The Economic and Social Review, Vol. 27, No. 2, January, 1996, pp. 145-159

Geary's Contiguity Ratio

ANTONY UNWIN
University of Augsburg

Abstract: Forty years ago Geary published a paper on spatial statistics introducing the contiguity ratio, c, to measure spatial pattern. He discussed c not only as a direct measure but as a regression diagnostic for assessing spatial association amongst regression residuals. c has continued to be used and this paper describes its properties, its current status and how work in this area has developed, emphasising particularly the complementary new approaches offered by interactive graphics tools.

I INTRODUCTION

Independence is a main tenet of much statistical theory but spatial data are very much not independent. Geary's interest in spatial statistics went back to an early study he carried out on TB rates in County Wexford (Geary, 1930). Identifying and assessing spatial patterns requires finding ways of modelling the spatial relationships. A first step is to construct some measure of the non-randomness of spatially distributed data. In his 1954 paper Geary considered "county" data (i.e., regional or area measures, not point data) and suggested the contiguity ratio, c, which is based on the squared differences between contiguous areas:

$$c = ((n-1)/2K_1)\Sigma'(x_t - x_{t'})^2 / \Sigma(x_t - \bar{x})^2 \qquad (1)$$

where n is the number of areas, x_t is the value for area t, $\bar{x}$ is the mean of all the values, k_t is the number of areas connected to area t and $K_1 = \Sigma k_t$ is twice the sum of all connections. Σ is the sum over all areas and Σ' is twice the

sum over all contiguous areas. Modern publications use the number of connections $A = K_1/2$ and a different summation notation, $\Sigma_{(2)}\partial_{ij}(x_i - x_j)^2$, for the numerator, but the formulae are exactly equivalent. (This is not apparent at first sight as ∂_{ij} is not the standard Kronecker delta but is defined as 1 if i and j are neighbours and 0 otherwise, with $\partial_{ii} = 0$. The summation $\Sigma_{(2)}$ counts each pair only once.)

The expected value of c is 1 under the null hypothesis of no spatial autocorrelation. This can be derived either by a randomisation argument or from classical normal theory. For the randomisation argument it is assumed that all n values may equally well have been allocated to the n areas in any one of n! possible ways. For the normality argument it is assumed that each value is an independent sample from the same normal distribution.

c is unusual for a measure of association in that a value of 1 suggests the values are distributed at random while values much less than unity or much more than unity suggest a pattern. If the data are positively spatially correlated then c will be small and close to 0 and if they are negatively spatially correlated then c will be bigger and close to 2. (It is surprising Geary did not suggest the statistic (1 – c) which would then be interpreted similarly to a standard correlation coefficient.) The maximum and minimum values which c can take are

$$\text{Max}(c) = \{(n-1)/2K_1\}m_{max} \text{ and } \text{Min}(c) = \{(n-1)/2K_1\}m_{min} \qquad (2)$$

where m_{max} and m_{min} are the largest and smallest eigenvalues of (I – C)W(I – C). W is the adjacency matrix ($w_{ij} = 1$ if areas i and j are contiguous and 0 otherwise) while each entry of C is 1/n (Haining, 1990, based on deJong *et al.*, 1984).

If the data are assumed to be a normal sample with no spatial structure then the variance of c will be

$$V(c) = \{K_1^2 + 2(K_1 + K_2)\}(n-1)/[(n+1)K_1^2] - 1 \qquad (3)$$

where $K_2 = \Sigma K_i^2$

and c is asymptotically normally distributed. The randomisation approach leads to a slightly different standard error. Geary describes the calculations but concentrates in his paper on the assumption of normality. He felt sure that c tended to normality fast, partly convinced by calculating the first four moments for the 26 Irish counties and finding that the skewness and kurtosis of c were close to those of a normal distribution. Simulations of small samples by Cliff and Ord (1981) support this view. A significance test is then carried

out by comparing $(1 - c)/\sqrt{V(c)}$ to the standard normal distribution.

