THE CONTIGUITY RATIO AND STATISTICAL MAPPING

by

R. C. GEARY

Introduction and Summary

The problem discussed in this paper is to determine whether statistics given for each "county" in a "country" are distributed at random or whether they form a pattern. The statistical instrument is the contiguity ratio c defined by formula (1.1) below, which is an obvious generalization of the Von Neumann (1941) ratio used in one-dimensional analysis, particularly time series. While the applications in the paper are confined to one- and two-dimensional problems, it is evident that the theory applies to any number of dimensions. If the figures for adjoining counties are generally closer than those for counties not adjoining, the ratio will clearly tend to be less than unity. The constants are such that when the statistics are distributed at random in the counties, the average value of the ratio is unity. The statistics will be regarded as *contiguous* if the actual ratio found is significantly less than unity, by reference to the standard error. The theory is discussed from the viewpoints of both randomization and classical normal theory. With the randomization approach, the observations themselves are the "universe" and no assumption need be made as to the character of the frequency distribution. In the "normal case," the assumption is that the observations may be regarded as a random sample from a normal universe. In this case it seems certain that the ratio tends very rapidly to normality as the number of counties increases. The exact values of the first four semi-invariants are given for the normal case. These functions depend only on the configuration, and the calculated values for Ireland, with number of counties only 26, show that the distribution of the ratio is very close to normal. Accordingly, one can have confidence in deciding on significance from the standard error.

The theory is also extended to regression problems. It is suggested that, if the dependent variables are found to be contiguous, the fact that the remainders after removal of the effect of independent variables are found to lack contiguity constitutes a prima facie case for regarding the independent variables included as *completely* explaining the dependent variables. There are, of course, other, and perhaps better, reasons for developing the regression aspects. If the theory is to be applied to problems of contagion (morbidity and

mortality rates or numbers), one cannot regard the fact of contagion in the narrow sense (i.e. that the disease has been transmitted by contacts) as established, or use the ratio as the measure of the strength of contagion, unless one has removed causative factors (independent variables) which may themselves have the property of contiguity. Contagion can only be established from the remainders when the effects of the causative factors have been duly allowed for. For instance, if a disease is known to vary according to social group, it is clearly necessary to correct for this effect which itself is very likely to be contiguous.

In the present paper, most of the applications are derived from Irish county data. Dublin (City and County) has been excluded because of the highly urbanized character of this area which renders most of the statistics exceptional. The greater number of the statistics examined exhibit the statistical property of contiguity in high degree. This property was fairly well known from ordinary mapping. For example, the counties to the north-west of a line from north Louth to west Cork have generally a higher proportion of the population in towns and villages than counties to the north-west. Even the agricultural characteristics of the two zones are significantly different. The writer, at this stage, is mainly concerned to see how the theory works with actual data, even though the fact of contiguity might have been anticipated. The ratio, moreover, establishes not only the fact but the relative strength of contiguity, and the exceptions have some interest.

In the final section of the paper proper, the statistical efficiency of the ratio, as a measure of contagion in the linear case, is discussed. It is found that the ratio is more efficient (by reference to a simple theoretical model of "contagion") than the "method of blocks" when the blocks contain but a few primary units, but the block method is to be preferred for larger sized blocks. In an appendix a method of orthogonalizing the independent variables by the use of latent roots and vectors is developed and applied to linear and quadratic terms of latitude and longitude of 25 Irish counties.

§ 1. The Contiguity Ratio—Randomization Aspect

Let the number of counties be n, the measure of the tth county z_t, with number of connexions k_t. The *contiguity ratio c* is given by

$$c = \frac{(n-1)}{2K_1} \frac{\sum'_{t \neq t'} (z_t - z_{t'})^2}{\Sigma_t (z_t - \bar{z})^2}, \tag{1.1}$$

where

$$\begin{cases} K_1 = \Sigma k_t \\ \Sigma \;= \text{sum over all counties;} \\ \Sigma' = \text{sum over contiguous counties.} \end{cases} \tag{1.2}$$

It is easy to show that

$$2a = \Sigma'(z_t - z_{t'})^2 = 2(\Sigma k_t z_t^2 - 2 \sum_{t < t'}{}' z_t z_{t'}). \tag{1.3}$$

The sampling theory of c can be discussed from two points of view: (i) randomization, or (ii) classical sampling theory which involves the assumption of universal normality of the z_t and also the concept of randomization.

Since the sum-product in the denominator of c, namely

$$b = \Sigma(z_t - \bar{z})^2, \tag{1.4}$$

is symmetrical in the z_t it assumes the same value for every permutation of the variables. Accordingly attention may be confined to

$$a = \Sigma k_t z_t^2 - 2 \sum_{t < t'}{}' z_t z_{t'}. \tag{1.5}$$

Denote the mean by the symbol M so that, where z and z' are any different pair from the series $z_1, z_2, \ldots, z_n$,

$$\begin{aligned} nM(z^2) &= \Sigma z_t^2 \\ n(n-1)M(zz') &= 2 \sum_{t < t'} z_t z_{t'} \end{aligned}$$

It is evident that, without loss of generality, it may be assumed that

$$nM(z) = \Sigma z_t = 0$$

Then

$$\begin{aligned}
&\text{(i)}\ n(n-1)M(zz') = 2\Sigma z_t z_{t'} = -\Sigma z_t^2 = -nM(z^2) \\
&\text{(ii)}\ n(n-1)M(z^2z') = \Sigma z_t^2 z_{t'} = -\Sigma z_t^3 = -nM(z^3) \\
&\text{(iii)}\ n(n-1)(n-2)M(z\,z'z'') = 6\Sigma z_t z_{t'} z_{t''} = 2\Sigma z_t^3 = 2nM(z^3) \\
&\text{(iv)}\ n(n-1)M(z^3z') = -\Sigma z_t^4 = -nM(z^4) \\
&\text{(v)}\ n(n-1)M(z^2z'^2) = 2\Sigma z_t^2 z_{t'}^2 = n^2[M(z^2)]^2 - nM(z^4) \\
&\text{(vi)}\ n(n-1)(n-2)M(z^2z'z'') = 2\Sigma z_t^2 z_{t'} z_{t''} \\
&\qquad\qquad = 2nM(z^4) - n^2[M(z^2)]^2 \\
&\text{(vii)}\ n(n-1)(n-2)(n-3)M(z\,z'z''z''') \\
&\qquad\qquad = 24\Sigma z_t z_{t'} z_{t''} z_{t'''} \\
&\qquad\qquad = 3n^2[M(z^2)]^2 - 6nM(z^4)
\end{aligned} \tag{1.6}$$

According to the randomization approach, the significance of the value of a, and hence of c, is judged by the position of the value actually found in the sequence of the $n!$ values of a (or c) found by permuting the n values of z in every possible way. From this point of view the $n!$ values are regarded as a frequency distribution with calculable moments. For simplicity of notation write

$$M(z^a z'^{a'} z''^{a''} \ldots) = (a\,a'a'' \ldots).$$

Then, from (1.5),

$$\begin{aligned} M(a) &= K_1\{(2) - (11)\} \\ &= K_1 n(2)/(n-1), \end{aligned} \tag{1.7}$$

from (1.6) (i), so that, from (1.1) and (1.3)

$$M(c) = 1 \tag{1.8}$$

Squaring a, as given by (1.5), and taking the mean of each term, bearing in mind that $\sum'_{t<t'}$ contains in all $K_1/2$ terms,

$$\begin{aligned} M(a^2) = {} & (4)\Sigma k_t^2 + 2(22) \sum_{t<t'} k_t k_{t'} - 4(31)\Sigma k_t^2 \\ & - 4(211)\Sigma k_t \left(\frac{K_1}{2} - k_t\right) + 2K_1(22) + 4(211)\Sigma k_t(k_t - 1) \\ & + 4(1111)\left\{\frac{K_1}{2}\left(\frac{K_1}{2} - 1\right) - \Sigma k_t(k_t - 1)\right\} \end{aligned}$$

$$\begin{aligned} = {} & K_2\{(4) - 4(31) - (22) + 8(211) - 4(1111) \\ & + K_1(K_1 + 2)\{(22) - 2(211) + (1111)\} \end{aligned} \tag{1.9}$$

where

$$K_1 = \Sigma k_t$$
$$K_2 = \Sigma k_t^2.$$

As a check it will be noted that the sums of the coefficients of the expressions in the brackets { } of (1.9) are zero. This is because when all the z are equal the moments from zero (say (4)′, (31)′, etc.) are all equal so that $M(a^2)$ is zero as it should be since, in this particular case, each of the a's is zero. Finally, using the last four relations of (1.6), $M(a^2)$ is given by

$$\begin{aligned} (n-1)(n-2)(n-3)M(a^2) = {} & K_2\{n(n^2 - n + 2)(4) \\ & - n(n^2 + 3n - 6)(2)^2\} + K_1(K_1 + 2)\{- n(n-1)(4) \\ & + n(n^2 - 3n + 3)(2)^2\} \end{aligned} \tag{1.10}$$

From (1.1), (1.3) and (1.10), the $M(c^2)$ may be computed as the product of $M(a^2)$ by $(n-1)^2/n^2K_1^2(2)^2$. It may be observed that, in practical applications, K_1 and K_2 will each be of order n, so that, as n tends towards infinity $M(c^2)$ will tend towards a finite limit. In fact, if

$$\begin{aligned} K_1 &= nk_1 \\ K_2 &= nk_2 \end{aligned} \tag{1.11}$$

$$M(c) = 1 \tag{1.12}$$
$$M(c^2) \sim 1 + \frac{1}{n}\left\{\left(\frac{k_2}{k_1^2} - 1\right)\left[\frac{(4) - (2)^2}{(2)^2}\right] + \frac{2}{k_1}\right\}$$

If the z can be regarded as a normal sample, (4) will be approximately equal to $3(2)^2$ so that

$$\operatorname{Var}(c) \sim \frac{2}{n}\left(\frac{k_2}{k_1^2} + \frac{1}{k_1} - 1\right) \tag{1.13}$$

In § 3 the regression aspects of the problem are discussed from the viewpoint of classical linear theory. The problem is to determine if there is a contiguity effect, i.e. if c has a significantly low value after the elimination of q independent variables by the least square method. As far as randomization is concerned, it would appear that the test developed in this section can be applied formally, the z being the remainders after the contributions of the independent variables have been removed. To a certain extent the writer shares the misgivings of some other students about the validity of the randomization approach in its application to regression remainders. As each successive independent variable is removed, should not the degrees of freedom be diminished? It does not seem so. What happens is that the variance (or range) of the remainders diminish as the effect of each independent variable is allowed for, the test becoming indeterminate when the number of independent variables (originally with mean zero) is one less than the number of observations n, i.e. when all the remainders are zero. Accordingly the formal application of the randomization procedure, without diminution of the number of degrees of freedom, does not result in *obvious* inconsistency: we can conceive of cases where c will be significantly low even after removal of the effect of $(n - 2)$ independent variables. Since doubts remain, however, the writer considered it desirable to examine the problem from the classical sampling aspect. In any case it will be interesting to compare the results of the two approaches. In the practical aspect the randomization method has the advantage that it can be applied without the assumption of universal normality in the n observations, regarded as a random sample.

