On the Effects of Preanalysis Standardization
Author(s): Ilse M. Hamann and Ute C. Herzfeld
Source: *The Journal of Geology*, Vol. 99, No. 4 (Jul., 1991), pp. 621-631
Published by: The University of Chicago Press
Stable URL: https://www.jstor.org/stable/30065006
Accessed: 10-05-2024 15:20 +00:00

GEOLOGICAL NOTES

ON THE EFFECTS OF PRE-ANALYSIS STANDARDIZATION[1]

ILSE M. HAMANN AND UTE C. HERZFELD[2]
NOAA National Marine Fisheries, 2570 Dole St., Honolulu, HI 96822-2396, USA

ABSTRACT

In spatial analysis the data are often assumed to be normally distributed, which may not be a property of the earth science phenomenon studied. Different types of pre-analysis standardization lead to disparate qualitative or quantitative interpretations: The standardization techniques compared are Gaussian normalization and proportion of range transformation. For two case studies: (1) oceanographic data investigated with Q-mode factor analysis, and (2) stratigraphic data analyzed by algebraic map comparison, proportion of range standardization is more appropriate than Gaussian standardization, in view of the scientific interpretation, with the results of factor analysis generally being more sensitive to pre-analysis standardization than the results of map comparison.

INTRODUCTION

In data analysis a particular problem arises as soon as more than one parameter needs to be investigated simultaneously, a typical situation when studying a complex earth science system. Data of different dimensions are encountered, which have to be considered on a common scale. A careless choice regarding standardizations breeds distortions and uncertainty with respect to the meaning of the results. Standardization is often understood as synonymous with z-score or Gaussian standardization (transformation of observed data to a mean of zero and a standard deviation of one). The parameters "mean" and "standard deviation," however, only make sense if the data show a Gaussian ("normal") distribution. In social sciences, economy, or ecology, this is an assumption that holds in many situations but not so much in geology or oceanography, a fact that is usually ignored. For example, in stratigraphy, the elevation of the base of a formation that cuts through the survey area is more likely uniformly distributed between minimum and maximum elevation than normally distributed. Computations based on the wrong assumptions concerning the frequency distribution of the variables may give rise to unreasonable oceanographic or geologic results. We use the term "standardization" in the (original) meaning of a transformation that makes several different variables comparable. For simplicity, such a transformation maps the original variable values into one of the intervals [0,1] or [−1,1].

In oceanography, many studies aim at explaining deviations from some "mean" state of the system. The data may, for example, be a set of variable values in a time-series or may originate from a synoptic spatial survey. In such applications the processes of interest often interact and produce superimposed responses, which constitute the measurable variables. The distributions of the responses are usually assumed to be normally distributed in the ensuing data reduction procedures, many of which are of a spectral or multivariate statistical nature. It is important to note, however, that the data transformations some of these analytical tools justify are not at all appropriate if the principal interest lies with the extreme values and the processes that create them; therefore the extremes should not be discarded beforehand as "outliers." There is a tendency in the current practice of statistical analysis and interpretation to suppress such data, either by omitting them or weighting them down to insignificance. Unless they are observational errors (a case we neglect here for simplicity), extreme values represent real information.

An oceanographic method that places more importance on the extremes rather than on the deviations from the mean is the use of Q-mode factor analysis (QMFA) to extract endmember-type components for a geochem-

[1] Manuscript received September 13, 1990; accepted January 17, 1991.

[2] Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093-0215, USA.

[JOURNAL OF GEOLOGY, 1991, vol. 99, p. 621–631]

0022-1376/91/9904-008$1.00

ically mixed oceanic system. Until now, this method was mainly used in geology as an aid in sedimentological and petrological mixing scenarios (e.g., Miesch 1976). The component associations lie at the end of a straight mixing line in property-property plots; hence the name "endmember." In example 1 below, the dependency of the results of this method on pre-analysis standardization is demonstrated. In a similar way, a second method of multivariate spatial data analysis is scrutinized with respect to its sensitivity to standardization. Thematic map comparison techniques aim at retaining spatial information when investigating multivariate relationships among different types of geological, geophysical, and geochemical data sets (e.g., Merriam and Jewett 1988; Herzfeld and Merriam 1990). The examples chosen are (1) the determination of water mass factors (called WMFs from here on) and the mapping of relative similarity of water samples with the WMFs, and (2) algebraic map comparison of geological and morphological data sets.

STANDARDIZATION CONCEPTS: PROPORTION OF RANGE STANDARDIZATION VERSUS GAUSSIAN STANDARDIZATION

(1) Proportion of Range.—If $xmin_j$ and $xmax_j$ are the minimal and maximal values observed on the variable x_j (for $j = 1, \ldots M$, M the number of variables), and x_{ij} are observations of x_j (for $i = 1, \ldots, N_j$, N_j the number of observations of x_j, e.g., water samples or geographical locations), then the proportion of range standardization z_p is defined by:

$$z_p(x_{ij}) = \frac{x_{ij} - xmin_j}{xmax_j - xmin_j} \tag{1}$$

If one variable generally increases while most of the others decrease, then inverse proportion of range z_{p-} is used:

$$z_{p-}(x_{ij}) = \frac{xmax_j - x_{ij}}{xmax_j - xmin_j} \tag{2}$$

Note that $z_{p-}(x_{ij}) = 1 - z_p(x_{ij})$. Both proportion of range transformations map the observed data range into the interval [0,1].