c is a global statistic, i.e., it is calculated for the whole of the data set. If c is significant the problem of interpretation still remains. In practice, as Geary wrote, contiguity properties may be well-known "from ordinary mapping" so c would be used more to assess the non-randomness of a pattern already observed and to measure the relative strength of the contiguity. In a large data set a non-significant value of c for the whole area might mask local concentrations of values. c might be calculated for sub-regions of the whole country separately. Had Geary been an American dealing with large numbers of more widespread geographic units rather than an Irishman dealing with the 26 counties would he have extended his ideas to local spatial statistics?

One of the curious effects of Geary's paper has been the amount of interest shown in analysing data from the 26 counties of the Republic of Ireland. It is obvious why he should have been primarily concerned with such data but it is typical, if disappointing, that other statisticians should have been so as well. There are three main disadvantages in working with Irish county data: the small sample size, the exceptional nature of data for Dublin (Geary left Dublin out because of this, reducing his sample size further from 26 to 25), and the topology. Obviously any small data set will have a high proportion of border areas, but this is compounded here by the border with Northern Ireland. Thus Donegal in the North-West only has one neighbour in the data set, Leitrim, and their common boundary is very short. In contrast, Tipperary has eight neighbours. Were Donegal to have an outlying value this would enter into the calculation only twice, while were the same outlier in Tipperary it would enter 16 times. An artificial example illustrates this. Let the 25 counties excluding Dublin all have the value 0, barring one which has the value 25. Then $c = 0.227$ if that county is Donegal and $c = 1.818$ if it is Tipperary.

Geary does not discuss alternative definitions of contiguity and their possible effect on c but in one respect his use is unexpected. Map II on p. 9 of his article shows the 26 counties, each labelled with the number of connections to neighbours. From this it can be seen that in the South-West, Kerry and Clare are regarded as contiguous although they are separated by the Shannon and in the South-East Waterford and Wexford are regarded as contiguous although they also have no common land border. In practice this made little difference as these are only two out of 55 connections. The most significant value of c reported by Geary is for milch cows and one of the least significant is for sheep. The values without Kerry/Clare and Waterford/Wexford are:

	Milch cows	*Sheep*
All 25 counties (excluding Dublin)	0.3415	0.8686
Excluding Kerry/Clare	0.3016	0.8696
Excluding Kerry/Clare and Waterford/Wexford	0.2848	0.8838

The standard error increases from 0.1512 to 0.1562 to 0.1600, the values for milch cows being all highly significant and the values for sheep being not significant.

II COMPARING GEARY'S c AND MORAN'S I

A few years earlier, in 1948, Moran had introduced another measure, I, to measure spatial autocorrelation.

$$I = (n / K_1)\Sigma'(x_i - \bar{x})(x_j - \bar{x}) / \Sigma(x_i - \bar{x})^2 \qquad (4)$$

I can take both positive and negative values and is close to zero when there is no spatial autocorrelation. It is more like a correlation coefficient although its expected value under the null hypothesis of no spatial autocorrelation is not 0 but $-1/(n-1)$. This may be derived either by assuming normality or by a randomisation argument. The maximum and minimum values it can take are similar in form to those for c (see Haining (1990), after deJong, *et al*. (1984)).

Geary lists Moran's paper in his references but does not mention it, a curious omission since he only lists five papers in all, including two of his own. It would be interesting to know whether the two had discussed the problem together, given the small community of statisticians at that time. Moran was in Oxford from 1946 until 1952 when he took up a chair in Australia.

The two statistics are often referred to in tandem. I looks more like a standard correlation measure while c can be traced to other ideas from time series (Geary refers specifically to Von Neumann's ratio although Cliff and Ord point out the relationship to Durbin and Watson's d statistic as well). Cliff and Ord come down slightly in favour of I on the basis of its appearing to be more robust, of its tending to normality faster and of asymptotic relative efficiency (ARE). As ARE "gives a fair guide to their relative power for alternatives not too far from the null hypothesis" (Cliff and Ord, 1981, p. 165) this does not seem useful for measures we are mainly going to consider when the data are far away from no spatial autocorrelation.