The Two-category Case

This is the case in which only two values (which, without loss of generality, may be taken as one and zero) appear in the county scheme. The present theory in its randomization aspect can be applied literally. The problem here is to determine if the two values are distributed at random in the pattern, or if, on the contrary, there is grouping or coagulation. If the number of counties is, as before, n and if the number of ones is np and the number of zeros nq, we require only the values of the moments from the mean, namely (2) and (4). These are

$$(1.14) \qquad \begin{aligned} (2) &= pq \\ (4) &= pq(p^2 - pq + q^2) \\ &\text{with } p + q = 1. \end{aligned}$$

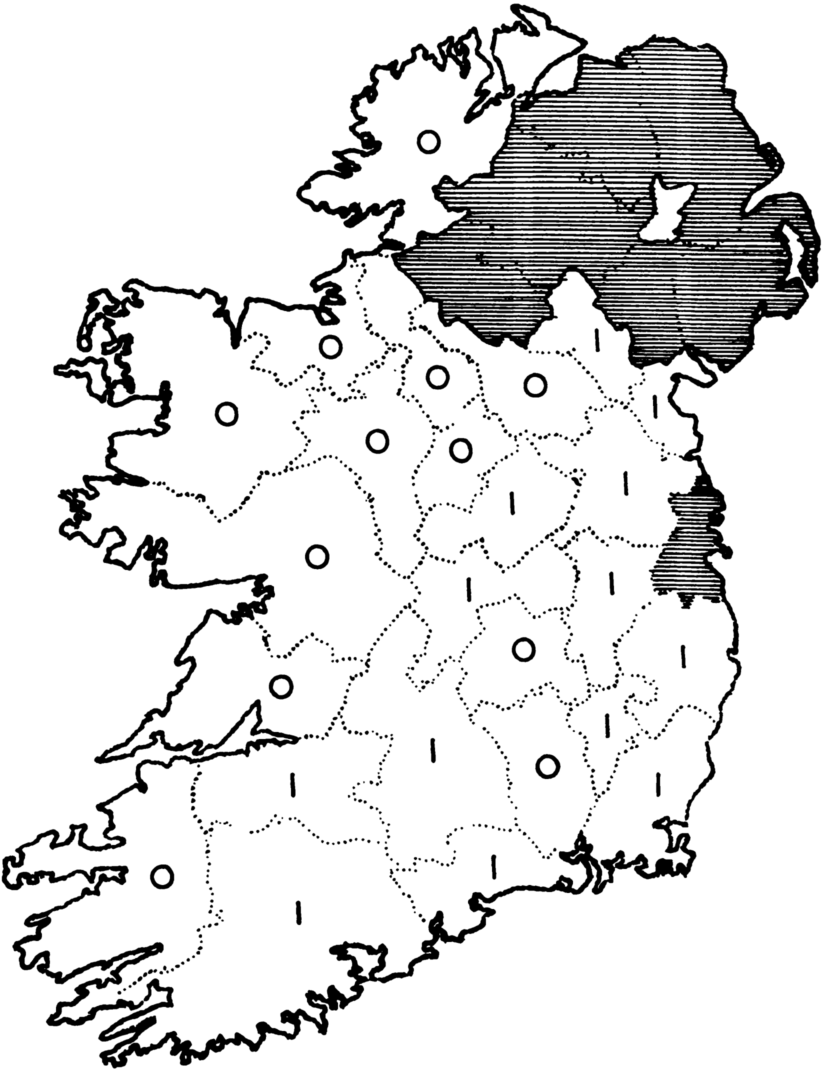
The randomization mean $M(c) = 1$ and $M(c^2)$ is found from (1.1), (1.3) and (1.10).

Example: Has the Irish 25-county scheme illustrated in Map I a pattern, or are the units and zero distributed at random?

For the 25-county scheme, $n = 25$, $K_1 = 110$, $K_2 = 544$. For the particular example $np = 13$, $nq = 12$ and from (1.1) the actual value of c is found to be 0·6993. The variance is 0·014069 so that

MAP 1

TWO-CATEGORY EXAMPLE FOR TWENTY-FIVE COUNTIES



the standard error is 0·1186 and $R = (1 - c)/\text{S.E.} = 2{\cdot}54$ which should be judged significant.

§ 2. Classical Theory: the Frequency Distribution of *c* in the Normal Case

Let the z_i be a random sample of n from a normal universe with mean zero and (unknown) variance σ^2. Since c is the quotient of a quadratic form by the estimated variance, it is known (Geary, 1933) that the moments from zero of the ratio are the quotients of the moments of the numerator and denominator for normal supplies. Following are the first four moments of c:

$$\begin{aligned}
&\mu'_1 = 1\\
&\mu'_2 = \{n^2k_1^2 + 2n(k_1 + k_2)\}(n - 1)/n^2(n + 1)k_1^2\\
(2.1)\quad &\mu'_3 = \{n^3k_1^3 + 6n^2k_1(k_1 + k_2) + 8n(3k_2 + k_3 - 6t')\}\\
&\qquad\qquad \times (n - 1)^2/n^3(n + 1)(n + 3)k_1^3\\
&\mu'_4 = \{n^4k_1^4 + 12n^3k_1^2(k_1 + k_2) + 4n^2(3k_1^2 + 30k_1k_2 + 8k_1k_3\\
&\qquad\qquad - 48k_1t' + 3k_2^2) + 48n(- k_1 + 2k_2 + 6k_3 + k_4\\
&\qquad\qquad - 2v_1 + 8v_2 - 8v_3 + 8q')\}\\
&\qquad\qquad\quad \times (n - 1^3/n^4(n + 1)(n + 3)(n + 5)k_1^4
\end{aligned}$$

where

$$\begin{aligned}
&n = \text{number of counties}\\
&nk_a = \Sigma_i k_i^a,\ a = 1, 2, 3, 4\\
&nv_1 = \sum_{i<j}{}' (k_i + k_j)^2\\
&nv_2 = \sum_{i<j}{}' k_ik_j\\
(2.2)\quad &nv_3 = \Sigma_i k_i T_i \ \left\{ \begin{array}{l}\text{where } T_i \text{ is the number of triads at county } i\\ \text{and } T \text{ the total number of triads in the system}\end{array}\right.\\
&nt' = T\\
&nq' = Q \text{ the total number of quartets in the system.}
\end{aligned}$$

The meaning of the terms *triad* and *quartet* (as well as connexions) will best be understood by reference to Map II (of Ireland—Twenty-six Counties).

The capital letters A—Z indicate the counties arranged in alphabetical order from A—Carlow to Z—Wicklow. The numbers on the map indicate the *connexions*. Thus A is connected with I, J, K, Y, Z, = 5. Clearly the total number of connexions in the system is half the total of the numbers of the map. As regards *triads*, there are 5 at A, namely AIK, AIZ, AJK, AJY, AYZ. Quartets at A number 9, namely AFIZ, AIJK, AIKS, AIKZ, AIYZ, AJKV, AJKY, AJWY, AJYZ. Triads and quartets enter into the calculations of moments because in raising $\sum_{ij} z_iz_j$ in a to the third and fourth powers respectively, the product terms like $(z_iz_j)(z_{i'}z_{j'})(z_{i''}z_{j''})$ and $(z_iz_j)(z_{i'}z_{j'})(z_{i''}z_{j''})(z_{i'''}z_{j'''})$ make non-zero contributions when triads and quartets are involved. For Ireland (26 Counties) the values of (2.2) are as follows:—

$$\begin{array}{rlrl} n = & 26 & nv_1 = & 6{,}168 \\ nk_1 = & 116 & nv_2 = & 1{,}472 \\ nk_2 = & 584 & nv_3 = & 519 \\ nk_3 = & 3{,}224 & nt' = & 33 \\ nk_4 = & 19{,}184 & nq' = & 42 \end{array}$$

The values of (2.1) are then:—

$$\begin{aligned} \mu_1' &= 1 \\ \mu_2' &= 1{\cdot}02226186 \\ \mu_3' &= 1{\cdot}06690364 \\ \mu_4' &= 1{\cdot}13553261 \end{aligned}$$

The values of the semi-invariants are

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &= 0{\cdot}0222619 \\ \lambda_3 &= 0{\cdot}0001179 \\ \lambda_4 &= 0{\cdot}0000028 \end{aligned}$$