(2) Gaussian Standardization.—The standardized data value $z_g(x_{ij})$ in the sense of Gaussian statistics is obtained by

$$z_g(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{x_{j,\mathrm{SD}}} \tag{3}$$

where $\bar{x}_j = 1/N \sum_{i=1}^{N} x_{ij}$ and $x_{j,\mathrm{SD}}$ denotes the standard deviation of variable x_j, computed as $x_{j,\mathrm{SD}} = (1/N \sum_{i=1}^{N} x_{ij}^2 - \bar{x}_j^2)^{1/2}$. In the set of standardized data (normalized data) each variable $(z_g)_j$ has a sample mean of zero and a sample standard deviation of one. The minimum and maximum values of the standardized data are not known in this case. Instead, if the data satisfy the Gaussian hypothesis, then 68% of the transformed data are expected to lie in the interval $[-1,1]$, and 95.5% in the interval $[-2,2]$ (cf. Kreyszig, 1970, p. 110). The Gaussian standardization assigns a special meaning to the mean and the standard deviation, and weights data according to the relation to those values. In contrast, the proportion of range transformation treats all data equally. In statistical language, the proportion of range transformation is associated with the uniform distribution. The actual distribution encountered in the data is usually unknown or very complex in geosciences. In such cases it is often a better approximation to assume a uniform rather than a Gaussian distribution. A histogram, for example, displays the distribution of sample values over their range but it dispenses with the regional component in the spatial data which is of principal interest as shown in the examples here.

EXAMPLE 1: DETERMINATION AND LOCALIZATION OF OCEANOGRAPHIC WATER TYPE ASSOCIATIONS USING Q-MODE FACTOR ANALYSIS (QMFA)

Problem Outline and Method of Factor Analysis.—In traditional geochemical and oceanographic mixing analysis, property-property plots are often used as guides to find characteristic water types that constitute the main ingredients of the mixture. We avoid calling the factors resulting from the analysis simply "water types" or "water masses," because these have strict definitions that do not exactly apply to the quantities we obtain (cf. Mamayev 1975 and Hamann and Swift 1991). The distances from the endpoints of linear portions of these property-property curves are inversely related to the proportion that the water type at the endpoint of the curve contributes to the mixture (e.g., Mamayev 1975, p. 120). If little is known about

the actual composition of the source waters, however, factor analytical techniques may be a better first step in water mass analysis because they do not need an a priori specification of contributing water types. Factor analytical computations consist of eigenanalysis algorithms of linear algebra as outlined in Joreskog et al. (1976). A data matrix X_{NxM} (N = number of samples, M = number of variables) is factored into a matrix F_{Mxp} ($p \leq M$) of scores (the WMFs) and a loadings matrix A_{Nxp} containing the weights that each of the p factors has in each of the N samples. Matrix F contains the rotated eigenvectors of the symmetric similarity matrix H_{NxN}—the major product moment of the scaled (standardized) and row-wise L_2-normalized data matrix W_{NxM} (a vector $x = (x_i, \ldots, x_n)$ is L_2-normalized if divided by its Euclid norm, which equals $(\Sigma_{i=1}^{N} x_i^2)^{1/2}$). The QMFA computations in example 1 were done using code CABFAC, listed in Klovan and Imbrie (1971). Results of this application here consist of a classification of oceanic water type associations on the basis of intervariable ratios found in the water samples. With QMFA one can determine locations of water sample vectors that are extremely dissimilar with respect to this criterion, i.e., these hypothetical sample vectors or factors depend on the eccentricities of the data set. Two sample vectors v_1 and v_2 are completely dissimilar if their scalar product equals zero:

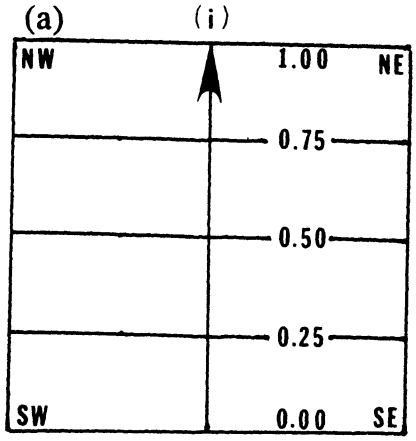
$$(v_1, v_2) = \cos(<v_1, v_2)\|v_1\| \, \|v_2\| = 0,$$

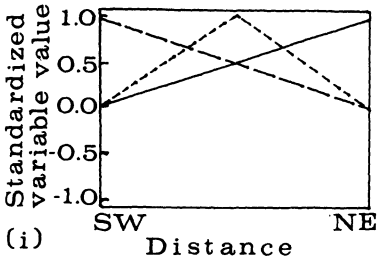
and completely similar if their scalar product equals one. The factor loadings can be mapped out that represent quantitative mixing proportions if we are dealing with a "closed" data set. In this example, however, the data set is "open," i.e., the sums of variable values in a sample do not equal the same constant, and consequently the loadings merely express the similarity of all samples with the water mass factors rather than mass conserving mixing proportions. In contrast, so-called R-mode factor analysis (RMFA) finds factors representing composite variables indicative of a common influence from some underlying natural process. These factors are determined on the basis of correlating variables. The Q-mode approach is independent of the degree of linear dependence between the variables. In fact, QMFA easily finds completely disjunct clusters (in variable space) that differ with respect to their intervariable ratio. In such cases the correlation coefficient is an invalid measure of similarity or linear dependence between the variables, and the use of RMFA is inappropriate. In general, in RMFA the exact spatial coordinate of the contribution to the similarity between variables and factors is lost. QMFA, however, identifies locations that represent unique mixing ingredient factors irrespective of what the correlation between the variables might be.

