COMPUTATION OF THE CORRESPONDENCE OF GEOGRAPHICAL PATTERNS

WALDO R. TOBLER, University of Michigan

Geographical research often requires estimates of the amount of agreement between patterns shown on geographical maps. These patterns may be considered to consist of points, lines, areas, intensities, or flows. The correspondence to be estimated may comprise various combinations of such elements.[1] For example, what is the percentage agreement between the observed distribution of cities within Europe and Christaller's theoretical arrangement of central places. A somewhat more complicated situation, for which conventional statistical methods seem inadequate, might require computation of the spatial correspondence between the pattern of railroads and the pattern of roads in the United States. The objective of a number of recent studies has been the elucidation of methods of computing some such spatial correlations.[2]

A procedure is presented here for the estimation of the degree of correspondence between two patterns of point locations. In more geographical terms, the technique allows computation of the amount of agreement between two dot distributions. The method is comparable to the ordinary product moment (Pearsonian) correlation and least squares regression procedures, extended for comparisons of two-dimensional distributions. The difference is that instead of paired one-dimensional observations (of the form x_j; y_j) one has paired locations (of the form x_j, y_j; u_j, v_j). From these paired couples one can compute a spatial correlation. Note that in the ordinary correlation one associates phenomena at the same location, for example, fertilizer applied and corn yields obtained at the same spot. In the spatial correlation, one associates locations, for example, place of current residence and place of birth for one individual. As formulated, the method requires an a priori pairing of locations.[3] The method

The author is indebted to Professors W. Bunge, Jr., A. Court, and J. Nystuen for discussions of this topic and for pointing out errors in an earlier version of this paper.

[1] W. W. Bunge, Jr., *Theoretical Geography* (Lund: Gleerup, 1963), p. 151. Emphasis in Bunge's work is on comparisons of observed patterns with theoretically derived patterns.

[2] R. Bachi, "Standard Distance Measures and Related Methods for Spatial Analysis," *Papers of the Regional Science Association*, X (1962), pp. 83-132.

D. Neft, "Statistical Analysis for Areal Distributions," Ph. D. thesis, Columbia University, 1962, 286 pp.

H. H. McCarty, *et al*, "The Measurement of Association in Industrial Geography," Department of Geography, State University of Iowa, 1956, mimeographed.

R. F. Minnick, "Measurement of Areal Correspondence," *Papers*, Michigan Academy of Science, Arts, and Letters (1964).

A. H. Robinson and R. A. Bryson, "A Method for Describing Quantitatively the Correspondence of Geographical Distributions," *Annals*, Association of American Geographers, XXXXVII (1957), pp. 379-91.

[3] Ordinary correlation, of course, also requires such a priori pairings. Bachi, *op. cit.*, p. 122, comments on a possible strategy in cases for which a natural pairing does not exist.

seems particularly useful to geographers but has applications in other fields.[4] The spatial regression is not new[5] but does not appear to have been applied in the social sciences. The development given here extends the geographical work of the past decade or so on centrographical methods, which have dealt with concepts of mean location and standard distance.[6] Bachi has come the closest to making the transition from these descriptive notions to the spatial regression model and gives an outline for weighting the observations, along with a number of examples including normative models comparing even and random distributions with observed patterns.[7]

For the mathematical development, let the symbols W_j and Z_j represent the j^{th} observation pair, where $j = 1, 2, \ldots, N$. These symbols can be interpreted as complex numbers of the form

$$W_j = (u_j + iv_j) \text{ and } Z_j = (x_j + iy_j) ,$$

or as vectors of the form

$$W_j = (u_j , v_j)' \text{ and } Z_j = (x_j , y_j)'$$

where the primes denote the transpose. These two interpretations lead to slightly different results. In both cases, however, the elements (u, v, x, y) can be considered rectangular coordinates on a geographical map.[8] The objective

---

[4] Aside from the obvious ecological situations, see the biological mappings given in D'Arcy W. Thompson, *On Growth and Form* (Bonner Abridgement), (Cambridge: University Press), 1961, pp. 268–325.