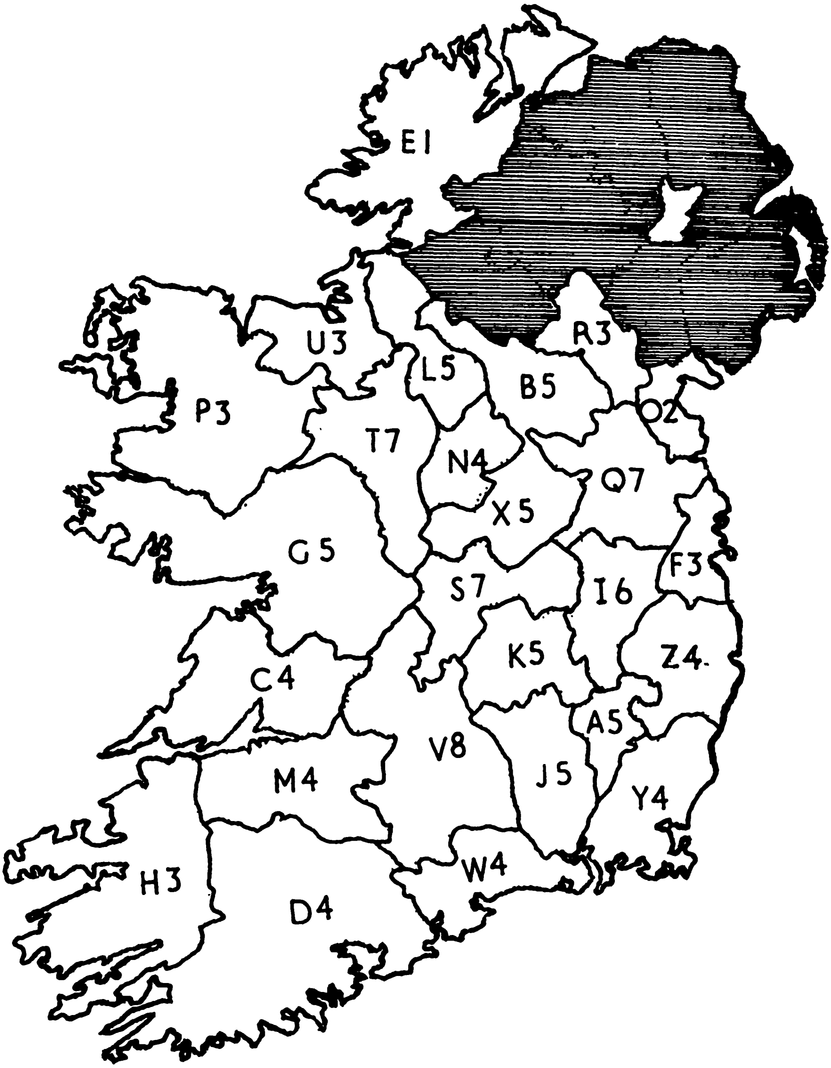
The values of $\sqrt{\beta_1} = \lambda_3/\lambda_2{}^{3/2}$ and $\beta_2 = \lambda_4/\lambda_2^2$ are respectively 0·0355 and 0·0057 so that, with n only 26 and in a system which may be regarded as well-diversified the distribution is obviously very close to normal. The writer does not consider that it is necessary to furnish a formal proof of the normality of c for n indefinitely large since, from the practical point of view, it is more important to establish that, for the sample sizes and the kinds of situation which are encountered, the assumption of normality is plausible, so that the standard deviation can be used as a test of significance.

To terms in n^{-4}, the values of the semi-invariants, computed from (2.1) and (2.2), are as follows:—

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &\sim \frac{2}{nk_1^2}(k_1 - k_1^2 + k_2) - \frac{2}{n^2k_1^2}(2k_1 - k_1^2 + 2k_2) \\ &\quad + \frac{2}{n^3k_1^2}(2k_1 - k_1^2 + 2k_2) - \frac{2}{n^4k_1^2}(2k_1 - k_1^2 + 2k_2) \\ \lambda_3 &\sim \frac{8}{n^2k_1^3}(2k_1^3 - 3k_1^2 - 3k_1k_2 + 3k_2 + k_3 - 6t') \\ &\quad - \frac{8}{n^3k_1^3}(8k_1^3 - 15k_1^2 - 15k_1k_2 + 18k_2 + 6k_3 - 36t') \\ &\quad + \frac{8}{n^4k_1^3}(26k_1^3 - 51k_1^2 - 51k_1k_2 + 66k_2 + 22k_3 - 132t') \end{aligned} \tag{2.3}$$

MAP II

TWENTY-SIX COUNTY MAP SHOWING NUMBER OF CONNECTIONS



$$
\begin{aligned}
\lambda_4 \sim \frac{48}{n^3k_1^4}&(5k_1^4 - 10k_1^3 - 10k_1^2k_2 + 2k_1^2 + 16k_1k_2 + 4k_1k_3 \\
&- 24k_1t' + 2k_2^2 + k_1 - 2k_2 - 6k_3 - k_4 + 2v_1 \\
&- 8v_2 + 8v_3 - 8q') \\
+ \frac{48}{n^4k_1^4}&(43k_1^4 - 96k_1^3 - 96k_1^2k_2 + 20k_1^2 + 172k_1k_2 \\
&+ 44k_1k_3 - 264k_1t' + 20k_2^2 + 12k_1 - 24k_2 - 72k_3 \\
&- 12k_4 + 24v_1 - 96v_2 + 96v_3 - 96q')
\end{aligned}
$$

Clearly β_1 and β_2 are $0(n^{-1})$, lending a measure of verisimilitude to the algebra. For Ireland (26 Counties) the approximation to λ_4 is $0{\cdot}337n^{-3} - 10{\cdot}42n^{-4}$. It may be remarked that the co-efficients of n^{-3} and n^{-4}, though they have a sharply increasing tendency, are very small in relation to the contributions of the different terms.

§ 3. Regression Aspects of Contiguity

The problem is to determine whether the value of c given by (1.1) is significantly small when the z_t are the remainders when the effect of a number of independent variables have been removed from the original dependent variable under examination. The typical procedure would consist in first establishing by the c test that the original observations were contiguous. The regression between the original observations and a series of correlative observations would be determined by least square procedure. The remainders would then be tested for contiguity. If the original observations were highly contiguous and the remainders not significantly so, this might be a good test for the thesis that the independent variables completely "explain" the observations. As a more practical application, if mortality or morbidity rates were being examined for evidence of contagion, given rates for n districts (parts of a city or county for example), it would be highly desirable to correct the rates for, say, independent variables such as income level, density of population, housing conditions, etc., each of which may also be significantly contiguous.

The model will be the usual linear one. In matrix notation, let $\boldsymbol{y}$ represent the $1 \times n$ matrix of original observations, $\boldsymbol{x}$ the $q \times n$ matrix of independent variables, q being the number of independent variables, $\boldsymbol{a}$ and $\boldsymbol{b}$ the matrices of coefficients determined by least squares. The absolute matrix will be (a, a, a, . . . n terms), i.e. $1 \times n$, whereas $\boldsymbol{b}$ will be $1 \times q$. The remainder matrix $\boldsymbol{z}$ will be $1 \times n$, i.e.

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{bx} + \boldsymbol{z} \tag{3.1}$$

Without loss of generality it may be assumed that the mean of each of the elements in $\boldsymbol{x}$ is zero. Then

$$\text{(3.2)} \qquad \boldsymbol{a} = \boldsymbol{yN}$$

where $\boldsymbol{N}$ is the $n \times n$ matrix, all of whose elements are $1/n$, and

$$\text{(3.3)} \qquad \boldsymbol{yx}' = \boldsymbol{bxx}'$$

where $\boldsymbol{x}'$ is the transpose of $\boldsymbol{x}$. Since the means of all the elements in $\boldsymbol{x}$ are zero

$$\text{(3.4)} \qquad \boldsymbol{xN} = \boldsymbol{0}$$

The matrix $\boldsymbol{y}$ is also given by

$$\text{(3.5)} \qquad \boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\beta x} + \boldsymbol{u},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the (usually unknown) population or universal mean values of $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively while $\boldsymbol{u}$, of dimensions $1 \times n$, is a random normal sample of mean zero and (unknown) variance σ^2. It is first necessary to express $\boldsymbol{z}$ in terms of $\boldsymbol{u}$. From the foregoing relations

$$\begin{aligned}\text{(3.6)}\ \boldsymbol{z} &= \boldsymbol{y} - \boldsymbol{a} - \boldsymbol{yx}'(\boldsymbol{xx}')^{-1}\boldsymbol{x} \\ &= \boldsymbol{\alpha} + \boldsymbol{\beta x} + \boldsymbol{u} - (\boldsymbol{\alpha} + \boldsymbol{\beta x} + \boldsymbol{u})\boldsymbol{N} - (\boldsymbol{\alpha} + \boldsymbol{\beta x} + \boldsymbol{u})\boldsymbol{x}'(\boldsymbol{xx}')^{-1}\boldsymbol{x} \\ &= \boldsymbol{u}[\boldsymbol{I} - \boldsymbol{N} - \boldsymbol{X}],\end{aligned}$$

where $\boldsymbol{I}$ is the unit $n \times n$ matrix and $\boldsymbol{X} = \boldsymbol{x}'(\boldsymbol{xx}')^{-1}\boldsymbol{x}$. Since the square of the symmetrical $n \times n$ matrix $[\boldsymbol{I} - \boldsymbol{N} - \boldsymbol{X}]$ is equal to itself the universal variance-covariance matrix of $\boldsymbol{z}$, namely $\boldsymbol{V}$, is given by

$$\text{(3.7)} \qquad \boldsymbol{V} = \sigma^2[\boldsymbol{I} - \boldsymbol{N} - \boldsymbol{A}],$$

or

$$\text{(3.8)} \qquad \frac{1}{\sigma^2} E(z_t z_{t'}) = \delta_{tt'} - \frac{1}{n} - \frac{1}{n} \sum_{i,i'} x_{it} s_{ii'} x_{i't'}$$

where x_{it} are the elements of x and $s_{ii'}$ the elements of $(xx')^{-1}$, and $\delta_{tt'} = 1$ for $t = t'$ and zero for $t \neq t'$. Using the relations $\Sigma_t x_{it} = 0$ it follows that

$$\text{(3.9)} \quad \frac{1}{\sigma^2} EM(z_t^2) = \frac{1}{\sigma^2} E\left(\frac{1}{n} \Sigma_t z_t^2\right) = 1 - \frac{1}{n} - \frac{q}{n} = \frac{(n - q - 1)}{n}$$

and, for $t \neq t'$,

$$\begin{aligned}\text{(3.10)} \quad \frac{1}{\sigma^2} EM(z_t z_{t'}) &= \frac{1}{\sigma^2} E\left(\frac{1}{n(n-1)} \sum_{t \neq t'} z_t z_{t'}\right) = -\frac{1}{n} + \frac{q}{n(n-1)} \\ &= \frac{-(n - q - 1)}{n(n-1)}\end{aligned}$$