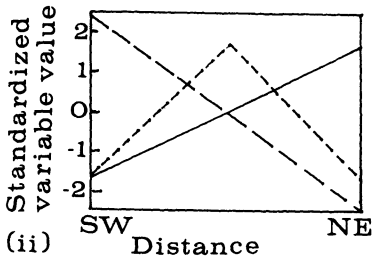
Prior to the analysis, the data need to be standardized. In the following we show that the water bodies found by the classification depend heavily, both in composition and spatial distribution, on the standardization. First the proportion of range transformation is applied, and the results are then compared with those obtained for input data that have been standardized assuming a Gaussian distribution (see eqns. 1 and 3).

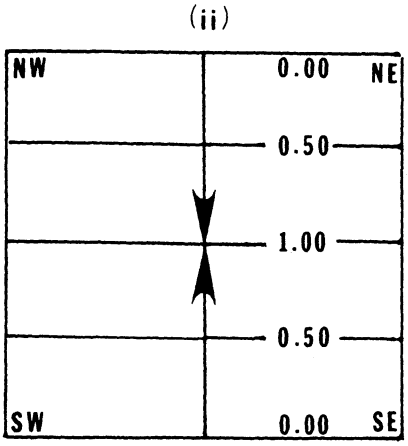
Numerical Example.—Consider a two-dimensional spatial domain in which the magnitude of three simulated variables (assumed to be error-free) is known for 676 samples (26 in the northern and 26 in the eastern direction, $N = 676$, $M = 3$). These synthetic variables may, for example, represent silicate, phosphate, and temperature (or salinity) on a density surface in the upper thermocline in the Pacific Ocean. Their spatial distribution in plane view, along a southwest-to-northeast transect, as well as histograms of the variable distributions are shown in figure 1. The gradients of the first two variables are meridional; that of the third variable points toward the southwest. The histograms are uniform for the first two and unimodal for the third variable (fig. 1*c*). The modality of the latter follows directly from the selected geometry of the spatial sampling area and has no real physical meaning (cf. example 2 below). Figure 2 shows property-property plots of variable 2 vs. variable 1 after both transformations [1] and [3] and before and after they have been L_2-normalized. Figure 3 displays the factor analytical results, i.e., maps of factor loadings or "pseudo-proportions" from reference vectors for both types of preanalysis data standardizations.

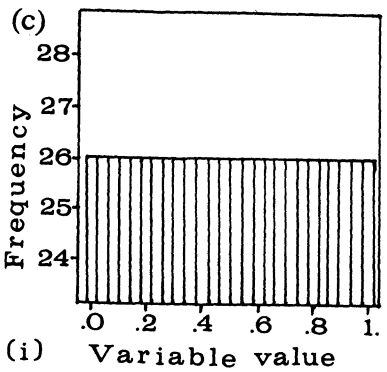
Samples that represent reference vectors are found at the vertices of property-property plots. The multi-variable nature of these typi-