[5] The formulation in terms of complex variables can be traced to C. F. Gauss. Also see:

M. Masuamya, "Correlation between Tensor Quantities," *Proceedings*, Physico-Mathematical Society of Japan, 3rd Series, XXI (1939), pp. 638–47.

T. H. Ellison, "On the Correlation of Vectors," *Quarterly Journal of the Royal Meteorological Society*, LXXX, 343 (1954), pp. 93–6.

A. Court, "Wind Correlation and Regression," *Scientific Report #3*, United States Air Force Contract 19 (604)-2060, AFCRC TN-58-230 (1958), 16 pp.

R. W. Lenhard, Jr., A. Court, and H. Salmela, "Reply," *Journal of Applied Meteorology*, II, 6 (1963), pp. 812–15.

H. U. Sverdrup, "Uber die korrelation zwischen Vektoren mit Anwendung auf meteorologische Aufgaben," *Metereologische Zeitschrift*, XXXIV (1917), pp. 285–91.

R. Detzius, "Ausdehnung der Korrelations methode und der Methode der kleinsten Quadrate auf Vektoren," *Sitzungsberichte der K. K. Akademie der Wissenchaften*, Wien, 125, IIa (1916), pp. 3–20.

[6] J. F. Hart, "Central Tendency in Areal Distribution," *Economic Geography*, XXX (1954), pp. 48–59.

B. G. Jones, "The Theory of the Urban Economy," Ph. D. thesis, University of North Carolina, 1960, 683 pp.

E. E. Sviatlovsky, and W. C. Eells, "The Centrographical Method and Regional Analysis," *The Geographical Review*, XXVII, 2 (1937), pp. 240–54.

[7] Nearest neighbor methods have also been employed for such comparisons; see M. F. Dacey, "Order Neighbor Statistics for a Class of Random Patterns in Multidimensional Space," *Annals*, Association of American Geographers, LIII, 4 (1963), pp. 505–15.

[8] The restriction to two components is suggested by the geographical subject matter. The development given here is for a plane.

in each instance is to use least squares methods to estimate the coefficients, A and B, in the transformation

$$\hat{W}_j = A + BZ_j$$

such that the residual is a minimum. The mathematical details are easily effected. The regression can be considered a mathematical mapping from one plane, the xy-plane, to another plane, the uv-plane, just as the ordinary regression line

$$\hat{Y} = a + bX$$

can be considered a mapping from one line, the X-axis, to another line, the Y-axis. The regression coefficients are given by the elements of B. Confidence limits may be established in a manner analogous to ordinary regression, if the necessary conditions are established. The coefficient of determination, R^2, can be defined as the ratio of the regression variance over the total variance, and the spatial correlation, R, is then given by the square root of this value. The sign of the correlation coefficient can be taken to be the sign of the determinant of the transformation.

When the observations are treated as complex numbers, the constants to be determined are

$$A = (a_1 + ia_2) \text{ and } B = (b_1 + ib_2)$$

and the equation to be minimized is

$$\sum_{j=1}^{n} | \hat{W}_j - W_j |^2 .$$

The complete transformation equations are then, separating the real and imaginary parts,

$$\hat{u}_j = a_1 + b_1 x_j - b_2 y_j$$
$$\hat{v}_j = a_2 + b_2 x_j + b_1 y_j .$$

This transformation consists of a rigid rotation, translation, and change of scale. It is proposed that this be referred to as the complex or Euclidean regression and correlation.

