Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2009;38:337–341
doi:10.1093/ije/dyn357

REPRINTS AND REFLECTIONS

Ecological Correlations and the Behavior of Individuals*

WS Robinson[1]

INTRODUCTION

AN INDIVIDUAL CORRELATION is a correlation in which the statistical object or thing described is indivisible. The correlation between color and illiteracy for persons in the United States, shown later in Table I, is an individual correlation, because the kind of thing described is an indivisible unit, a person. In an individual correlation the variables are descriptive properties of individuals, such as height, income, eye color, or race, and not descriptive statistical constants such as rates or means.

In an *ecological correlation* the statistical object is a *group* of persons. The correlation between the percentage of the population which is Negro and the percentage of the population which is illiterate for the 48 states, shown later as Figure 2, is an ecological correlation. The thing described is the population of a state, and not a single individual. The variables are percentages, descriptive properties of groups, and not descriptive properties of individuals.

Ecological correlations are used in an impressive number of quantitative sociological studies, some of which by now have attained the status of classics: Cowles' ''Statistical Study of Climate in Relation to Pulmonary Tuberculosis'';[1] Gosnell's ''Analysis of the 1932 Presidential Vote in Chicago,''[2] Factorial and Correlational Analysis of the 1934 Vote in Chicago,''[3] and the more elaborate factor analysis in *Machine Politics*;[4] Ogburn's ''How women vote,''[5] ''Measurement of the Factors in the Presidential Election of 1928,''[6] ''Factors in the Variation of Crime Among Cities,''[7] and Groves and Ogburn's correlation analyses in *American Marriage and Family Relationships*;[8] Ross' study of school attendance in Texas;[9] Shaw's *Delinquency Areas* study of the correlates of delinquency,[10] as well as The more recent analyses in *Juvenile Delinquency in Urban Areas*;[11] Thompson's ''Some Factors Influencing the Ratios of Children to Women in American Cities, 1930'';[12] Whelpton's study of the correlates of birth rates, in ''Geographic and Economic Differentials in Fertility;''[13] and White's ''The Relation of Felonies to Environmental Factors in Indianapolis.''[14]

Although these studies and scores like them depend upon ecological correlations, it is not because their authors are interested in correlations between the properties of areas as such. Even out-and-out ecologists, in studying delinquency, for example, rely primarily upon data describing individuals, not areas.[15] In each study which uses ecological correlations, the obvious purpose is to discover something about the behavior of individuals. Ecological correlations are used simply because correlations between the properties of individuals are not available. In each instance, however, the substitution is made tacitly rather than explicitly.

The purpose of this paper is to clarify the ecological correlation problem by stating, mathematically, the exact relation between ecological and individual correlations, and by showing the bearing of that relation upon the practice of using ecological correlations as substitutes for individual correlations.

The Anatomy of an ecological correlation

Before discussing the mathematical relation between ecological and individual correlations, it will be useful to exhibit the structural connection between them in a specific situation. Figure 1 shows the scatter diagram for the ecological correlation between color and illiteracy for the Census Bureau's nine geographic divisions of the United States in 1930. The X-coordinate of each point is the percentage of the divisional population 10 years old and over which is Negro. The Y-coordinate is the percentage of the same population which is illiterate.[16] The Pearsonian correlation for Figure 1, i.e. the ecological correlation, is .946.

Table 1 is a fourfold table showing for the same population the correlation between color and illiteracy considered as properties of individuals rather than geographic areas. The Pearsonian (fourfold-point) correlation for Table I, i.e., the individual correlation, is .203, slightly more than one-fifth of the corresponding ecological correlation.

* American Sociological Review, Vol 15. No 3 (Jun., 1950), 351–357. Reprinted with permission.

[1]University of California at Los Angeles.