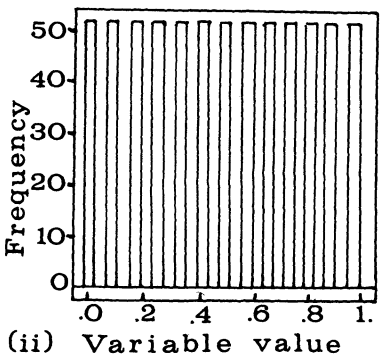


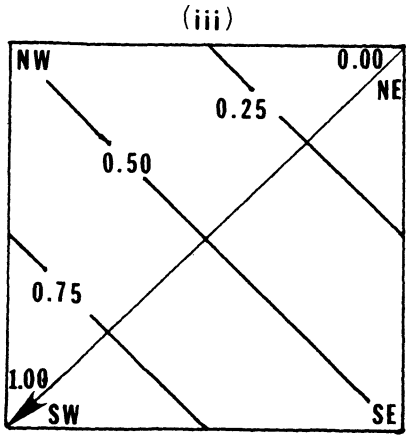


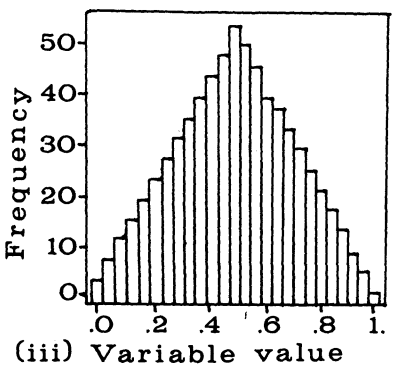












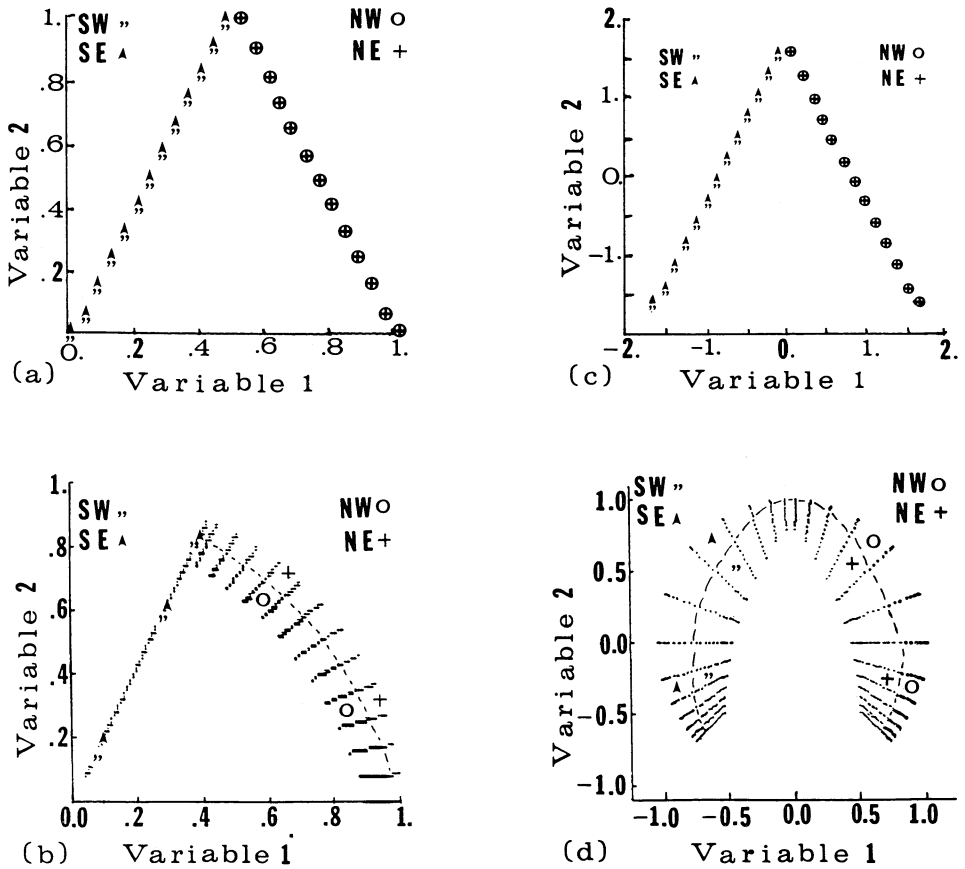


FIG. 2.—Property-property plots of variable 2 vs. variable 1. (*a*) proportion of range data; (*b*) same, but L_2-normalized; (*c*) Gaussian-standardized data; (*d*) same, but L_2-normalized. The symbols indicate the geographic occurrence of variable values: circle = northwestern quadrant, plus = northeast, quotes = southwest, arrow tip = southeast.

cal samples is reflected in the factor scores diagrams next to the factor loadings maps (fig. 3*a*). Positive (negative) values for a variable's score on a reference factor signify above (below) average property values in the ingredient source water. The rank of the data matrix determines the number of WMFs of the factor model. For the proportion-of-range transformed data three factors are found characterized by a high individual proportion of variable 1, 3, and 2 in the first, second, and third water mass factors, respectively (labeled F1, F2, and F3 in fig. 3*a* i, ii, iii). Loadings maxima occur where one (group of) variable(s) dominates the rest, exemplified by the positively scoring variables. Maxima are found in the eastern part of the northern margin, the western part of the southern boundary, and the eastern equatorial region where the first, third, and second variables dominate over contributions from the remaining two variables. High factor loadings values are to be expected for samples (locations) that form the tips of line or curve segments that near the axes in the two-dimensional property-property plots (fig. 2*a* and *b*). The

FIG. 1.—(*a*) Spatial distribution of synthetic data (plane view). (i) silicate (variable 1), (ii) phosphate (variable 2), (iii) temperature (variable 3). (*b*) Standardized synthetic data along a southwest to northeast diagonal transect through the domain (continued line—variable 1; short dashed line—variable 2; long dashed line—variable 3). (i) proportion of range transformation; (ii) Gaussian standardization. (*c*) Histograms of variables transformed to proportion of their range. (i) variable 1, (ii) variable 2, (iii) variable 3.

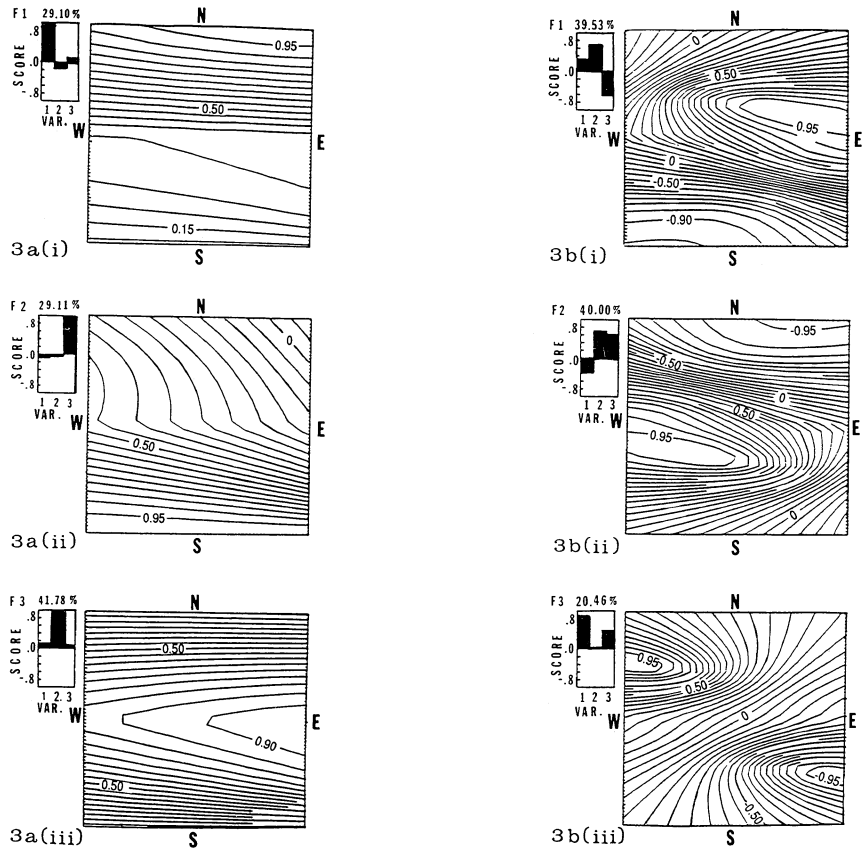


FIG. 3.—Spatial distribution of L_2-normalized varimax factor loadings. (*a*) of proportion of range data; (*b*) of Gaussian-standardized data (i), (ii), (iii) water mass factors 1, 2, 3 respectively. Above the bar-diagrams of normalized factor scores the percentage of the total compositional information is given.

