The Measurement Of Spatial Association: A Comment

G. B. Norcliffe*

This comment addresses Latham's [6] review of techniques for measuring spatial association in which he identifies three approaches, namely: distance based measures; dichotomous measures of relative location; and continuous measures of relative location. Two purposes motivate what follows. First, to elaborate on some of the difficulties encountered in using measures based on distances. And second, to argue that a correlation coefficient based on proportional data, of which Latham is rather critical, may in many circumstances be a better indicator of spatial association than is a correlation coefficient based on absolute values, this being the measure preferred by Latham.

MEASURES BASED ON DISTANCES

Latham's criticisms of distance based measures relate mainly to a study by Hale [4] in which mean shipment distances to suppliers and markets are calculated for industrial sectors as a whole. He makes three cogent criticisms of Hale's study. First, the mean shipment distances in the case study are so long that they give little indication of geographical association. Second, they are computed for industrial sectors as a whole, and not for individual establishments: since industries as a whole typically have interindustry relations with many other sectors, average shipment distances become inflated. And third, use of the arithmetic mean of shipment distances may conceal close association between a particular pair of industries. Latham suggests that the median distance is a better measure.

These criticisms are not pervasive. They relate to a particular use of distance measurements. If distance measures were calculated not for industrial sectors, but for individual establishments, and if modal distances or indeed the entire frequency distribution of shipment distances were used, then these criticisms would not carry much force. Latham correctly points out that shipment distance data at the establishment level are not readily available, but this is a problem of data availability, not of the method per se.

If data were available at the establishment level, then shipment distances would be a more useful indicator of geographical association than are the same measures based on industry aggregates. Nevertheless they would still be descriptive statistics with no inferential basis, and therein lies the most

*Department of Geography, York University, Canada.
telling criticism of this measure. Clearly, the shipment distances for certain pairs of industries will be less than for others, but designating which are geographically associated and which are not is purely arbitrary. It may be possible to deduce some sort of expected distribution of shipping distances using randomization procedures, but the inherent spatial autocorrelation present in all human distribution patterns [2], the influence of the shape of territories [4], and the different intensities of economic activity from one region to another suggest that this would be a formidable task.

CONTINUOUS MEASURES OF RELATIVE LOCATION

Latham's discussion of continuous measures of geographical association deals largely with the relative merits of absolute versus proportional data in computing correlation coefficients. Latham agrees with Richter [12] that correlations will be high if regions are large and if the industries are aggregated into heterogeneous groups. In particular, if there is a high correlation between each industry's regional employment distribution and the regional distribution of population, then there will be high correlations amongst the industries concerned.

This argument can be re-phrased to put the emphasis on magnitudinal differences. It is well known that the number of employees in many industries shows a fair degree of correspondence with the population of a region, that correspondence being quite close amongst market-oriented industries. It follows that in any situation where the population magnitudes vary greatly from one region to another (such as was the case in the study by McCarty, Hook and Knos [8]) pairwise correlations amongst industries will tend to be high.

Latham argues that the problem of inflated correlations is reduced when data are disaggregated more finely by industry and by region, demonstrating this with an example drawing from U.S. data. Using data for 199 industries at the 4 digit level of the S.I.C. measured over a total of 377 regions including both S.M.S.A's and non-S.M.S.A's, he finds that, in contrast to the McCarty study, correlations between population and regional employment are not uniformly high. Nevertheless, very few negative correlations are recorded, and fully sixty percent of his correlations are greater than +.3. By using a sample of 377 regions, magnitudinal effects appear to be reduced. However, that reduction is not an intrinsic property of the number of regions; it is attributable to the greater homogeneity of size of regions in his sample.

Richter [12] and Lever [7] have sought to eliminate the magnitudinal effect entirely by expressing employment in each industry as a ratio of the total employment in a region. Latham considers this possibility, but does not adopt it for three major reasons. First, the resulting correlations are affected by a redistribution of the population base, whereas correlations based on absolute data are invariant under such redistributions. Second, small outputs in small regions are weighted as heavily as large outputs in large regions when ratio data are used. And third, ratio correlation can lead to a reversal of signs, indeed this occurs in a simple example in which
he compares employment in manufacturing with employment in services for ten cities.

In a recent study of local industrial complexes [11], I was faced with the choice between using absolute and ratio data to measure geographical associations amongst industries [10]. Having opted for the latter, it may be useful to argue the case in favour of ratio correlation. This will be done by examining Latham’s three criticisms.