When the observations are treated as components of a vector the constants become

$$A = (a_1, a_2)' \text{ and } B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

and the equation to be minimized is

$$\sum_{j=1}^{n} (\hat{W}_j - W_j) \cdot (\hat{W}_j - W_j)' .$$

The complete transformation is therefore of the form

$$\begin{pmatrix} \hat{u}_j \\ \hat{v}_j \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} x_j \\ y_j \end{pmatrix} .$$

These are the equations of an affine transformation, and it is proposed that the regression, and correlation, be referred to by this name. The term vector correlation, however, has been applied to this transformation.[9] In terms of the simple correlations and variances, this affine correlation is given by Court[10] as

$$R_{WZ} = \frac{\sigma_u^2(r_{xu}^2 + r_{yu}^2 - 2r_{xu}r_{yu}r_{xy}) + \sigma_v^2(r_{xv}^2 + r_{yv}^2 - 2r_{xv}r_{yv}r_{xy})}{(\sigma_u^2 + \sigma_v^2)(1 - r_{xy}^2)} .$$

This is not symmetrical, and R_{WZ} is not equal to R_{ZW}.

Hotelling also gives a different definition of vector correlation in conjunction with his canonical correlation.[11] In this case, the problem is to find α_k such that the (ordinary) correlation between the canonical variates

$$w_j = \alpha_1 u_j + \alpha_2 v_j \text{ and } z_j = \alpha_3 x_j + \alpha_4 y_j$$

is as large as possible. From our point of view, this approach has the disadvantage that, given x_j, y_j, and α_k, it is not possible to obtain a unique pair

$$\hat{u}_j, \hat{v}_j$$

as is the case with the spatial regression models given above.

Curvilinear spatial regression can also be established[12] but is more complicated. One reason the ordinary linear regression is so simple is that there is only one straight line which can be fitted to the data, given the least squares criterion.[13] In fitting a transformation from one plane to another plane, the general linear mappings defined here yield similarly unambiguous results. Choice of the model in ordinary curvilinear regression is considerably more difficult, since there are many curves from which to choose. The same is true of mappings from one plane to another. Knowing only that the mapping is to be curvilinear does not specify the transformation equations sufficiently. Polynomials seem advantageous since solution of the normal equations is relatively simple. On the other hand, the form of the curvilinear model should be chosen on the basis of theoretical expectations. This is perhaps more difficult for two-dimensional transformations than in the case of ordinary regression, but geographers have an advantage in that they are acquainted with the closely related

---

[9] See Court, *op. cit.*, and Lenhard, *et al*, *op. cit.*

[10] Court, *op. cit.*, equation 6.7, p. 12.

[11] H. Hotelling, "Relations between Two Sets of Variates," *Biometrika*, XXVIII (1936), pp. 321–77.

[12] Ellison, *op. cit.*, p. 95, gives the treatment for vectors. The conformal transformation defined by a complex polynomial is employed in photogrammetry; see G. H. Schut, "Development of Programs for Strip and Block Adjustment at the National Research Council of Canada," *Photogrammetric Engineering*, XXX, 2 (1964), pp. 283–91. Bunge has suggested as geographically interesting a least squares approach to the problem of forcing distributions into Christaller's hexagonal central place pattern. A differential equation defining the restrictions on the transformation for this problem is given in W. R. Tobler, "Geographical Area and Map Projections," *The Geographical Review*, LIII, 1 (1963), pp. 59–78.

[13] This is not strictly true. See R. L. Miller and J. S. Kahn, *Statistical Analysis in the Geological Sciences* (New York: John Wiley & Sons, 1962), p. 204.

subject of map projections. As an example, if one wishes to estimate the correspondence between current residence and immediately previous residence within an urban area in order to investigate whether or not migration to the suburbs is within the same sector, then it might be appropriate to employ a curvilinear mapping, since one suspects that people in the suburbs are more mobile than people living near the center of town; this is a fairly common feature of geographical movement.