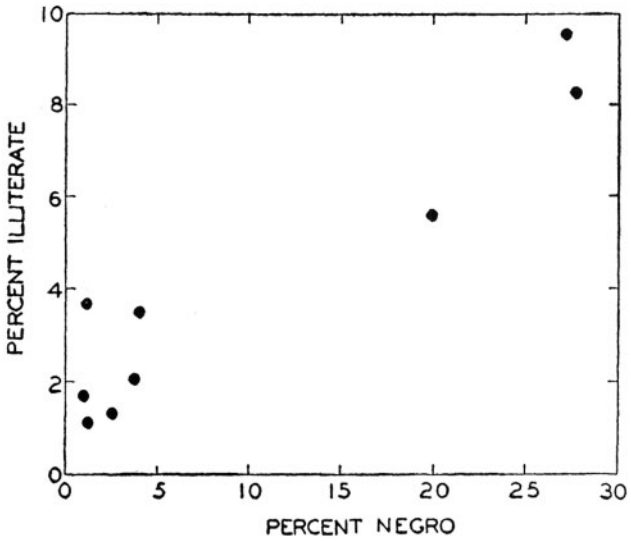

Figure 1.

Ordinarily, such an ecological correlation would be computed on a county or state basis, instead of the divisional basis used here to simplify numerical presentation. Whether the ecological areas are counties, states, or divisions, however, the results are similar. Figure 2, for example, shows the ecological correlation on a state rather than a divisional basis. When the ecological areas are states, as in figure 2, the ecological correlation is .773, to be compared with .946 when the ecological areas are divisions.

The connecting link between the individual correlation of Table 1 and the ecological correlation of Figure 1 is the individual correlations between color and illiteracy *within* the nine geographic divisions which furnish the nine observations for the ecological correlation. These are the *within-areas individual correlations*, a selection from which is given in Table 2.

Both the individual correlation and the ecological correlation depend upon the within-areas individual correlations, but in different ways. The individual correlation (Table 1) depends upon the internal or cell frequencies of the nine within-areas individual correlations. Its cell frequencies are sums of the nine corresponding divisional cell frequencies. For example, in the upper left cell of Table 1 the frequency is $1{,}512 = 4 + 32 + 36 + \ldots + 2$.

The ecological correlation (Figure 1) also depends upon the nine within-areas individual correlations, *but only upon their marginal totals*. For example, in Table 2 the marginal total for the first table shows 76,000 Negroes in the New England division. Since the total population for this division is 6,702,000, the percentage of Negroes is 100(76)/6,702 – 1.1. The percentage of illiterates in New England is computed from the other marginal total in the same way.

In brief, the individual correlation depends upon the *internal* frequencies of the within-areas individual correlations, while the ecological correlation depends upon the *marginal* frequencies of the within-areas individual correlations. Moreover, it is well known that the marginal frequencies of a fourfold table do not determine the internal frequencies. There is a large number of sets of internal frequencies which

Table 1 The individual correlation between color and illiteracy for the united states, 1930 (for the population 10 years old and over)[17]

	Negro	White	Total
Illiterate	1,512	2,406	3,918
Literate	7,780	85,574	93,354
Total	9,292	87,980	97,272

Table 2 The within-areas individual correlations between color and illiteracy for the united states, 1930[18]

		Negro	White	Total
New England	Illiterate	4	240	244
	Literate	72	6,386	6,458
	Total	76	6,626	6,702
Middle Atlantic	Illiterate	32	719	751
	Literate	836	19,958	20,794
	Total	868	20,677	21,545
East North Central	Illiterate	36	392	428
	Literate	735	19,443	20,178
	Total	771	19,835	20,606
Pacific	Illiterate	2	71	73
	Literate	75	6,332	6,407
	Total	77	6,403	6,480