There is an element of arbitrariness in the choice of a base for ratio correlation. For instance, the denominator could be total population, total employment in all industries, or total employment in manufacturing (or in the sub-set of industries being considered). Typically, choice amongst these alternative bases will have only a small effect on the resulting correlations. The other problem relating to the choice of base is where a redistribution of the base occurs. Latham uses a hypothetical example with the closure of a military base in one town, and a redistribution of part of its population amongst nine other towns, while no change takes place in manufacturing and service employment. The argument is not convincing because such instantaneous adjustments are most exceptional. Commonly a more gradual redistribution occurs, and interindustry dependencies lead one to expect associated changes in employment structure. Moreover if the base is the sum of the industries being correlated, then a redistribution of the base must also affect some of the industries.

The arguments for and against weighting areal units according to size has been debated in the geographical literature for some time. The issue should be considered in the context of ecological correlations [3]. The question is: what size of areal unit best measures the relationship between two variables? Latham would apparently support Robinson’s [13] argument that greater weight should be given to large regions. However, Curry [2] has turned the argument round by suggesting that smaller regions will give a more precise measure of the relationship between two variables. His argument relates to the characteristic spectrum of the spatial series in question: small regions, rather like a finely meshed fishing net, trap more information about the relationship between two variables than do large regions. Accordingly, there is a case for putting a greater weight on small regions.

Latham’s third criticism of correlations based on proportional data is that a reversal of signs may occur. What is raised is the problem of closed number systems, percentage and proportional data being cases in point. The major contributions to this topic appear to have come from stratigraphers. Latham’s example concerns the correlation between two variables in a three variable system (the unspecified third variable is the remainder of the population not employed in manufacturing and services). At least two of the correlation coefficients in a three variable closed system will always be negative, and the correlation between pairs of variables can be predicted from the sample variances without any knowledge of the covariation present. Clearly, in a three variable closed number system, correlations need interpreting with great care. Krumbein and Graybill [5, p. 52] note that as the number of variables (sectors) increases,
so more degrees of freedom are entered into the closed number system and the effect on correlation coefficients diminishes. Usually, studies of geographical association amongst industries are based on data sets with a large number of industries, so that the effect upon correlations will be quite small.  

In measuring geographical association amongst industries using ratio correlation, it would seem that the problems referred to by Latham will not commonly be serious. On the other hand, the problem of inflated correlations due to magnitudinal differences amongst regions is a very common problem. In large measure this is attributable to city systems generally having skewed frequency distributions, often approximating the lognormal density. Clearly, in exceptional cases where all regions are of uniform size, absolute and ratio data will yield essentially identical correlation coefficients. In other cases, correlations based on absolute values will tend to be inflated: *ceteris paribus*, the more unequal is the size of regions, the greater is the inflationary effect on the correlation coefficients.

Bearing in mind the foregoing argument, should correlations amongst industries be based on proportional or absolute data? In essence, it depends upon the meaning one attaches to the term "geographical association." Latham hints that magnitude itself should be taken into account. Comparing employment in manufacturing and services using proportional data he obtains [6 p. 26] "a negative value which seems intuitively to be misleading because there is no reason to expect that services will be repelled from a region as large as an S.M.S.A. by manufacturing." In absolute terms, yes, but in relative terms, not necessarily yes.

The point is that Latham's and McCarty's correlations are simultaneously measuring two effects, the spatial coincidence of pairs of industries, and their tendency to agglomerate. For those who include both effects in their definition of geographical association, absolute data should be used. For those who discount agglomerative effects and define geographical association purely in relative terms, proportional data should be used.

**FOOTNOTES**

1. The modal distance, and not the median distance, is the measure of central tendency most appropriate for the purposes in hand.
2. See Chayes [1].
3. Relationships in a three variable closed system are summarized in Norcliffe [9, Ch. 14].
4. In Latham’s example, there are 199 sectors; in my study of local industrial complexes [11], 44 sectors are used.
5. For instance, 39 urban centres were included in the study of local industrial complexes in Ontario. Forty-six percent of the total workforce resided in the Toronto C.M.A., and sixty-eight percent in the four largest urban centres so that these few large towns overwhelmingly influence measures of geographical association based on absolute data.

**REFERENCES**