The foregoing formulae in this section are, of course, well known. It has been judged expedient to develop them in some detail because they illustrate in the simplest case the processes by which the variance of c is derived. The universe which imparts variability to c is the resultant of two processes (1) the independent variables being fixed and ordered the variability being due to the normal variate u, superimposed on which is (2) variability due to the $n!$ permutations of the independent variables as a group. The combined process as it relates to the derivation of universal means is represented by the operational "product" EM, E relating to process (1) and M to process (2). It will be noted from (3.9) and (3.10) that the process yields the classical results in its application to the variance-covariance.

Formula (3.6) shows that, in the conditions specified, z is a normal variate with variance-covariance matrix given by (3.8) which, it will be noted, is a function of the order subscripts t and t'. Given the order the moments from zero of c, with

$$c = P\,a/b', \tag{3.11}$$

where

$$\begin{aligned} a &= \Sigma_t k_t z_t^2 - 2\sum_{t<t'}{}' z_t z_{t'} \\ b' &= \Sigma_t z_t^2 \\ P &= (n-1)/K_1, \end{aligned} \tag{3.12}$$

the moment of any degree of c, i.e. $E(c^k)$, is the quotient of the moment of the numerator by the moment of the denominator. The final universal moments, on permutation of the orders represented by the process $ME = EM$, is found as the simple average of the E process since at the second stage of averaging the denominators, which are symmetrical functions of the variables z and hence of the orders, are equal for all permutations. Symbolically, as regards any symmetrical function, $EM = E$. We now have

$$\begin{aligned} \mu_1' &= EM(c) = PEM(a)/E(b') \\ &= \frac{\sigma^2(n-1)}{K_1} \times \frac{K_1(n-q-1)}{(n-1)}/\sigma^2(n-q-1) \end{aligned} \tag{3.13}$$

or

$$\mu_1' = 1$$

The second moment μ_2' of c is given by

$$\mu_2' = ME(c^2) = P^2\{ME(a^2)\}/E(b')^2 \tag{3.14}$$

Since b' is a normal variance with $(n-q-1)$ degrees of freedom it is distributed as $\sigma^2\chi^2$, so that, in the last term of (3.14)

$$\frac{P}{E(b')^2} = \frac{(n-1)^2}{K_1^2\sigma^4(n-q+1)(n-q-1)} \tag{3.15}$$

To find $ME(a^2)$ we take the square of a given by (3.12) and perform in succession the operations E and M on the various terms. The result is an expression of the form (1.9) where $(a\ b\ c\ .\ .\ .)$ represents $ME(z_t^a\, z_{t'}^b\, z_{t''}^c\ .\ .\ .)$, $t < t' < t''\ .\ .\ .\ .$ Now z_t as a linear function of the normal variable u (see 3.6) is itself normally distributed with mean zero. Hence its moments of any dimension and of any even degrees are functions of the variances and covariances of the z_t where the order of the subscripts, prior to the operation M, is fixed. We require only the following relations

$$
\begin{aligned}
\mu(4) &= 3[\mu(2)]^2 \\
\mu(31) &= 3\mu(20)\mu(11) \\
\mu(22) &= \mu(20)\mu(02) + 2[\mu(11)]^2 \\
\mu(211) &= \mu(200)\mu(011) + 2\mu(101)\mu(110) \\
\mu(1111) &= \mu(1100)\mu(0011) + \mu(1010)\mu(0101) \\
&\qquad + \mu(1001)\mu(0110)
\end{aligned}
\tag{3.16}
$$

The operation M is then performed on the various product terms on the right of (3.16). After M it is obvious by symmetry that each of the three product terms in $\mu(1111)$ yields the same result.

To illustrate the process, consider $\mu(211)$. It will be convenient to regard the independent variables as orthogonalized by a non-singular linear transformation,* the same for all t, so that the matrix xx' reduces to a diagonal $q \times q$ matrix $n\sigma_1^2, n\sigma_2^2, \ldots, n\sigma_q^2$ and (3.8) becomes

$$
\begin{aligned}
E(z_t^2) &= \frac{\sigma^2}{n}(n - 1 - \Sigma_i \xi_{it}^2) \\
E(z_{t'} z_{t''}) &= \frac{\sigma^2}{n}(-1 - \Sigma_i \xi_{it'} \xi_{it''}) \\
\xi_{it} &= x_{it}/\sigma_i \\
n\sigma_i^2 &= \Sigma_t x_{it}^2.
\end{aligned}
\tag{3.17}
$$

Therefore

$$
\begin{aligned}
&\frac{n^3(n-1)(n-2)}{2} M\{\mu(200)\mu(011)\} \\
&= n^2 \Sigma_t E(z_t^2) \sum_{t' \neq t'' \neq t} E(z_{t'} z_{t''}) \\
&= n \Sigma_t E(z_t^2) \left\{ -(n-1)(n-2) + \sum_i \sum_{t' \neq t} \xi_{it'}(\xi_{it} + \xi_{it'}) \right\},
\end{aligned}
\tag{3.18}
$$

using the second of (3.17) and the relation $\Sigma_t \xi_{it} = 0$. Using the further relation $\Sigma_t \xi_{it}^2 = n$ the second term in the brackets { } becomes

$$qn - 2\Sigma_i \xi_{it}^2.$$

Then, on substituting in the right side of (3.18) for $E(z_t^2)$ given by the first formula in (3.17) and summing for t, we find

$$
\begin{aligned}
M\{\mu(200)\mu(011)\} = \frac{\sigma^2}{n^2(n-1)(n-2)} \{&-(n-1)^2(n-2) \\
&+ 2q(n-1)(n-2) - q^2 n + 2\gamma\}
\end{aligned}
\tag{3.19}
$$

with

* See Appendix.

$$n\gamma = \Sigma_t(\Sigma_i \xi_{it}^2)^2 \tag{3.20}$$

All the other expressions required are derived in a similar way. Finally

$$\begin{aligned} ME(a^2) \times \frac{n(n-1)(n-2)(n-3)}{\sigma^4} &= K_2\{2n(n-1)(n-2)(n-3) - 2q(2n^3 - 9n^2 - 15n - 6) \\ &\quad - q^2(n^2 + 3n - 6) + 3\gamma(n^2 - n + 2)\} \\ &\quad + (K_1^2 + 2K_1)\{n(n-1)(n-2)(n-3) \\ &\quad - 2q(n^3 - 6n^2 + 9n - 3) + q^2(n^2 - 3n + 3) \\ &\quad - 3\gamma(n-1)\} \end{aligned} \tag{3.21}$$

As a check on (3.21), consider the use of $q = n - 1$. Then the least squares fit to the observations y_t will be exact, so that all the remainder terms z_t will be zero and, from the first formula of (3.17), $\Sigma_i\xi_{it}^2 = n - 1$. Hence from (3.20) $\gamma = (n-1)^2$. Substitution of these expressions for q and γ in (3.21) gives zero for the coefficients of K_2 and $(K_1^2 + 2K_1)$ as it should. From (3.14), (3.15) and (3.21) $\mu_2' = ME(c^2)$ is found. The final expression is, of course, free of the error variance σ^2.

In (3.21) γ is a function of the orthogonalized independent variables. In terms of the original variables γ is given in matrix notation by

$$n\gamma = \Sigma_t\{\boldsymbol{x}_t'(\boldsymbol{x}\boldsymbol{x}')^{-1}\boldsymbol{x}_t\}^2 \tag{3.22}$$

where $\boldsymbol{x}_t$ is the $q \times 1$ column matrix in the $q \times n$ matrix $\boldsymbol{x}$. It may be of interest to add that $\gamma = 42{\cdot}169$ when $n = 25$, $q = 5$ when the ξ_{it} are the quadratic orthogonal latitude-longitude terms, the "quolls" described in the Appendix.

§ 4. Contagion: Efficiency of Contiguity Ratio in the Linear Case

The contiguity ratio c can, of course, be used to establish the fact, and to measure the degree, of contagion. The object of the present section is to compare the statistical efficiency of c as a measure of contagion ("Method I") with that of another method ("Method II") which will presently be described.