For an equivalent to the ordinary multiple regression, one considers locations in the uv-plane as being dependent on locations in an xy-plane *and* on locations in an mn-plane, and so on. That is, one wishes to explain, in the least squares sense, one geographical pattern on the basis of several, k, other geographical patterns. The observed information then consists of $k + 1$ pairs of coordinates possibly for several time periods. The mathematical details are fairly simple; Ellison gives the treatment for vectors.[14]

Information made available through the courtesy of Pitts, of the University of Pittsburgh, has been employed for a numerical example. This information consists of 767 pairs of latitudes and longitudes giving the premarital residential locations of brides and grooms in the rural area south of Takamatsu City, Japan, for the year 1951. The material was collected by Professor Pitts in 1962. Discussion of the appropriateness of the model for these data is deferred until presentation of the numerical results. The form of the data, after conversion to cartesian coordinates in miles from an arbitrary origin at 34°N latitude and

TABLE 1

OBSERVATIONS

Observation Number	Groom		Bride	
	u	v	x	y
1	56.23	13.86	55.99	13.23
2	62.42	20.74	59.99	22.15
3	57.23	15.58	59.45	21.36
4	58.70	17.16	59.23	21.11
5	57.11	23.76	59.68	21.67
6	75.34	17.49	62.99	23.77
7	66.30	19.30	61.83	21.42
8	57.08	18.59	61.32	21.08
9	62.39	20.50	61.26	21.32
10	61.18	20.74	60.10	21.10

First 10 of 767 observations.

x bride easting y bride northing

u groom easting v groom northing

Values are in miles east and north of 34°N, 133°E.

[14] Ellison, *op. cit.*, p. 95.

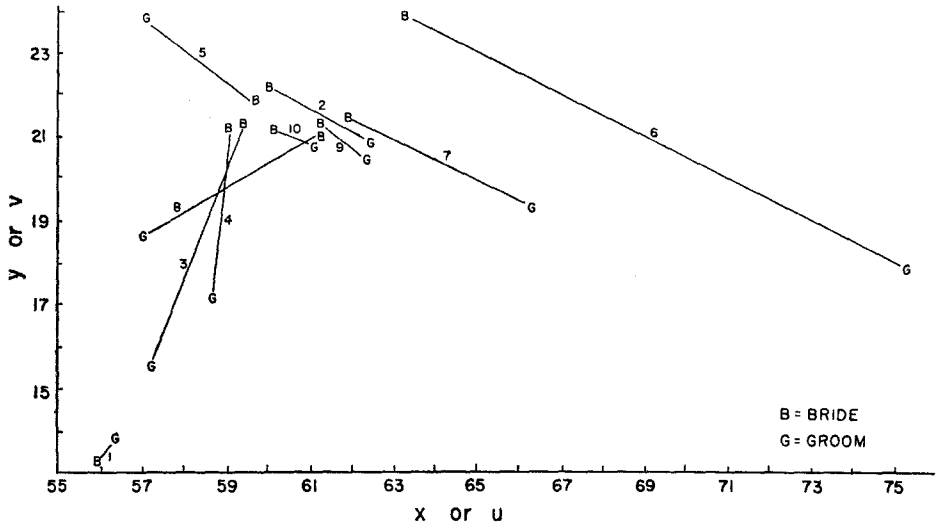

FIGURE 1. OBSERVED LOCATION PAIRS

133°E longitude, is given in Table 1 and Figure 1. The summary measures are given in Table 2. The affine regression equation is

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} 30.11 \\ 22.84 \end{bmatrix} + \begin{bmatrix} 0.5047 & -0.0105 \\ -0.1372 & 0.2569 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

and has a standard error of (5.84, 3.56) miles.[15] This equation can be interpreted as transforming every grid square in the plane of the independent ob-

TABLE 2

CORRESPONDENCE OF LOCATIONS OF BRIDES AND GROOMS
(RESULTS OF COMPUTATION)