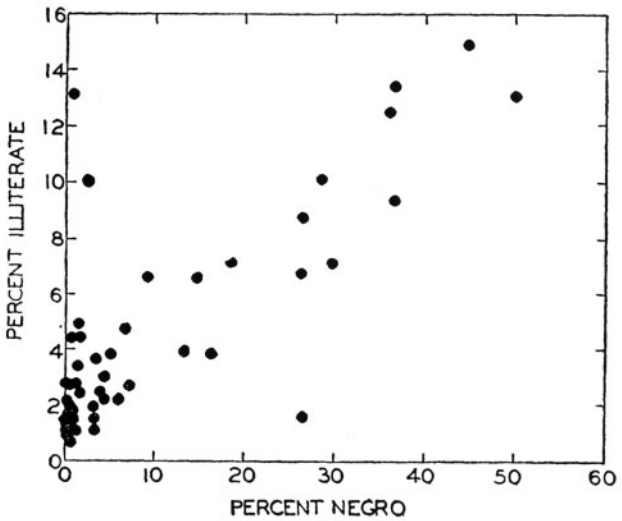

Figure 2.

will satisfy exactly the same marginal frequencies for any fourfold table. Therefore there are a large number of individual correlations which might correspond to any given ecological correlation, i.e. to any given set of marginal frequencies. In short, the within-areas marginal frequencies which determine the percentages from which the ecological correlation is computed do not fix the internal frequencies which determine the individual correlation. Thus there need be no correspondence between the individual correlation and the ecological correlation.

An instance will document this conclusion. The data of this section show that the individual correlation between color and illiteracy is .203, while the ecological correlation is .946. In this instance, the two correlations at least have the same sign, and that sign is consistent with our knowledge that educational standards in the United States are lower for Negroes than for whites.

However, consider another correlation where we also know what the sign ought to be, viz, that between nativity and illiteracy. We know that educational standards are lower for the foreign born than for the native born, and therefore that there ought to be a positive correlation between foreign birth and illiteracy. This surmise is corroborated by the individual correlation between foreign birth and illiteracy, shown in Table 3. The individual correlation for Table 3 is .118. However, the ecological correlation between foreign birth and illiteracy, shown in Figure 3, is .−619! When the ecological correlation is computed on a state rather than a divisional basis, its value is −.526.

Table 3 The inidivdiual correlation between nativity and illiteracy for the united states, 1930 (for the population 10 years old and over)

	Foreign born	Native born	Total
Illiterate	1,304	2,614	3,918
Literate	11,913	81,441	93,354
Total	13,217	84,055	97,272

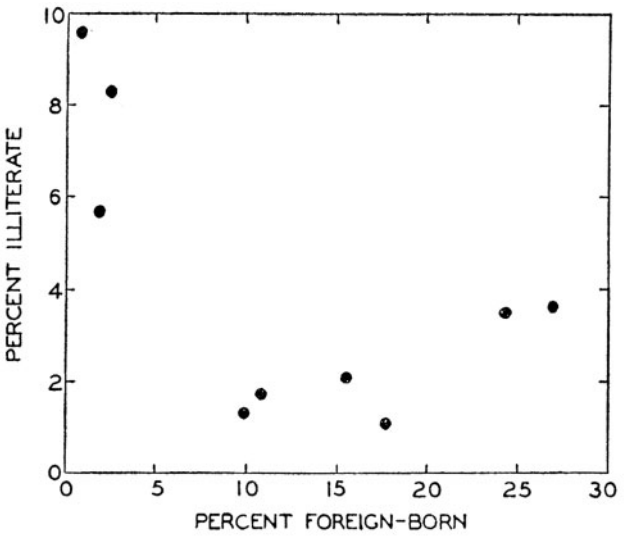

Figure 3.

(1) There is a total group of N persons, who are characterized by two variable properties X and Y. These properties may be genuine variables such as age or income, or they may be dichotomous attributes such as sex or literacy.
(2) The N members of the total group can be put into *m* distinct sub-groups according to their geographic position, whether by census tracts, townships, counties, states, or divisions. It is convenient to think of these *m* sub-groups as defined by *m* values of a third variable A (=Area) which is really an attribute, viz, geographic region.