isolines of figure 3 do not signify additive mixing proportions, but fractional units of similarity whose squares add up to unity, the total compositional information content of each sample. (While in previous applications of QMFA to "closed" data sets, in which the variables add to the same constant in each sample, truly quantitative mixing proportions can be obtained by re-scaling the factor loadings; this is not possible for this particular, heterogeneous oceanographic data set where the original variables have incompatible measurement units [see, for example, Miesch 1976]). The numerical values indicated for both the factor scores and loadings (numbers on contours) in figure 3 are L_2-normalized, attributing each reference factor or individual sample vector with a compositional variability equal to one. Imbrie and van Andel (1964, p. 1138) explain what is meant by compositional variability: it is supposed that each sample has an information content of 1, which is based on the proportions between the variable values measured for the sample. Numerically, this number is obtained by summing the squares of the variable values of a sample, i.e., the elements of a row vector of W, the standardized data matrix, whose L_2-norm equals unity. The total information content in the data set therefore equals N. The fractional compositional variability for the whole data set associated with a factor is calculated in relation to this quantity (see percentage values over the factor scores diagrams in fig. 3). Note that this is a different kind of variability than the "variance" of R-mode methods where the summation of all deviations from the mean is over all samples.

The performance of Q-mode analysis of the (Gaussian) standardized data set is examined next. While the property-property plot of the non-normalized data looks similar to that of the proportion of range data (compare fig. 2*a* and *c*), the L_2-normalized plot of the (Gaussian) standardized data features no obvious vertices that would aid the identification of dissimilar reference samples (fig. 2*b* and *d*). Consequently, there exists no unique dominance of any single variable in any of the three endmember associations, i.e., high absolute values of more than one factor score occur (see small diagrams next to fig. 3*b*). The interpretation of the endmembers becomes less "simple," which is considered a definite disadvantage (Harman 1976, p. 98, reviews criteria for "simple structure" in R-mode factor analysis, which were originally formulated by Thurstone 1947). The locations of extremes, indicative of most dissimilar inter-variable ratios, are offset by a considerable distance from the property extremes on the northern and southern boundary and the equator. Completely different maximum/minimum loadings regions are detected whose meaning can only be explored by a convoluted argument that involves more than one variable. It is particularly disturbing that the spatial gradients in these factor loadings distributions deviate by a large angle from that obtained for the proportion-of-range transformed data (compare panels of fig. 3*a* with those of 3*b*). Consequently, the proportion of range standardization is preferable because we want to maintain and make obvious a close link between extremes of small subsets of variables pertaining to some phenomenon that creates distinctive inter-variable ratios.

EXAMPLE 2: THEMATIC COMPARISON OF GEOLOGICAL AND MORPHOLOGICAL MAPS

Algebraic map comparison methods are used as an aid to quantify relationships between several variables and localize regions of similarity. The subject of map comparison is of considerable practical interest, since different types of maps are necessary virtually everywhere in geosciences. Exploration geology, for instance, uses map comparison in "favorability mapping": to determine potential future exploration sites, data from geological, geophysical, geochemical, and petrographical surveys are combined, which are supplied by the various investigators in the form of maps.

An overview of computer based comparison methods is given in Merriam and Jewett (1988). All of these methods require standardization of the input maps. Merriam and Sondergard (1988) noted that the significance of their correlation coefficient is slightly better if proportion of range standardization is used instead of Gaussian standardization. When results of their reliability index (a moving window local correlation coefficient) maps were compared, the patterns of the comparison map remained similar, but varied in location, depending on the standardization method chosen. We use an algebraic map comparison method that permits any number of input maps and results in a comparison that is itself a map, showing areas of joint co-variation vs. areas of dissimilarity (Herzfeld and Merriam 1990; Herzfeld and Sondergard 1988).

Method of Algebraic Map Comparison.—The algebraic map comparison algorithm used is based on the calculation of weighted averages between standardized values of any pair of input maps, for each point of the map area (in practice, for each node of a common underlying grid). Let M denote the map area, n the number of input maps, and $M_1, \ldots, M_n$ the input maps with $m_k(x)$ the (standardized) value of map M_k at location x in the map area. Then the comparison map F is defined by:

$$F(x) = \frac{\sum_{s<t,t=1}^{n} w_s w_t |m_s(x) - m_t(x)|}{\sum_{s<t,t=1}^{n} w_s w_t} \quad (4)$$

where w_k is a weight assigned to map k (with $1 \le k \le n$). The denominator is the number of comparisons, if all maps are weighted equally by 1.

$F(x)$ is a distance measure (a semi-norm in a space of dimension of the number of pairwise comparisons possible). Consequently areas of low values indicate high similarity, while high values mean poor relationships between the input maps. If proportion of range-standardization is used, then the values of

$F(x)$ are in the interval [0,1]. For Gaussian standardization, this is obviously not the case.

Geological Problem and Data.—From an area in south-central Kansas, geological and morphological maps were chosen. To demonstrate a prediction or model-searching problem, we consider the depth of the Precambrian sedimentary surface, the topography, and the residual thickness of sedimentary layers between the Precambrian surface and the topography. Thus the relationship is known in principle, and regional deviations can be estimated from a visual comparison.