$\bar{x} = 59.74$	$\bar{y} = 21.16$	$\bar{u} = 60.04$	$\bar{v} = 20.07$
$\sigma_x = 3.38$	$\sigma_y = 2.90$	$\sigma_u = 6.08$	$\sigma_v = 3.65$
cov $(xy) = 1.33$	$r(xy) = .14$		$\sigma_{xy} = 4.45$
cov $(xu) = 5.74$	$r(xu) = .28$		$\sigma_{xu} = 6.96$
cov $(xv) = -1.22$	$r(xv) = -.10$		$\sigma_{xv} = 4.97$
cov $(yu) = 0.58$	$r(yu) = .03$		$\sigma_{yu} = 6.74$
cov $(yv) = 1.98$	$r(yv) = .19$		$\sigma_{yv} = 4.67$
cov $(uv) = 4.98$	$r(uv) = .22$		$\sigma_{uv} = 7.09$

Complex correlation: $R(xy, uv) = 0.007$.
Affine correlation: $R(xy, uv) = 0.266$, $R(uv, xy) = 0.263$.
Sverdrup's vector correlation = .003.
Hotelling's vector correlation = .003.
Bachi's index of association = .22.
Bachi's index of nonrandomness = .25.

[15] The computations were performed with the assistance of the University of Michigan Computing Center. Copies of the computer program can be made available to interested parties for a limited time.

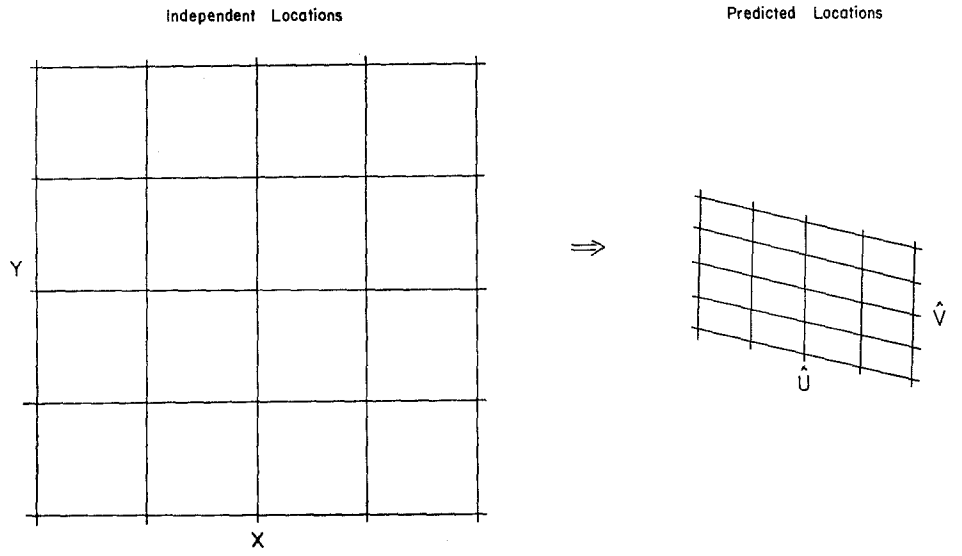

FIGURE 2. GRAPHIC ILLUSTRATION OF THE REGRESSION
(Mapping of the xy-plane onto the uv-plane)

servations (brides) into a parallelogram to obtain an estimate of locations of the dependent observations (grooms), as illustrated in Figure 2. The predicted locations for a number of observations are given in Table 3 and Figure 3.

The results of this numerical example are such that one can predict the location of a groom, with an average error of about seven miles, when given the location of a bride in a certain region in Japan. This is presented solely as a numerical example and is of little serious theoretical interest, except per-

TABLE 3

PREDICTED GROOM LOCATIONS

Observation Number	Predicted Groom		Deviation		Miles
	$\hat{u}$	$\hat{v}$	$\hat{u}-u$	$\hat{v}-v$	
1	58.23	18.55	2.00	4.69	5.10
2	60.16	20.29	−2.27	−0.45	2.31
3	59.89	20.16	2.66	4.58	5.30
4	59.78	20.13	1.08	2.97	3.16
5	60.00	20.21	2.89	−3.55	4.58
6	61.65	20.30	−13.69	2.81	13.97
7	61.09	19.85	−5.07	0.55	5.10
8	60.84	19.84	3.76	1.24	3.96
9	60.81	19.91	−1.59	−0.59	1.69
10	60.22	20.01	−0.96	−0.73	1.21

First 10 of 767 observations.