Whether these arguments are of much practical influence is unclear, particularly as I and c measure slightly different aspects of spatial correlation. A more convincing reason for preferring I may be found in Anselin's work

(1993) in which he describes how I may be decomposed to show the local effects. He observes that I may be interpreted as measuring the degree of linear association between a vector of observed values y and a weighted average of the neighbouring values, Wy. (c may be decomposed too but does not have this interpretation.) He recommends using a scatterplot of Wy against y with a linear regression of slope I. The usual regression diagnostics may then be used to identify outliers and investigate other aspects of the fit and hence to interpret I and deviations from no correlation.

Both the I and c statistics were introduced for regions and use continguity to define adjacency. This means that natural barriers are ignored as are length of common boundary and other geographical features. In geostatistics, when the data locations are points, adjacency measures use Euclidean distance. O'Loughlin *et al.* (1994) used a mixture of the two in an effort to cope with the variety of scale differences in their analysis of the 1930 German elections because of the mixture of urban and rural voting areas. Results were calculated using both contiguity and a distance band of 56 kilometres, which was chosen to ensure that no area was unconnected. The definition used will affect the results although the same kind of tests based on the normal distribution may be used. Cliff and Ord give the theory for general weighting matrices $\{w_{ij}\}$ (which need not be adjacency matrices at all). A further generalisation is based on work by Hubert *et al.* (1981) who show that the weighted versions of c and I are both special cases of the general cross product statistic after some suitable normalisation

$$\Gamma = \Sigma_i \Sigma_j G_{ij} C_{ij} \tag{5}$$

where G_{ij} is a measure relating locations i and j and C_{ij} is a function of the variable values. For c, $G_{ij} = w_{ij}$ and $C_{ij} = (x_i - x_j)^2$ while for I, $C_{ij} = (x_i - \bar{x})(x_j - \bar{x})$. Results will also be affected by the areal framework used in that different frameworks will give different sets of values and generate different statistics. This is a classic problem in quantitative geography known as the Modifiable Areal Unit problem. An interesting reference from this point of view is Openshaw and Taylor (1981).

Missing values are a further problem as they are for all statistical methods. In the case of c it would be valuable to look at the influence they could have. Geary left out Dublin although values were available and nowadays one would calculate statistics both with and without outliers. On the other hand, Anselin analysed African data on strife but without having any value for Angola. Sensitivity analyses based on ranges of possible values should be easy to implement.

III LOCAL STATISTICS

The inappropriateness of using a global statistic to assess pattern across a large area is obvious, too much information has to be encapsulated in a single number. The work of Anselin referred to above is one way of responding to this, though curiously this is not the approach he adopted in a later application (O'Loughlin *et al.*, 1994). Pre-war Germany was divided into six regions and statistics on Nazi voting strength in the 1930 Reichstag election were calculated separately for each. Another recent proposal has been that of Getis and Ord (1992) who describe G statistics to assess local pattern, in particular to identify local concentrations of high and low values. G was originally defined for point patterns and is calculated for each unit individually based on all local points, where local is defined as within a specified distance. G has since been applied to areal data too. Its value depends, of course, on the definition of distance/adjacency used (as does its significance, see Ding *et al.*, 1992). The value for point i as in Getis and Ord is

$$G_i(d) = \sum w_{ij}(d)x_j / \sum x_j \qquad j \neq i \tag{6}$$

where $\{w_{ij}\}$ is a symmetric one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i. (The version for areal data obviously just uses an appropriate $\{w_{ij}\}$ matrix.) Note that the summation in the denominator does not include x_i so that it is different for each i. The expected value and variance of G_i are

$$E(G_i) = W_i / (n-1) \tag{7}$$

$$V(G_i) = \left[W_i(n-1-W_i) / \left\{(n-1)^2(n-2)\right\}\right](Y_{i2} / Y_{i1}^2) \tag{8}$$

where $W_i = \sum w_{ij}(d)$, $Y_{i1} = \sum x_j / (n-1)$ and $Y_{i2} = \sum x_j^2 / (n-1) - Y_{i1}^2$.