Attention will be confined to the linear case. The Method I mathematical model in the case of no contagion is as follows. A straight line of given length is divided into n equal divisions (e.g. a street or streets, each division representing a house). Each division is assigned a number 1 or 0 with probabilities p and q $(= 1 - p)$ respectively, each assignment being independent. In the linear case

$$\begin{aligned} K_1 &= 2 + 2(n-2) = 2(n-1) = nk_1 \\ K_2 &= 2 + 4(n-2) = 2(2n-3) = nk_2 \end{aligned} \tag{4.1}$$

The moments (2) and (4) are given by

$$(4.2) \qquad \begin{aligned}(2) &= pq \\ (4) &= pq(p^2 - pq + q^2)\end{aligned}$$

The mathematical model of contagion will be constructed by dividing the line into two parts* containing n_1 and n_2 divisions so that

$$n_1 + n_2 = n.$$

In the first part the probability of assigning 1 in each division will be p_1 and in the second part p_2, so that the over-all value of p as computed from the "observations" will be

$$(4.3) \qquad p = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$$

If the probability p obtained uniformly throughout the $n = n_1 + n_2$ divisions the average value of $M(c)$, from (1.1) and (1.7) will, of course, be unity. This is the "nul-hypothesis" case, in which, from (1.7),

$$(4.4) \qquad \begin{aligned}M_0(a) &= 2(n-1)npq/(n-1) \\ &= 2npq\end{aligned}$$

In the "actual" case the value will be

$$(4.5) \qquad M(a) \sim 2n_1p_1(1 - p_1) + 2n_2p_2(1 - p_2)$$

when both n_1 and n_2 are large, as they will be assumed to be in the rest of this section. Then, from (4.3) and (4.4)

$$(4.6) \quad M_0(a) - M(a) \sim \frac{2n_1n_2}{(n_1 + n_2)}(p_1 - p_2)^2,$$

and

$$(4.7) \quad M_0(c) - M(c) = 1 - M(c) \sim \frac{n_1n_2(p_1 - p_2)^2}{n^2pq}$$

Finally we need the standard error in the nul-hypothesis case. From (1.12), since $k_2/k_1^2 \sim 1$ and $2/k_1 \sim 1$,

$$(4.8) \qquad \mathrm{Var}\,(c) \sim 1/n$$

We now introduce the *sensitivity S* defined as the ratio of the average deviation (4.7) to the nul-hypothesis standard error, so that

$$(4.9) \qquad S \sim n_1n_2(p_1 - p_2)^2/n^{3/2}pq$$

Method II envisages the divisions grouped in m blocks of d divisions each, so that $n = dm$, d being a small fixed number so that m may be regarded as of the same order of magnitude as n. The test of contagion according to this method consists in comparing the number of blocks in $(d + 1)$ classes according to the number of units in each block with the theoretical distribution on the

* The theory applies to a division into any number of parts when n (and m) are large, provided that such number remains finite as n tends towards infinity and provided, of course, that only two probabilities, p_1 and p_2 apply to the different parts.

assumption of no contagion using χ^2 with d degrees of freedom. In this "nul-hypothesis" case the probability of x will be the binomial*

$$(4.10) \qquad \phi_x = \binom{d}{x} p^x q^{d-x}$$

For the "contagion" case, the line, as before, is divided into two parts containing now m_1 and m_2 blocks, so that $n_1 = dm_1$ and $n_2 = dm_2$. The theoretical probability of x units in this case will be

$$(4.11) \qquad \phi'_x = \binom{d}{x} (\pi_1 p_1^x q_1^{d-x} + \pi_2 p_2^x q_2^{d-x})$$

where $\pi_1 = m_1/(m_1 + m_2)$, $\pi_2 = m_2/(m_1 + m_2)$, $\pi_1 + \pi_2 = 1$.

If the corresponding "actual" proportionate frequency (i.e. that found by classifying the blocks in a single experiment) be f'_x then the appropriate value of χ^2 is

$$(4.12) \qquad \chi^2 = m \sum_{x=o}^{d} (f'_x - \phi_x)^2/\phi_x,$$

which, since $\Sigma_x f'_1 = 1 = \Sigma_x \phi_x$, gives

$$(4.13) \qquad (\chi^2/m) + 1 = \Sigma_x f'^2_x/\phi_x.$$

Since

$$(4.14) \qquad m^2E(f'^2_x) = m(m-1)\Sigma\phi'^2_x + m\phi'_x,$$

$$(4.15) \qquad m^2E(1 + \chi^2/m) = m(m-1)\Sigma\phi'^2_x + m\Sigma\phi'_x/\phi_x.$$

Substituting for ϕ_x and ϕ'_x as given by (4.10) and (4.11), the right side of (4.15) becomes

$$(4.16) = m(m-1)\left\{\pi_1^2\left(\frac{p_1^2}{p} + \frac{q_1^2}{q}\right)^d + 2\pi_1\pi_2\left(\frac{p_1p_2}{p} + \frac{q_1q_2}{q}\right)^d + \pi_2^2\left(\frac{p_2^2}{p} + \frac{q_2^2}{q}\right)^d\right\} + m\left\{\pi_1\left(\frac{q_1}{q}\right)^d\left[1 - \left(\frac{p_1q}{pq_1}\right)^{d+1}\right]\Big/\left(1 - \frac{p_1q}{pq_1}\right) + \pi_2\left(\frac{q_2}{q}\right)^d\left[1 - \left(\frac{p_2q}{pq_2}\right)^{d+1}\right]\Big/\left(1 - \frac{p_2q}{pq_2}\right)\right\}$$

The required value of $E\chi^2$ is found at once from (4.15) and (4.16). It will be seen that if $p_1 = p_2 = p$, $E\chi^2 = d$, the number of degrees of freedom, as, of course, it should.

To compare the efficiency of the two methods, calculations of the associated probabilities were made for various sets of values of p_1 and p_2 and for block sizes 2, 4, and 8. The nul-hypothesis value of p was taken as $(\pi_1p_1 + \pi_2p_2)$ throughout. To find the number of divisions n, S, given by (4.9), is written in the form

$$(4.17) \qquad S = \pi_1\pi_2(\ p_1 - p_2)^2n^{\frac{1}{2}}/pq$$

* By analogy with Method I, randomization procedure should also be envisaged in Method II, in which case the frequency is $\phi_x + 0(m^{-1})$, when both p and q have fixed values.

and equated to its assumed critical value 2, which corresponds to a normal probability of 0·0455, i.e. it is assumed that c is approximately normally distributed. This gives n as

$$n = 4p^2q^2/\pi_1^2\pi_2^2(p_1 - p_2)^4 \tag{4.18}$$

and then m is taken as n/d. Finally the value of $E\chi^2$ is computed from (4.16) and its value compared with the χ^2 corresponding to the probability 0·0455. For each example $\pi_1 = 0{\cdot}7$, $\pi_2 = 0{\cdot}3$.

The results of the computation are shown in Table 1.

TABLE 1

Comparison of Values of χ^2 Corresponding to Probability 0·0455 with Values of $E\chi^2$ (Formula (4.16)) for Five Examples each with Three Block Sizes (d)

Example	Assumed Values of p_1, p_2		Number (d) of Divisions in Block					
			2		4		8	
	p_1	p_2	χ^2	$E\chi^2$	χ^2	$E\chi^2$	χ^2	$E\chi^2$
1	0·1	0·2	6·19	3·57	9·71	8·98	15·79	19·23
2	0·2	0·1		3·95		10·43		26·23
3	0·2	0·4		4·08		11·25		38·30
4	0·4	0·2		4·05		10·61		26·83
5	0·4	0·6		4·18		10·46		19·98

Method I will be regarded as more efficient than Method II at the probability level 0·0455 if the χ^2 corresponding to this probability is greater than $E\chi^2$. It will be seen that for block sizes up to about 4, Method I is at least as efficient as Method II, but for $d > 4$ Method II is to be preferred. This is a rather unexpected result: one might have thought that Method I would always be superior because it *seems* to use more information than does Method II which "blankets" all the units within the block into a single figure (total number of units), whereas Method I takes account of contagion (or contiguity) within blocks. For instance, if 3 units are found in a block of 6 houses, Method II simply takes account of the total of 3 whereas in Method I the contiguity total can range from 1 (when the 3 units are together at either end) to 5 (when no two units are contiguous). The writer does not understand why the efficiency of Method II seems to increase with block size. There must, in practice, be some limit to the block size of greater efficiency. It will be recalled that the number of blocks decreases in inverse proportion to block size. A point will be reached when there will not be a sufficient number of blocks for the χ^2-distribution to be deemed to apply: this may be the solution of the anomaly. Of course, the theory of infection being investigated will probably impose its own block size, but, though a very simple model has been assumed in the present case, the investigator should lean towards favouring larger rather than smaller block sizes, if he can.

It is realized that a power function analysis would be more rigorous for comparing the efficiencies of the two methods. This would have involved the determination of the approximate frequency distribution of χ^2 given by (4.12) or (4.13). The mean value of χ^2 has been given. For students who might be interested in pursuing the power function aspect, it will be useful to place on record the value of the second moment. It is derivable from

$$(4.19)\quad m^4E(1+\chi^2/m)^2$$
$$= m_4(\Sigma_x\phi_x'^2/\phi_x)^2 + 2m_3\left(\Sigma\frac{\phi_x'^2}{\phi_x}\Sigma\frac{\phi_x'}{\phi_x} + 2\Sigma\frac{\phi_x'^3}{\phi_x^2}\right)$$
$$+ m_2\left\{\left(\Sigma\frac{\phi_x'}{\phi_x}\right)^2 + 6\Sigma\frac{\phi_x'^2}{\phi_x^2}\right\} + m_1\Sigma\phi_x'/\phi_x^2$$

with

$$m_r = m(m-1)(m-2)\ .\ .\ .\ (m-r+1)$$

Furthermore the author fully realizes that the model of "contagion" for the present purpose is far too rudimentary as a theory of contagion, considered *per se*. His object was only to form some impression of the efficiency of c in this application, for c can be applied in actual cases, however complicated the manner of spread of contagion.