The structure of Kansas geology is fairly simple: the stratigraphic layers in south-central Kansas lie conformally on top of each other, all dip west to west-southwest and consequently outcrop with from east to west in decreasing age. The subsea structure on top of the Precambrian surface (fig. 4*a*) shows a prominent southward plunging anticlinal feature (the Nemaha Anticline) with a parallel southerly plunging syncline just to the east. The configuration west of the anticline is a series of northeast-trending anticlinal and synclinal features of lesser magnitude, the map subarea shows mainly the general dipping trend of the Precambrian surface. The present day topography (fig. 4*b*) dominantly shows the Flint Hills scarp trending north-south in the eastern third of the map (Merriam and Sondergard 1988). The residual thickness (fig. 4*c*) consequently increases from the eastern map edge toward the west and the southwest. The highest gradients occur in locations of the Nemaha Anticline and the Flint Hill escarpments. In our example, we want to find areas where the maps enhance each other versus areas of anomalies. The anomalies should be caused by relief, which appears, geometrically, as deviation from the tilted plane, and algebraically, as deviation from linearity.

The data used are digitized from published maps: Precambrian (Cole 1962), present-day topography (U.S. Geological Survey 1963). The grid refers to the U.S.G.S. grid of townships and ranges, the area lies between T.18S and T.34S and between R.4W and R.12E. Digitization of 17 by 17 grid nodes was done on a 6 mile intergrid distance. The grid is coarse but sufficient for the illustrative purpose here.

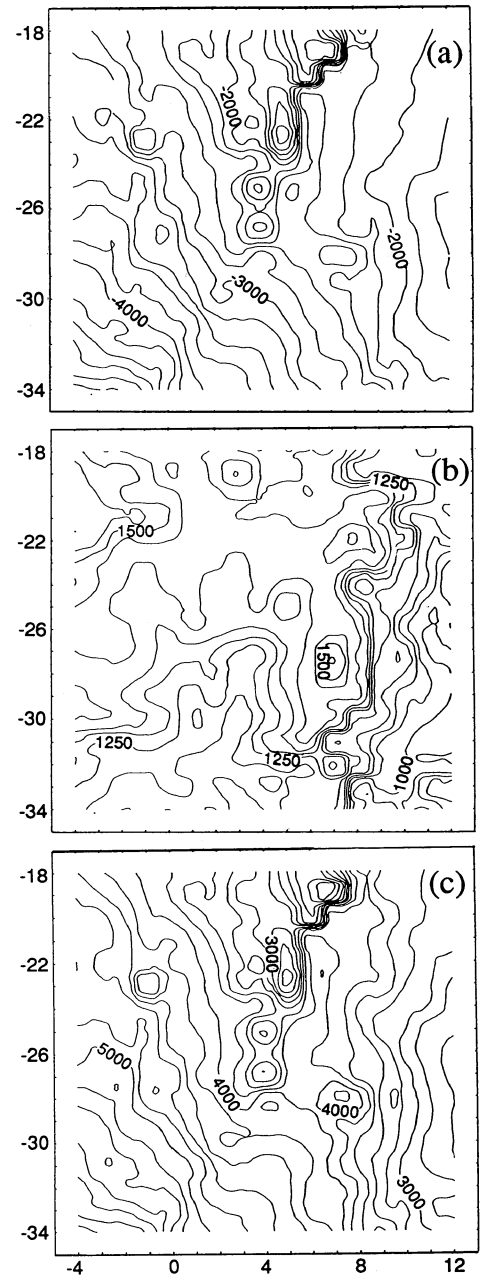


FIG. 4.—Geological and morphological maps from south-central Kansas. The grid refers to the U.S.G.S. grid of townships and ranges, area is 6 miles by 6 miles (1 mile = 1580 m). (*a*) Subsurface elevation of the Precambrian in [m] (top of the formation, digitized after Cole 1962). (*b*) Topography in meters (digitized after U.S.G.S. 1963). (*c*) Total thickness in meters of sedimentary layers above the top of the Precambrian [difference of (a) and (b)].

Histograms.—Histograms, calculated from the grid node values, are shown for the three input maps in figure 5 (*a*—Precambrian, *b*—topography, *c*—total thickness). The topography histogram is close to a uniform distribution. The histograms of Precambrian and total thickness have a unimodal shape: the histogram of the Precambrian has a maximum at −2000, and minima at the lower (−5000) and the higher (−1000) end of its data range. This does not, however, imply that the surface follows a Gaussian distribution, but is rather due to the geometry of the mapping window. Disregarding the Nemaha Anticline, the top of the formation follows a west-southwest- to west-dipping trend surface, and the shape of the histogram results from the selection of the subarea: As is obvious from the map, the elevation of the formation decreases linearly in the direction of the dip, while it is constant along its striking. Data on the trend surface would correspond to a pure uniform distribution. A subarea, rectangular with edges parallel to dip and strike, would produce a uniform histogram, while a rectangular one rotated by 45° with respect to dip and strike will produce a unimodal shape (compare the histogram and situation in variable 3 in example 1). Similar considerations hold for the total thickness, as the absolute range of the topography is much less than that of the Precambrian surface. Thus the histogram is not very meaningful, geologically.