Although adjacent values of G are highly correlated, each $G_i(d)$ can be taken to have a normal distribution as $n \to \infty$, provided that d is not too small (so that there are no neighbours) or too large (so that all elements are neighbours). Getis and Ord suggest standardising the G_i's individually to Z_i and comparing with a standard normal distribution. A large positive Z_i implies many large values close to i and a large negative Z_i implies many small values close to i. The main disadvantage of G is that it is only sensible for a positive variable with a natural origin. Even when these two conditions are satisfied it is not clear how useful G is compared to straightforward use of interactive graphics. In principle G could be used to check whether patterns identified by visual inspection could be taken to be non-random. In practice G only checks for a limited form of pattern (for instance it would not identify

patterns along borders) and is univariate. More research is needed.

Anselin (1994) has proposed a class of local statistics, LISA (Local Indicators of Spatial Association), which, however, do not include the G statistics. These indicators are derived by decomposing global statistics (Anselin uses Moran's I primarily) into their local components.

IV SPATIAL AUTOCORRELATION AND INTERACTIVE GRAPHICS

It is common in geographical analyses to display variable values for regions using chloropleth maps. These may be shaded or coloured in a variety of ways in order to obtain better representations, but all versions suffer from their dependence on the classification scheme used to discretise the data and their static nature. Interactive tools which permit direct map interrogation and linking of the map to histograms, bar charts, scatterplots and other statistical displays enable a flexible exploration of the data both in its geographic and statistical contexts.

To illustrate, consider another Irish data set, albeit a more recent one than Geary's. The data are the Irish election results for the 41 constituencies which existed through the 1980s and early 1990s. For each constituency the data set includes the first preference percentage support for each of the major parties over the five elections of the 1980s, the first and second count percentages for the 1990 Presidential election which Mary Robinson won and the percentage "yes" votes in the four referenda during the period (Divorce, Single European Act, Right to Life, Maastricht). In the following figures the map has been drawn with the eleven Dublin constituencies (including Dun Laoghaire) magnified and placed in the Irish Sea to the East of their true position. The interactive graphics software used to analyse the data is REGARD (Unwin, 1994) which has been developed to extend the tools found in software such as Data Desk, JMP and SAS Insight to spatial data.

Not surprisingly, all of the political results show evidence of strong spatial pattern. The following table gives the values of c for a selection of them:

Fianna Fáil 1981	0.206
Fianna Fáil 1989	0.430
Mary Robinson First Preferences	0.214
Single European Act	0.249
Maastricht	0.294
Divorce	0.117
Right to Life	0.119

As the standard error under the assumption of normality is 0.125, all are highly significant. The most significant value is for the divorce referendum

and it is easy to see why from the REGARD screen in Figure 1. The group of constituencies which were most favourable to the introduction of divorce have been selected in the histogram and are automatically highlighted in the map. They were all in and around Dublin.