§ 5. Applications

Twelve county series were selected for examination: they are displayed in the accompanying Table 2. Many of the figures for Dublin (City and County) are so exceptional on account of the highly urbanized character of that area that it was decided to exclude it from the calculation of the contiguity c, shown at the foot of the page. The number 0·1512 used for computing the significance R is the standard error computed from formula (2.1). It will be recalled that for this formula universal normality is assumed. On the other hand the randomization procedure gives for c, for 25 Irish counties, mean unity and variance

$$\text{Var}\ (c) = 0{\cdot}00498737 \times \frac{(4)}{(2)^2} + 0{\cdot}00904961$$

This, in contradistinction to the normal value, varies with each series. If $(4)/(2)^2$ has the normal value 3 the variance (randomization theory) is 0·02401172 so that the standard error is 0·1550, very similar to the normal theory value. Except in the "unity-zero" (two category) and other cases where the distribution is extremely non-normal, the difference between normal theory and randomization

TABLE 2

Twelve Statistical Series for Irish Counties, showing Value of Contiguity c and Significance Ratio R

Serial Letter	County (incl. Co. Borough)	Percentage Number Agricultural Holdings in Valuation Groups (1950)			Per 1,000 Acres Crops and Pasture (1952)				Town and Village Population as Percentage of Total (1951)	Per 1,000 Population (1951)		Retail Sales £ per Person (1951)	Single Males as % of all Males Aged 30–34 (1951)
		£2–£10	£10–£50	Above £50	Milch Cows	Other Cattle	Pigs	Sheep		Private Cars Registered (1952)	Radio Licences (1952)		
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A	Carlow	31·8	46·9	21·3	67	252	56	531	40·2	43	169	66	60·3
B	Cavan	40·1	56·0	3·9	99	231	97	56	17·3	26	56	49	73·4
C	Clare	38·8	54·4	6·8	110	285	32	116	24·4	22	67	28	68·3
D	Cork	33·2	50·4	16·4	146	256	137	148	52·6	38	130	66	60·1
E	Donegal	69·8	25·9	4·2	102	248	22	463	18·9	21	80	45	62·4
F	Dublin*	41·2	33·4	25·4	108	268	110	236	94·8	49	185	117	40·8
G	Galway	45·7	50·9	3·4	69	239	44	801	28·1	22	87	40	69·1
H	Kerry	51·4	45·1	3·5	194	283	84	354	26·7	20	76	41	68·2
I	Kildare	34·0	41·5	24·5	52	290	28	184	29·2	40	123	54	53·5
J	Kilkenny	25·0	50·6	24·4	91	283	63	157	31·1	41	82	45	64·8
K	Laoighis	32·7	51·7	15·6	69	269	54	87	26·7	38	121	46	67·2
L	Leitrim	60·2	38·4	1·4	102	231	37	84	13·7	20	70	29	73·4
M	Limerick	33·3	47·5	19·2	181	277	68	36	48·2	32	158	53	55·2
N	Longford	40·4	51·8	7·8	74	290	49	75	21·3	32	111	44	68·6
O	Louth	36·0	48·2	15·8	69	285	55	204	63·0	37	200	78	51·0
P	Mayo	68·0	30·8	1·2	97	289	50	393	18·5	17	84	37	67·4
Q	Meath	32·0	48·8	19·2	55	351	23	252	17·5	49	116	53	62·1
R	Monaghan	31·8	61·9	6·3	85	235	101	39	24·6	32	80	70	69·7
S	Offaly	31·2	55·7	13·1	55	262	50	112	35·6	33	110	55	65·2
T	Roscommon	44·6	51·7	3·7	66	275	24	299	13·2	22	115	28	74·9
U	Sligo	48·9	48·1	3·0	92	266	30	205	29·7	24	102	42	67·0
V	Tipperary	28·3	52·1	19·6	107	312	52	140	36·5	41	127	56	62·8
W	Waterford	34·3	39·1	26·6	122	292	96	199	56·4	41	164	74	54·6
X	Westmeath	28·4	54·4	17·2	43	323	25	188	35·8	37	157	57	56·9
Y	Wexford	27·0	52·1	20·8	64	219	68	288	34·6	34	122	66	56·4
Z	Wicklow	34·7	46·3	19·0	79	212	44	528	49·8	36	102	65	50·4
Value of contiguity c		0·4193	0·8828	0·6160	0·3415	0·7876	0·6533	0·8686	0·6148	0·5185	0·8141	0·5267	0·6465
Significance $R = (1 - c)/0{\cdot}1512$		3·84	0·78	2·54	4·36	1·40	2·29	0·87	2·55	3·18	1·23	3·13	2·35

* Excluded from calculation of contiguity—see text.

values of the variance are not important. It is immaterial which is used.

When the quolls (the five quadratic orthogonal latitude-longitude terms—see Appendix) have been removed by regression, the residuals have a normal theory variance of 0·027291 or a standard error of 0·1652. As formula (3.21) shows, in this computation due allowance is made for degrees of freedom involved in the quolls (i.e. $q = 5$) while no such allowance has been made in the "normal" randomization value of 0·1550 already quoted. This comparison goes far to justify the use of the randomization method with the residuals without making allowance for degrees of freedom.

It will be remembered that the constants in c have been so determined that its mean value is unity. The 25-county values of k_1 and k_2 are respectively 4·4 and 21·76.

Cows. Though for the original data $c = 0{\cdot}3415$ with R having the highly significant value of 4·36, after removal of the quolls the value of c for the residuals is 1·03145, actually greater than unity though not significantly so since the randomization standard error of the residuals is 0·1724. The distribution of cows in Ireland at the county level is very largely due to geography alone! See analysis of variance in the Appendix.

Pigs. After removal of the quolls the value of c is 0·8504 with a standard error of 0·1570. While the difference from unity is not significant, R is less than unity, and this fact, coupled with the fact that the original significance 2·29 was not very emphatic, does not permit us to asseverate with the same confidence as in the case of cows that the distribution is due to location.

Town and Village Population. The original ratio, as shown in the table, gives a significance of 2·55. This value is appreciably affected by the fact that Co. Meath, with some of the best land in Ireland, has a very low town and village population. While this is partly due to propinquity to Dublin, it is also influenced by the fact that the large border town of Drogheda is administratively assigned to Co. Louth which has partly in consequence a very high town population ratio. When the county borders are "redrawn" so that Drogheda is assigned to Co. Meath, the significance becomes 2·83. When the quoll terms are removed the value of c is 0·9375, with significance of 0·38 so that the geographical situation goes far towards explaining the distribution of the town and village population of Ireland. Incidentally the residuals after removal of the linear terms of latitude and longitude give a value of c of 0·8750 with a significance of 0·72.

Private Cars. While the contiguity is highly significant, this is due entirely to motor cars on farms: as series (3) of Table 2 has shown, the larger farms have the property of contiguity in high degree. When attention is confined to non-agricultural motor cars (per 1,000 non-agricultural population), the value of c becomes 0·9648 which is not significant. It is curious that, for county units, there is no significant correlation between estimated numbers of non-agricultural superior personnel and non-agricultural motor cars per 1,000 population.

Retail Sales are highly correlated at the county level with town and village percentage (series (8)) as would be expected: $r = 0{\cdot}78$. What is very strange is that this correlation does not explain the contiguity of retail sales for when town and village effect is removed by regression the residuals have a value of c of 0·5325 (nearly equal to the original value) with a (randomization) significance R of 2·71. It may be observed that the "one-zero" example given in an earlier section of the paper was based on the retail sale distribution by assigning 1 to all counties with sales above the general (simple) average and 0 to the remaining counties. It will be seen that this simplification of the pattern has the effect of reducing R from 3·13 to 2·54. On the other hand if a five-grade classification be used and the numbers 1, 2, 3, 4, 5 assigned to the different grades, the apparent significance is greatly increased: in fact $R = 4{\cdot}04$. This is probably due to the exaggeration of the differences between the counties in the arbitrary numerical system used, thus increasing the general variance more than the contiguous variance.

Single Males show a significance R of 2·35. As was well known from the Census analysis into town and rural areas, this percentage is highly correlated ($r = -0{\cdot}80$) with the town and village percentage. Different from retail sales, the contiguity is largely explained by the town and village contiguity. In fact, when the latter is removed, $c = 0{\cdot}8938$ and $R = 0{\cdot}69$.

The only data, other than Irish county data, which the author has examined for contiguity are the death rates from tuberculosis (all forms) in the 22 Registrars' Districts in County Wexford, in the 21 years from 1906 to 1926 inclusive. This data has little current interest since the death rates have fallen considerably since the period in question and conclusions valid then may not hold now. In a paper of many years ago (Geary, 1930) it was shown that in the 8 Registrars' Districts with marl sub-soil there was a strongly linear relationship between the TB rate and marl area as percentage of total area.

The contiguity was not very significant in the original rates:

$c = 0{\cdot}8314$, $R = 1{\cdot}09$. Even after allowing for the effects of the marl percentage and the town and village percentage, the contiguity ratio for the residuals was not perceptibly improved for then $c = 0{\cdot}8579$, $R = 1{\cdot}02$. This is another example of the fact that independent variables, even if highly correlated with the dependent variables, do not necessarily "explain" the contiguity of the latter.

APPENDIX

Method of Orthogonalizing Independent Variables in Regression Analysis

The practical convenience of having independent variables orthogonal is obvious. For if the dependent variable is z_t $(t = 1, 2, \ldots, n)$ and the orthogonalized independent variables ξ_{it} $(i = 1, 2, \ldots, q)$, i.e. so that $\Sigma_t\xi_{it} = 0$, $\Sigma_t\xi_{it}^2 = n$, $\Sigma_t\xi_{it}\xi_{it'} = 0$, $t \neq t'$, the regression coefficients are found simply as

$$(1) \qquad b_i = \Sigma_t\xi_{it}z_t/\Sigma_t\xi_{it}^2$$

Furthermore, using orthogonal terms the total variance of the dependent variable z can readily be analysed to show the contribution of each term, and its statistical significance assessed from the residual variance.