The numerical values from which the ecological correlation is computed describe these *m* sub-groups. They may be means, medians, or percentages, and in fact all three are sometimes involved in a single ecological correlation analysis. Usually, however, they are percentages. While the mathematics applies to means as well, and approximately to medians also, it will simplify the present discussion to assume that X and Y are dichotomous properties, and therefore that the ecological correlation is a correlation between *m* pairs of percentages.

In the preceding section, three distinct correlations were shown to be involved in the ecological correlation situation. In mathematical terms, these correlations are described as follows:

The *total individual correlation*® is the simple Pearsonian correlation between X and Y for all N members of the total group, computed without reference to geographic position at all. If X and Y are dichotomous properties, the total individual correlation will be a fourfold-point correlation based on a fourfold table (Table 1).

The *ecological correlation* (r_e) is the weighted correlation between the *m* pairs of X- and Y-percentages which describe the sub-groups. In the example of Section 2, r_e is the correlation between the nine percentages of Negroes and the nine corresponding percentages of illiterates. However, each cross-product of an X- and Y-percentage is weighted by the number of persons in the group which the percentage describes, to give it an importance corresponding to the number of observations involved.

Ordinarily, ecological correlations are computed without the refinement of weighting. While the weighted form is theoretically more adequate, and is required by the mathematics of this section, the numerical difference between the two is negligible. The weighted ecological correlation for Figure 1, which involves few observations and should therefore

be very sensitive to weighting, is .946, while the corresponding unweighted value is .944.

The *within-areas individual correlation* (r_w) is a weighted average of the *m* within-areas individual correlations between X and Y, each within-area correlation being weighted by the size of the group which it describes.

Two correlation ratios, η_{XA} and η_{YA}, are also involved in the relation. Their purpose is to measure the degree to which the values of X and Y show clustering by area. If X is a dichotomous property, say illiteracy, then a large value of η_{XA} indicates wide variation in the percentage of illiterates from one area to another.

With these definitions, the relation between individual and ecological correlations may be written as

$$r_e = k_1 r - k_2 r_w, \tag{1}$$

where

$$k1 = 1\eta_{XA}\eta_{YA} \tag{1a}$$

and

$$k_2 = \sqrt{1 - \eta_{XA}^2}\sqrt{1 - \eta_{YA}^2}\eta_{XA}\eta_{YA}. \tag{1b}$$

That is, the ecological correlation is the weighted difference between the total individual correlation and the average of the *m* within-areas individual correlations. In this weighted difference, the weights of the total individual correlation and the within-areas individual correlation depend upon the degree to which the values of X and Y show clustering by area.

Investigation of the relation given in (1) shows that an individual and ecological correlation will be equal, and the equivalency assumption will therefore be valid, when

$$r_w = k_3 r, \tag{2}$$

where

$$k_3 = \frac{1 - \eta_{XA}\eta_{YA}}{\sqrt{1 - \eta_{XA}^2}\sqrt{1 - \eta_{YA}^2}} \tag{2a}$$

However, the minimum value of k_3 in (2) is unity. Therefore (2) will hold, and the individual and ecological correlations will be equal, only if the average within-areas individual correlation is not less than the total individual correlation. But all available evidence is that (whatever properties X and Y may denote) the correlation between X and Y is certainly not larger for relatively homogenous sub-groups of persons than it is for the population at large. In short, the equivalency assumption has no basis in fact.

The consistently high numerical values of published ecological correlations in comparison with the smaller values ordinarily got in correlating the properties of individuals suggest that ecological correlations have some reason for being larger than corresponding individual correlations. The relation given in (1) shows what this reason is, for it gives as the condition for the numerically larger value of the ecological correlation

$$r_w < k_3 r, \tag{3}$$

where k_3 is given by (2a). Since the minimum value of k_3 is unity, equation (3) implies that the ecological will be numerically greater than the individual correlation whenever the within-areas individual correlation is not greater than the total individual correlation, and this is the usual circumstance.