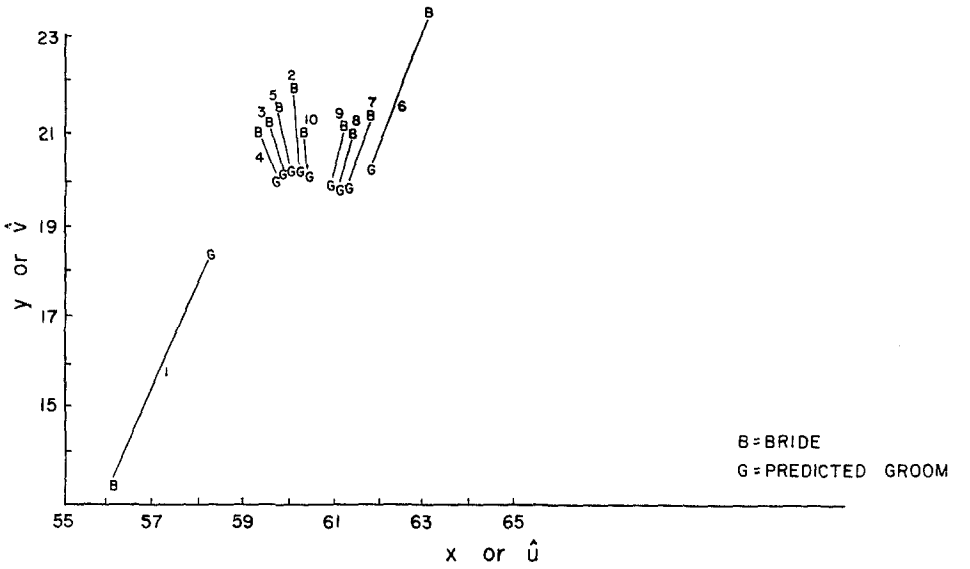

FIGURE 3. PREDICTED LOCATION PAIRS

haps for matrimonially inclined females. The affine correlation is low, 0.266, and the theoretical justification meager. From a mathematical point of view, a least squares fit of this type can always be obtained and can be employed for empirical estimates of correspondence and prediction. This is comparable to the fitting of trend lines in order to extrapolate population growth in some region. The method is employed in practice, but suffers from a paucity of theoretical insight.[16] This can be contrasted with the stimulus response interpretation of a regression equation, where, for example, it can be anticipated on theoretical grounds that an increased application of fertilizer will be followed by an increase in crop yields.

The two-dimensional interpretation might be that a change in one geographical pattern is followed by a change in a second geographical pattern. Unless the form of such a change can be anticipated on theoretical grounds, the analysis must remain statistical and devoid of fruitful substantive interpretation. The interesting situations are most generally nonlinear, and formulation in terms of curvilinear least squares mappings seems appropriate, but these also are more complicated mathematically.

The general method presented here has the advantages and disadvantages of any single equation regression model. One should not attempt to fit a straight line to data which, when plotted on a scatter diagram, appear to lie on a circle. If one does, the resulting low linear correlation cannot be interpreted as implying that no relation exists. Similarly, the intermarital distance frequencies in the foregoing numerical example display the decline with increasing distance typical of many geographical interactions. One must also

[16] W. Isard, *Methods of Regional Analysis* (New York: John Wiley & Sons, 1960).

avoid many-to-one situations, as for example, place of employment and place of residence when large employment centers are involved in the analysis. Substitution of the mean employee residence may be appropriate in such cases.