Comparison of Similarity Results Depending on Standardization Methods.—The map of comparison after Gaussian standardization (fig. 6*a*) shows dominating features (maxima of 2.4, 2.2, 1.0, 0.8) following the Nemaha Anticline and extending farther southward. The Nemaha Anticline shows well on two input maps. The Flint hill escarpment does not seem to leave a noticeable imprint on the comparison map. Instead, values increase to maxima along the eastern map edge (up to 2.2). Another maximum (> 3) appears in the southwest corner, where no relief is found. The locations of the maxima are therefore generally not consistent with the relief in the Precambrian and surface layers or with gradients in thickness.

The pattern of the map based on proportion of range standardization shows subtle differences in shape (fig. 6*b*). Maxima are also located along the Nemaha anticline, the eastern

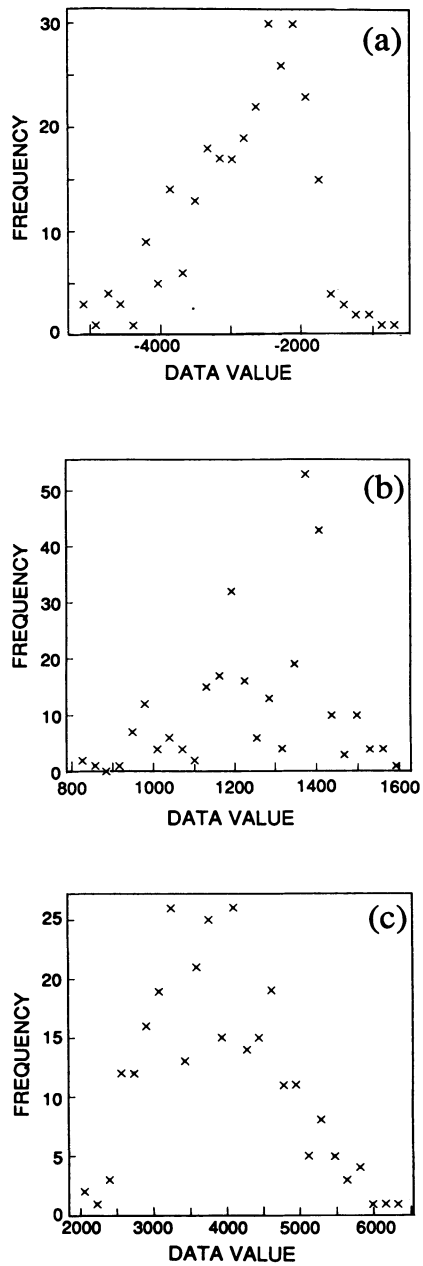


FIG. 5.—Histograms of geological and morphological maps. (*a*) Histogram of subsurface elevation of the top of the Precambrian (fig. 4*a*). (*b*) Histogram of the topographical elevation (fig. 4*b*). (*c*) Histogram of thickness of layers between Precambrian and topography (fig. 4*c*).

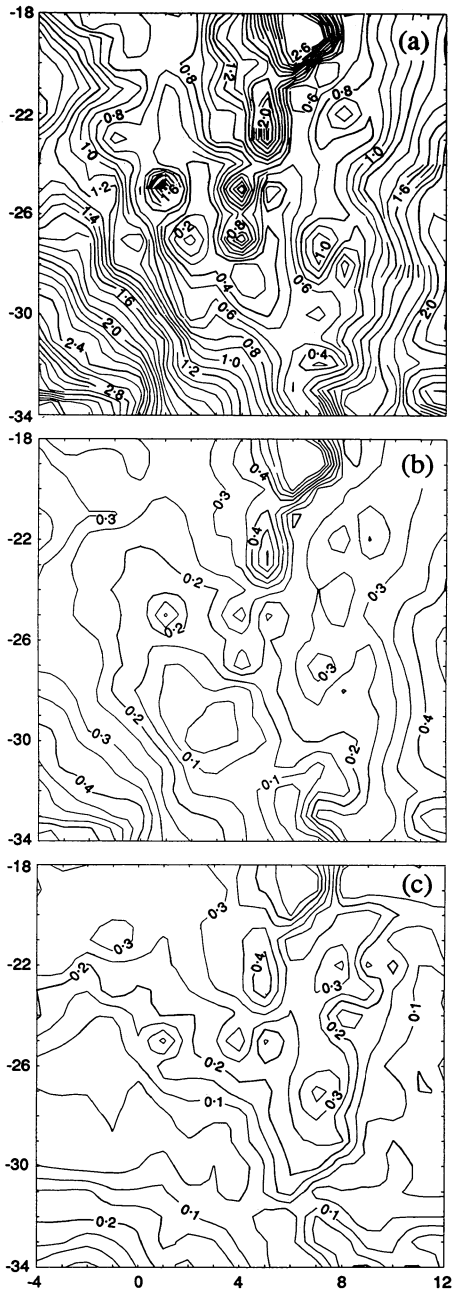


FIG. 6.—Comparison maps, depending on different standardizations: Precambrian subsurface elevation (fig. 4*a*) vs. topography (fig. 4*b*) vs. total thickness (fig. 4*c*). (*a*) Similarity using Gaussian standardization of all three variables (contoured with distance of 0.1). (*b*) Similarity using proportion of range-standardization of all three variables (contoured with distance of 0.05). (*c*) Similarity using proportion of range-standardization of Precambrian, and inverse proportion of range of both topography and total thickness (contoured with distance of 0.05). The grid refers to the U.S.G.S. grid of townships and ranges. Low values indicate good similarity. To facilitate a visual comparison of the two standardization methods, similarity maps based on proportion of range standardization (equations 1 and 2) have a 0.05 contour interval, while in those based on Gaussian standardization (eqn. 3) an isoline distance of 0.1 was used. This should account for the fact that proportion of range maps into [0,1], while normalization has the standard interval [−1,1]. But at first glance one notes that the isolines in figure 6*a* are much denser than on 6*b* and 6*c*, and that the Gaussian type yields higher values. Since the values are not bounded, it is harder to get a feeling for good and poor matches, as is possible with the proportion of range transformed input.

edge, and the southwest corner, but the gradients in the comparison map are not nearly as strong in the latter two locations. Other than in the first case (fig. 6*a*), an area of good covariation extends in the southwest quadrant of the map (an area where the three input maps do not show much variation). This result is slightly better than that in figure 6*a* with respect to the similarity pattern. Its major advantage is the comparability of the scale (see caption of fig. 6*a*).