C. R. Rao (1952) has suggested a method of orthogonalizing the original variables x_{it} which consists in taking new variables x'_{it} as follows:

$$\begin{aligned} x'_{it} &= x_{it} \\ x'_{2t} &= x_{2t} - a_{21}x'_{2t} \\ x'_{3t} &= x_{3t} - a_{32}x'_{2t} - a_{31}x'_{1t} \end{aligned}$$

The constants a_{ij} are chosen in successive stages so that $\Sigma_t x'_{1t}x'_{2t} = 0$, $\Sigma_t x'_{2t}x'_{3t} = 0$, etc. This method is very easy to apply in practice—far easier than the latent root method which will presently be described—but it supplies $q!$ different orthogonal transformations depending on the order of the variables, facing the computer with the problem of choice. Perhaps the logical order would be according to the correlation between the dependent and the several original independent variables, or the average of these for each independent variable when one is dealing with more than one dependent variable.

While the original independent variables can be transformed into orthogonal variables in an infinity of ways, it has seemed to the writer that the most logical method, and that which in certain cases imparts an objective meaning to the transformed variables, consists in determining the principal components of the original independent variables, having first standardized them, i.e. converting them individually so that each variance is unity. The procedure has the merit that it is symmetrical: it does not imply any particular order in the original variables and gives equal weight to all of them. As is well known the process consists in finding the values of the coefficients a_i which maximize

(3) $$2w = \Sigma_t(\Sigma_i a_i x_{it})^2$$

subject to

(4) $$\Sigma_i a_i^2 = 1$$

Introducing the Lagrange multiplier $n\lambda$, the equations to determine the a_i are as follows:

(5) $$\Sigma_j a_j m_{ij} = \lambda a_i,\ i = 1, 2, \ldots, q,$$

where $$nm_{ij} = \Sigma_t x_{it} x_{jt}.$$

The values of λ are found from (5) as the roots of the determinantal equation

(6) $$|m_{ij} - \lambda\delta_{ij}| = 0,\ \delta_{ij} = 0 \text{ if } i \neq j,\ = 1 \text{ if } i = j$$

In the kind of applications contemplated we shall not have to trouble about (6) having multiple roots. The orthogonal transformation required is

(7) $$\xi'_{it} = \Sigma_j a_{ij} x_{jt}$$

where a_{ij} ($j = 1, 2, \ldots, q$) are the solutions of (5) in a_i corresponding to the root λ_i of (6), arranged in descending order of magnitude. The ξ'_{it} are mutually orthogonal. These are then standardized by multiplication by $1/\sqrt{\lambda_i}$ to give the ξ_{it} with $\Sigma_t \xi^2_{it} = n$ and, of course, $\Sigma_t \xi_{it} = 0$. When $q = 2$ the transformation is

(8) $$\xi_{it} = (x_{it} + x_{2t})/\sqrt{2}$$
$$\xi_{2t} = (x_{it} - x_{2t})/\sqrt{2}$$

If x_{it} and x_{2t} are both positively or both negatively correlated to the dependent variable z, we can envisage the greater part of the variance being taken up by the ξ_{1t} term in the regression, the ξ_{2t} playing a subsidiary role; and inversely if the correlations of the x_{1t} and x_{2t} with y are of different signs.

This procedure of analysing the independent variables into dependent components is a purely algebraic (or even arithmetic) one: it has no stochastic implications whatsoever. The stochastic element, in regression theory, enters via the dependent variable. As a more general point it may be recalled that *any* non-singular linear transformation results in the remainders for *each* dependent element which are identical with those which would have been obtained from regression on the original independent variables. Furthermore the regression coefficients are consistent with the transformation in the sense that the two series of coefficients (i.e. on the original and, e.g. the orthogonal transforms) change into one another through the transformation as if the dependent variables were *exact* linear functions of the independent variables.

TABLE 3

Latitude-Longitude—25 Irish Counties. Quadratic Orthogonal Transforms Standardized—the "Quolls"

Serial Letter	County	Latitude °	Longitude °	ξ_{1t}	ξ_{2t}	ξ_{3t}	ξ_{4t}	ξ_{5t}
A	Carlow	52·70	6·80	− 0·575	0·896	− 0·131	− 0·538	− 1·257
B	Cavan	53·97	7·30	− 0·624	− 0·360	− 0·464	0·888	0·799
C	Clare	52·87	8·97	1·083	0·482	0·978	− 0·180	0·937
D	Cork	51·97	8·73	2·020	− 0·323	− 1·810	− 0·860	− 0·081
E	Donegal	54·90	7·92	− 0·408	− 3·739	− 1·148	0·081	− 1·192
G	Galway	53·35	8·75	0·226	0·192	1·308	− 0·402	1·059
H	Kerry	52·13	9·58	3·644	0·127	0·066	1·100	− 0·348
I	Kildare	53·20	6·78	− 0·609	0·946	0·044	0·591	− 0·319
J	Kilkenny	52·55	7·22	− 0·443	0·612	− 0·660	− 1·179	− 0·544
K	Laoighis	53·00	7·32	− 0·616	0·749	− 0·269	− 0·426	0·523
L	Leitrim	54·12	8·00	− 0·646	− 1·060	0·107	− 0·079	0·744
M	Limerick	52·50	8·75	1·207	0·359	− 0·223	− 0·534	0·854
N	Longford	53·73	7·71	− 0·688	− 0·090	− 0·027	0·059	1·212
O	Louth	53·93	6·53	− 0·159	0·217	− 0·391	2·559	− 0·373
P	Mayo	53·95	9·33	0·483	− 0·922	3·281	− 0·466	− 1·002
Q	Meath	53·63	6·67	− 0·447	0·626	− 0·103	1·633	− 0·142
R	Monaghan	54·15	6·93	− 0·356	− 0·558	− 0·811	1·869	0·258
S	Offaly	53·20	7·58	− 0·609	0·577	− 0·146	− 0·348	1·097
T	Roscommon	53·73	8·27	− 0·466	− 0·284	0·677	− 0·391	1·134
U	Sligo	54·17	8·67	− 0·345	− 1·420	1·428	− 0·642	− 0·104
V	Tipperary	52·62	7·88	− 0·068	0·490	− 0·682	− 1·131	0·714
W	Waterford	52·22	7·58	0·040	0·120	− 1·423	− 1·819	− 0·646
X	Westmeath	53·55	7·47	− 0·702	0·289	− 0·136	0·170	1·112
Y	Wexford	52·50	6·60	− 0·503	0·828	− 0·027	− 0·789	− 2·434
Z	Wicklow	52·98	6·35	− 0·439	1·246	0·562	0·834	− 1·999
	Sum	—	—	0	0	0	0	0
	Sum Squares	—	—	25	25	25	25	25

In order to determine the "quolls" (quadratic orthogonal components of latitude and longitude) for the 25 counties of the Irish Republic (i.e. excluding Dublin) the latitude and longitude of the centre of each county was assessed by inspection from a large map. It was not considered necessary for the present purpose to have exact Ordnance Survey readings. For each county the three products $x_{it}x_{jt}$ $(i, j = 1, 2;\ t = 1, 2, \ldots, 25)$ were computed; these also were standardized. The latter three series (x_{3t}, x_{4t} and x_{5t}) together with the standardized x_{it} and x_{2t} constituted the five independent variables. The latent root equation (6) was

(8)

$$\begin{vmatrix} 1-\lambda & -0{\cdot}102478 & 0{\cdot}415695 & -0{\cdot}258652 & 0{\cdot}289645 \\ -0{\cdot}102478 & 1-\lambda & -0{\cdot}252129 & 0{\cdot}365408 & 0{\cdot}086717 \\ 0{\cdot}415695 & -0{\cdot}252129 & 1-\lambda & -0{\cdot}326526 & 0{\cdot}049330 \\ -0{\cdot}258652 & 0{\cdot}365408 & -0{\cdot}326526 & 1-\lambda & -0{\cdot}220296 \\ 0{\cdot}289645 & 0{\cdot}086717 & 0{\cdot}049330 & -0{\cdot}220296 & 1-\lambda \end{vmatrix} = 0$$

or

$$(9)\quad \lambda^5 - 5\lambda^4 + 9{\cdot}303706\lambda^3 - 8{\cdot}113070\lambda^2 + 3{\cdot}335669\lambda - 0{\cdot}520925 = 0,$$

the roots of which are

$$\begin{aligned} \lambda_1 &= 1{\cdot}938867 \\ \lambda_2 &= 1{\cdot}167814 \\ \lambda_3 &= 0{\cdot}849644 \\ \lambda_4 &= 0{\cdot}561003 \\ \lambda_5 &= 0{\cdot}482672 \end{aligned} \tag{10}$$

It may be useful to place the transformed variables on record for students who may wish to compute regressions and remainders for the series in Table 2 or other Irish county series of which there are many. It should be pointed out that, while, in principle, the quolls have the properties $\Sigma\xi_{it}\xi_{jt} = 25\delta_{ij}$ ($\delta_{ii} = 1$, $\delta_{ij} = 0$, $i \neq j$), the actual totals are deemed correct only to the second decimal place.