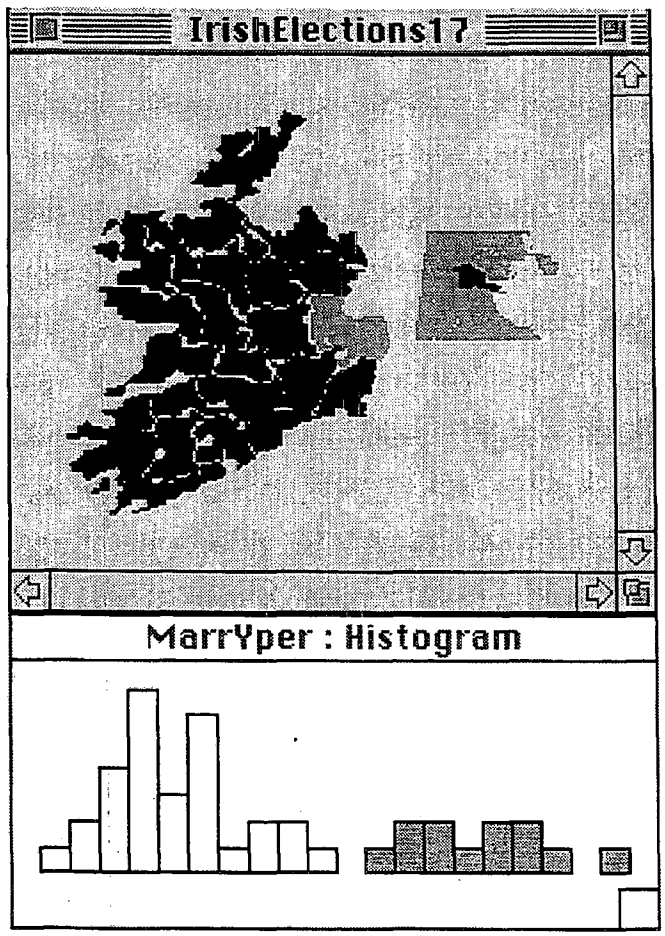

Figure 1: *Constituencies which had the Strongest Vote for Divorce in the Referendum of 1986*

The least significant of the values was for the Fianna Fáil first preferences in 1989. Figure 2 shows a similar kind of picture to Figure 1 but with rather less pattern and with the selection made from the map. It is important to realise that when using REGARD, selections may be changed directly by a mouse-click, either choosing areas on the map or parts of the histogram. Selections are automatically linked to any other open display. Thus Figures 1

and 2 are particular examples of many screens that were explored to understand better why c is significant for these data. While it only takes a few seconds to run through such interactive analyses, it would take a large number of printed pages to convey the same information and would not, of course, provide the same flexibility of exploration.

The automatic linking to other displays in REGARD encourages a multivariate approach because of the ease with which many variables may be considered simultaneously. (An example is shown in the next section.) Geary