Habitual users of ecological correlations know that the size of the coefficient depends to a marked degree upon the number of sub-areas. Gehlke and Biehl, for example, commented in 1934[20] upon the positive relation between the size of the coefficient and the average size of the areas from which it was determined. This tendency is illustrated in Section 2, where the ecological correlation between color and illiteracy is .773 when the sub-areas are states and .946 when the sub-areas are the Census Bureau's nine geographic divisions. The same tendency is shown by the correlations between nativity and illiteracy, the value being −.526 on a state basis and −.619 on a divisional basis.

Equation (1) shows why the size of the ecological correlation depends upon the number of sub-areas, for the behavior of the ecological correlation as small sub-areas are grouped into larger ones can be predicted from the behaviour of the variables on the right side of (1) as consolidation takes place. As smaller areas are consolidated, two things happen:

(1) The average within-areas individual correlation increases in size because of the increasing heterogeneity of the sub-areas. The effect of this is to *decrease* the value of the ecological correlation.
(2) The values of $\eta_{XA}\eta_{YA}$ decrease because of the decrease in the homogeneity of values of X and Y within sub-areas. The effect of this is to *increase* the value of the ecological correlation.

However, these two tendencies are of unequal importance. Investigation of (1) with respect to the effect of changes in the values of $\eta_{XA,}\eta_{YA}$, and r_w indicates that the influence of changes in the η's is considerably more important than the influence of changes in the value of r_w, The net effect of changes in the values of the η's and of r_w taken together, therefore, is to increase the numerical value of the ecological correlation as consolidation takes place.

Conclusion

The relation between ecological and individual correlations which is discussed in this paper provides a

definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot. While it is theoretically possible for the two to be equal, the conditions under which this can happen are far removed from those ordinarily encountered in data. From a practical standpoint, therefore, the only reasonable assumption is that an ecological correlation is almost certainly not equal to its corresponding individual correlation.

I am aware that this conclusion has serious consequences, and that its effect appears wholly negative because it throws serious doubt upon the validity of a number of important studies made in recent years. The purpose of this paper will have been accomplished, however, if it prevents the future computation of meaningless correlations and stimulates the study of similar problems with the use of meaningful correlations between the properties of individuals.

[1] Journal of the American Statistical Association, 30 (Sept., 1935), 517–536.
[2] American Political Science Review, 24 (Dec., 1935), 967–984.
[3] Journal of the American Statistical Association, 31 (Sept.,1936), 507–518.
[4] Chicago, 1938.
[5] Political Science Quarterly, 34 (Sept., 1919), 413–433.
[6] Social Forces, 8 (Dec., 1929), 175–183.
[7] Journal of the American Statistical Association, 30 (Mar., 1935), 12–34.
[8] New York, 1928.
[9] School Attendance in the United States: 1920 a supplementary report to the 1920 U.S Census, Washington, 1924.
[10] Chicago, 1929.
[11] Chicago, 1942.
[12] American Journal of Sociology, 45 (Sept., 1939), 183–199.
[13] Annals of the American Academy of Political and Social Science, 188 (Nov., 1936), 37–55.
[14] Social Forces, 11 (May, 1932), 498–513.
[15] In Shaw's Delinquency Areas, for example.
[16] These percentages were computed from the marginal totals of the fourfold tables given in Table 2.
[17] The source for this and all following tables is the 1930 U.S. Census. All figures are in thousands.
[18] The tables for the West North Central, South Atlantic, East South Central, West South Central, and Mountain divisions are omitted to save space.
[19] The derivation of this equation is not given here because of space limitations. Readers wishing a copy may secure one by sending a stamped, self-addressed envelope to WS Robinson, Department of Anthropology and Sociology, University of California, Los Angeles 24, California.
[20] Certain effects of grouping upon the size of the correlation coefficient in census tract material,'' Journal of the American Statistical Association, 24 (Mar., 1934, supplement), 169–170.