Regression on the quolls have been computed for two series only for this paper, number milch cows and pigs per 1,000 acres of crops and pasture—series (4) and (6) of Table 2. The five term regression on percentage of population in towns and villages were worked from the original variables. The regressions were as follows:

Milch Cows

$$z_t = 91{\cdot}60 + 30{\cdot}705\xi_{1t} - 7{\cdot}226\xi_{2t} - 6{\cdot}331\xi_{3t} - 6{\cdot}003\xi_{4t} + 0{\cdot}346\xi_{5t} + u_t$$

Pigs

$$z_t = 55{\cdot}56 + 12{\cdot}740\xi_{1t} + 2{\cdot}729\xi_{2t} - 13{\cdot}644\xi_{3t} - 2{\cdot}263\xi_{4t} - 2{\cdot}894\xi_{5t} + u_t$$

The analysis of variance in each case is shown in the following table:

Term	Degrees of Freedom	Milch Cows		Pigs	
		Sum Squares	Ratio	Sum Squares	Ratio
1	1	23,570	62·79	4,058	7·14
2	1	1,305	3·48	186	0·33
3	1	1,002	2·67	4,654	8·18
4	1	901	2·40	128	0·23
5	1	3	0·01	209	0·37
u—total	19	7,133	—	10,805	—
u—*mean*	—	*375·4*	—	*568·7*	—
z	24	33,914		20,040	

With 1 and 19 degrees of freedom the 5 per cent and 1 per cent points of the ratio are respectively 4·38 and 8·18 so that only the ξ_1

component in the case of cows and the ξ_1 and ξ_3 components in the case of pigs would be adjudged significant.

In the latitude-longitude case exemplified above the five components appear to have but little objective significance: the method is merely a computational device. It would probably be otherwise if the original series were economic variables. It is intended, in Ireland, to analyse into orthogonal components an extended series of economic variables available for (*a*) counties (including contiguous county boroughs) and (*b*) rural districts (including contiguous urban districts). Less elaborate research shows that there is a large degree of consistency between different economic statistics throughout Irish counties so that the likelihood is that a few of the latent roots will be so much greater than the rest that the rest can be ignored. These components will then be available as independent variables to workers in many fields, e.g. for market research.

DISCUSSION

MR. H. W. GEARING said that the Conference had been very greatly honoured by Dr. Geary presenting to it some of the results of his original work on this subject. In the limited time since the text of the paper was circulated he had been unable to find any reference to the Contiguity Ratio in text-books, but more serious students of this topic would already be acquainted with the similar problem covered in Mr. Moran's Paper of 1948 concerning attributes of adjoining areas. On a different plain a recent book by Mr. T. Cauter and Mr. J. S. Downham entitled *Communication of Ideas* would also be relevant to this subject.

Mr. Gearing expressed the view that this technique would be used primarily as a complement to other methods of studying this kind of information, e.g.:

(*a*) The geographical method where one started with a map showing physical features and urban concentrations and added shading to help understand the problem, or

(*b*) The analysis of variance and the critical division of variability in samples between the variance arising from errors of sampling and the variance between strata of the sample. He would like to see some account given of the variability within each county before considering contiguity between counties and his first impression was that only if the variability within counties is a small proportion of the total variability would we be justified in pursuing this kind of analysis.

Mr. Gearing also asked what arrangements could be made in this method to take account of contagion which arose from facilities for communication between areas which were geographically removed from each other, e.g. tropical diseases could be carried by aircraft and our ordinary domestic diseases would tend to spread as the result of social contacts which need not necessarily be made between geographical neighbours.

Geographers would no doubt continue to prefer the map approach to such problems, but it was the job of the statistician to condense data into concise forms and the Conference was very greatly indebted to Dr. Geary for a very stimulating and original contribution.

DR. WISHART said that Dr. Geary had provided us with a thoughtful and learned paper which made an important contribution to the statistical mapping problem by defining and illustrating a measure of contiguity, the random sampling distribution of which he had to a considerable extent worked out. Dr. Geary had been bold enough to treat the Association as a very learned society, and in that respect had helped to make history. Contributors to the discussion would have difficulty in keeping on the author's high plane of achievement, one reason being that there had not been much time to study the circulated version of the paper. On the spur of the moment he had only two aspects of the paper on which to comment. The actual point at which reference should be made to Moran's paper of 1948 was not indicated although Dr. Geary said that his contiguity ratio was an attempt to do for variable measures what Moran did for attributes. Moran dealt, as a matter of fact, with a two-dimensional problem which had defeated Dr. Hartley and himself following on their one-dimensional attack in 1936 (*J. Lond. Math. Soc.*, **11**, p. 227) on the problem of the distribution of the joins between line segments of two kinds (black and white). Since then an extensive literature had emerged on this problem of what was generally known as the theory of runs, but a solution in two dimensions, let alone three, which is an outstanding problem of the nuclear physicist, was still awaited in full generality. Dr. Geary may have known of the interesting contribution made to the problem in 1950 by Dr. R. C. Bose (*Sankhya*, **10**, p. 13), but the physicist was still waiting for the actual probabilities associated with a given pattern of black and white patches.

His other comment concerned the interesting suggestion in the Appendix relating to the orthogonalizing of the independent variables in a regression analysis, with a reference to Rao's method. He had for long been a believer in this method of working out multiple regressions, chiefly to provide a very easy computational technique to the non-mathematician, and had had some success in teaching it to a class of agriculturists (see Communication No. 15—*Field Trials II—The Analysis of Covariance*—of the Commonwealth Bureau of Plant Breeding and Genetics, 1950). He welcomed the further publicity which Dr. Geary had given to the method, and commended the ingenuity with which he dealt with the vexed question of the order of choice of variable by bringing in principal component analysis.

MR. G. PRYS WILLIAMS thanked the speaker for propounding a principle which might prove very valuable in tackling problems arising from filtration and extrusion through multiple orifices. It appeared to him to offer a means of exploring a field which had proved difficult to approach with existing techniques and concepts.

MR. I. F. HENDRY felt that the contiguity ratio migh tbe very much more efficient than grouping in two dimensions. It gave us a different sort of information, information which was most important. There were several obvious industrial applications of this technique, e.g. the examination of materials such as paper, plastics and textiles which are made in a continuous web. When any of these fails under load, e.g. a paper bag bursts, a textile splits, or a piece of plastic gives way, it is important to know whether the surrounding areas have contributed to the defect of that part which failed.

There are many weak spot theories which have been applied on yarn and thin strips of metal to show the effect of length of strip on the strength. One of the most important tests on paper is the burst test and he felt that it would be an advance if we could work out a two-dimensional weak spot theory to show the effect of contiguous area on the impact strength of the paper at any one point.

He believed that this theory could be used for the extruded plastic industry and for the paper industry where the raw material is forced out of a narrow but long orifice and one would like to know whether irregularities are due to corresponding faults in contiguous areas.

MR. H. PALCA referred to Mr. Hendry's point about the relative efficiencies of the grouped and contiguity analyses. He wondered whether the former was unfairly helped by the presupposition of a Poisson Distribution: would not a Negative Binomial Distribution have been an equally valid hypothesis?

MR. A. MUIR referred to the fact that some people would be sceptical about using an estimate based on only a few observations. He pointed out that when sampling from a rectangular distribution the best estimate of the mean was given by the mid-range and not the arithmetic mean. Although it appeared that only two observations were used, i.e. the smallest and the largest, in fact, all the observations were used to decide which were the smallest and largest.

DR. GEARY thanked the speakers for their reception of his paper. He esteemed it a great honour to be invited to address the Conference.

He agreed with Mr. Gearing that variability within counties should be analysed by the contiguity ratio or, more simply, by analysis of variance. Certainly smaller units should be used when studying contagion. County boundaries were only accidentally (if at all) economic or social boundaries. In relation to a particular problem the boundary should be drawn so as to maximize the ratio of mean variance between counties to variance within counties. The question of contagion between non-adjoíning areas could easily be written into the formula for the contiguity ratio.

Following a train of thought prompted by Dr. Wishart's interesting observations, Dr. Geary emphasized that, as stated in the paper, the contiguity ratio was a rather obvious generalization of the von Neumann ratio. He was himself more interested in the extension of the idea to regression remainders, to consideration of the efficiency of the ratio in the "linear contagion" case (Section 4), and to the specific applications to Irish data. As to the orthogonalizing of the independent variables by the use of the latent root equation on which Dr. Wishart was so good as to comment, in honesty the speaker must confess himself a bit worried about one aspect, namely the validity of implicitly assigning equal weights to each of the original variables. If, as we intend to do in Ireland, we wish to orthogonalize say seven economic statistics given for counties, are we justified in attributing equal importance to an obviously prime indicator like retail sales per head as to a "secondary," say motor cars registered per head, if we wish to accord objective significance to the principal component, i.e. that associated with the largest latent root?

The lecturer was very interested in the suggestions of Mr. Prys Williams and other speakers about applying the contiguity ratio to industrial problems.

On the observations of Mr. Muir and other speakers about efficiency of estimates which generally arose out of Section 4 of the paper, it might be that the reason why the text which *seemed* to use more information (i.e. the contiguity ratio) was less sensitive for larger groupings than the so-called Method II was because the basic distribution was rectangular. Queer things happen with rectangular distributions. In the fairly well-known "Cars in a Town" problem (given that the cars are numbered consecutively, to estimate the number of cars in the town from the recorded numbers on a given sample of cars, assuming that the cars pass in random order), there are, of course, an infinity of solutions. The lowest variance solution is a constant times the largest observation, though this does not *seem* to use as much information as say, the average. In actual fact the S.D. of the largest value solution is $\mathbf{2}(1/\delta)$ whereas the S.D. of the solution based on the average is $\mathbf{2}(1/\delta\tfrac{1}{2})$.

REFERENCES

GEARY, R. C. (1930). The mortality from tuberculosis in Saorstát Éireann—A statistical study: *Journal of the Statistical and Social Inquiry*

Society of Ireland, 83rd Session, 1929–30.

Geary, R. C. (1933). A general expression for the moments of certain symmetrical functions of normal samples: *Biometrika*, **25**, 184.

Moran, P. A. P. (1948). The interpretation of statistical maps: *Journal of the Royal Statistical Society*, Series B, X, No. 2.

Neumann, J. Von (1941). Distribution of the ratio of the mean square successive difference to the variance: *The Annals of Mathematical Statistics*, XII, No. 4.

Rao, C. R. (1952). Advanced statistical methods in biometric research, 345.