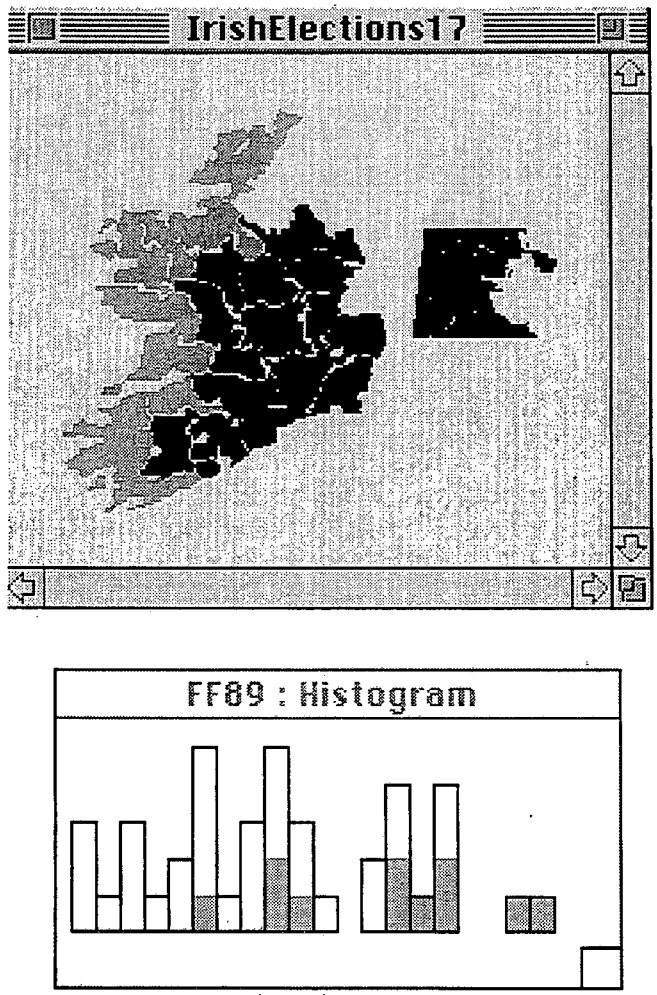

Figure 2: *Constituencies on the West Coast and First Preference Percentages for Fianna Fáil in 1989*

was well aware of the need to look at more than one variable in his paper but was limited by the lack of any computing power. Many later researchers, despite having considerable computing power at their disposal, have similarly, but unnecessarily, limited themselves. c is a univariate measure while all spatial data sets are multivariate.

Geary himself suggested one multivariate approach. This involves initially ignoring the spatial component and fitting a regression or other statistical model to the data. Then the model residuals are examined using c to see if a spatial component remains. As the results for the Divorce and Right to Life referenda were highly negatively correlated (r=–0.967), (because of the way the questions in the referenda were phrased), the residuals from a linear model of Divorce voting on Right to Life voting were examined. The value of c obtained was 0.194, still highly significant. Figure 3 shows the reason. In other cases that were examined c also proved a useful tool for assessing the spatial pattern in the residuals.

In an appendix to his paper Geary constructed "quolls" (quadratic orthogonal components of latitude and longitude) to use in regressions to remove geographic effects. This is not a convincing section as it is hard to interpret the resulting terms in any useful way. Geary himself noted this and explained that he planned to calculate orthogonal components of an extended series of economic variables for Ireland. Orthogonal components were much more popular in those days because of their computational properties. Geary comments that the method is very easy to apply in practice but is dependent on the order of the variables, "facing the computer with the problem of choice". In Ireland in the early 1950s a computer was obviously still a person.

V c AND DISPLAYS OF DIFFERENCES

The advantage of c lies in its interpretation as the sum of all squared adjacent differences. Differences have a much more natural local interpretation than products (which Moran's I uses) and this consideration is important in discussing and conveying results. For c the natural supporting graphical display to use is a version of the variogram cloud (although this does not seem to have been suggested before), in which squared differences are plotted against distance (Haslett *et al.*, 1991). Either plotting may be restricted to contiguous pairs or distance may be measured by the degree of neighbourliness (so that all contiguous pairs are at distance 1). The latter could equally well be achieved by a dotplot of the contiguous differences (or perhaps a histogram) ignoring the rest, but then the context of the other information on differences would be lost. The same applies to the version restricted to contiguous pairs in which it would be better to plot all points but

