. Rev.*, **109,** 587–598.

Richman, M. B., 1981: Obliquely rotated principal components: An improved map typing technique? *J. Appl. Meteor.*, **20,** 1145–1159.

Rummel, F. J., 1970: *Applied Factor Analysis.* Northwestern University Press, Evanston, IL, 617 pp.

Sellers, W. D., 1968: Climatology of monthly precipitation patterns in the western United States, 1931–1966. *Mon. Wea. Rev.*, **96,** 585–595.

Walsh, J. E., and M. B. Richman, 1981: Seasonality in the associations between surface temperatures over the United States and the North Pacific Ocean. *Mon. Wea. Rev.*, **109,** 767–783.

——, and A. Mostek, 1980: A quantitative analysis of meteorological anomaly patterns over the United States, 1900–1977. *Mon. Wea. Rev.*, **108,** 615–630.