In the last comparison (fig. 6*c*), topography and total thickness have been scaled inversely (equation 2), and Precambrian with ordinary proportion of range (equation 1), according to the trends of the surfaces: The Precambrian surface dips west to west-southwest as typical for most stratigraphic layers in the area, while the topography generally decreases toward the east, and a little to the south. The residual thickness, being the difference of the former two, decreases toward the east. Figure 6*c* also shows the Nemaha Anticline location (maximum values of 0.45), but east of the Flint hill location now the similarity values decrease to a minimum. Good similarity is also shown in the southwestern quadrant, extending about as far south as to the southern drainage step in the topography map. The low is congruent with the Precambrian decrease in elevation, while topography and total thickness increase here; therefore this similarity map does account for the relief anomalies that are expected. Consequently, as we improved the way of standardizing the input data, the results improved also, inasmuch as they can now be interpreted geologically. Strong features, such as the

Nemaha Anticline, remain independent of the standardization, because the comparison algorithm is defined locally, thereby retaining the regional structure.

CONCLUSIONS

The two examples show that spatial data analysis is affected by pre-analysis standardization, and that results do depend on the standardization method chosen. A histogram does not generally provide a valid description of the properties of the spatial data, as the regional information is lost. A data set that yields a unimodally shaped histogram can actually belong to a uniform process. In both examples, the Gaussian hypothesis leads to less reasonable conclusions.

From the diversity of the two presented data sets it becomes apparent that the problem of standardization is not specific to a certain field but relevant to most different disciplines of geosciences. However, the degree of sensitivity to standardization depends on the data analysis method used.

Algebraic map comparison is more robust with respect to standardization than Q-mode factor analysis. The comparison method is itself defined locally, while in factor analysis the locations are determined by the eigenspace analysis. In both case studies, the proportion of range standardization with positive and inverse scaling, corresponding to the spatial distribution of the studied objects, proves to be superior to Gaussian standardization, in view of an oceanographical or geological analysis.

When investigating a new data set, several standardization methods should be considered. The two examples explored here demonstrate that one should carefully scrutinize the problem at hand and make sure that the chosen analysis method does not adulterate the raw data by means of inherent transformation and normalization constraints, but instead expresses the salient information in the least biased form. An analysis of the geological and oceanographic relationships and properties of the variables helps to find a suitable standardization that permits a meaningful interpretation of the data analysis results.

REFERENCES CITED

COLE, V. B., 1962, Configuration of top of Precambrian basement rocks in Kansas: Kansas Geol. Survey, Oil and Gas Invest. 26 (map).

HAMANN, I. M., and SWIFT, J. H., 1991, A consistent inventory of water mass factors in the intermediate and deep Pacific Ocean derived from conservative tracers: Deep Sea Research, in press.

HARMANN, H. H. 1976, Modern factor analysis: Chicago, The University of Chicago Press, 487 p.

HERZFELD, U. C., and MERRIAM, D. F., 1990, A map comparison technique utilizing weighted input parameters, *in* GAAL, G., and MERRIAM, D. F., eds., Computer Applications in Resource Estimation: Oxford, Pergamon Press, p. 43–52.

———, and SONDERGARD, M. A., 1988, MAPCOMP—A FORTRAN 77 program for weighted thematic map comparison: Computers Geosci., v. 14, p. 699–713.

IMBRIE, J., and VAN ANDEL, T. H., 1964, Vector analysis of heavy-mineral data: Geol. Soc. America Bull., v. 75, p. 1131–1156.

JORESKOG, K. G.; KLOVAN, J. E.; and REYMENT, R. A., 1976, Geological Factor Analysis: Amsterdam, Elsevier, 178 p.

KLOVAN, J. E., and IMBRIE, J., 1971, An algorithm and FORTRAN-IV program for large-scale Q-mode factor analysis and calculation of factor scores: Jour. Math. Geol., v. 3, p. 61–77.

KREYSZIG, E., 1970, Introductory mathematical statistics: New York, Wiley, 470 p.

MAMAYEV, O. I., 1975, Temperature-salinity analysis of World Ocean Waters: Amsterdam, Elsevier, 374 p.

MERRIAM, D. F., and JEWETT, D. G., 1988, Methods of thematic map comparison *in* MERRIAM, D. F., ed., Curent Trends in Geomathematics: New York, Plenum, p. 9–18.

———, and SONDERGARD, M. A. 1988, A reliability index for the pairwise comparison of thematic maps: Geol. Jahrbuch (Hanover), A 104, p. 433–446.

MIESCH, A. T., 1976, Q-mode factor analysis of geochemical and petrological data matrices with constant row-sums: U.S. Geol. Survey Prof. Paper 574-G, 47 p.

THURSTONE, L. L., 1947, Multiple Factor Analysis: Chicago, The University of Chicago Press, 535 p.

U.S. GEOLOGICAL SURVEY, 1963, State of Kansas: U.S. Geol. Surv. Map, 1:500000.