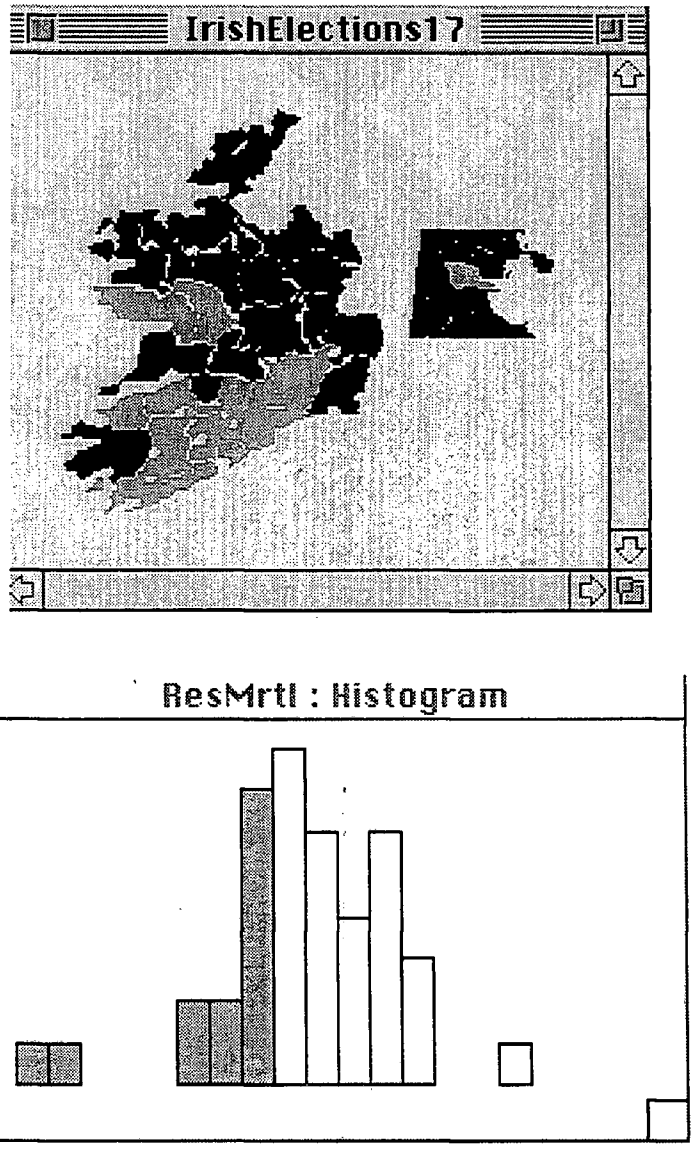

Figure 3: *The Largest Negative Residuals from a Linear Regression of the Results of the Divorce Referendum on those of the Right to Life Referendum*

draw the contiguous pairs with a different symbol for identification purposes.

The key element which makes the displays valuable is the ability to link the points in the variogram cloud to the map. Either the endpoints (or areas) associated with particular differences may be highlighted, or, to be more

specific, lines joining the adjacent pairs may be highlighted. Fully interactive graphical software is still not common but the necessary difference calculations and point highlighting can be accomplished with a little effort using the linking and relational facilities of the Macintosh software Data Desk (Velleman, 1992). No map can be shown but a scatterplot of latitude and longitude can be used to provide the linking to spatial position. Loading the differences into a line layer in REGARD allows linking with proper maps and also representation of areas as areas (stored in a regional layer in REGARD) and not as points. As is common in GIS systems, REGARD allows the overlaying of several layers of information, in this case a regional layer and a line layer have been used in tandem and are linked together. A line is drawn between the centres of each pair of contiguous areas and the values for the squared differences are associated with the line objects.

Figure 4 shows an example in which boxplots of the squared differences between adjacent constituencies for the four referenda and for the support for Mary Robinson in the first count of the Presidential election have been drawn. The current selection is of the highest squared differences in Mary Robinson's vote. The corresponding lines have been highlighted on the map and boxplots for them have been drawn on top of the boxplots for all 41 constituencies. Limerick East stands out as a constituency very different from the four around it. Anyone who recalls the rise of the Progressive Democrats will recognise the influence of their leader, Des O'Malley.

It is interesting to observe that the voting patterns for the referenda are not the same as for the Presidential election. Further analysis using linked scatterplots and other interactive tools of REGARD may be used to study the data in more depth. While it is clear that with these interactive graphical tools Geary's c can be understood better, it is also clear that there is a need for multivariate measures to complement multivariate graphical analyses. Extending Geary's c to a multivariate measure could be worthwhile.

VI CONCLUSION

Geary was obviously aware of the need for tools to analyse spatial data from his early work on TB. His 1954 paper introduces a simple global statistic and investigates its properties. This part of the paper has had a major effect because c is comprehensible and easy to understand. The rest of the paper has had less effect, as the methods he suggests for more sophisticated analyses of spatial data are not attractive. Here he was undoubtedly restricted by the level of computing power available to him. Subsequent researchers have developed many new methods for spatial analysis, mainly heavily dependent on modern computing power, though none has managed

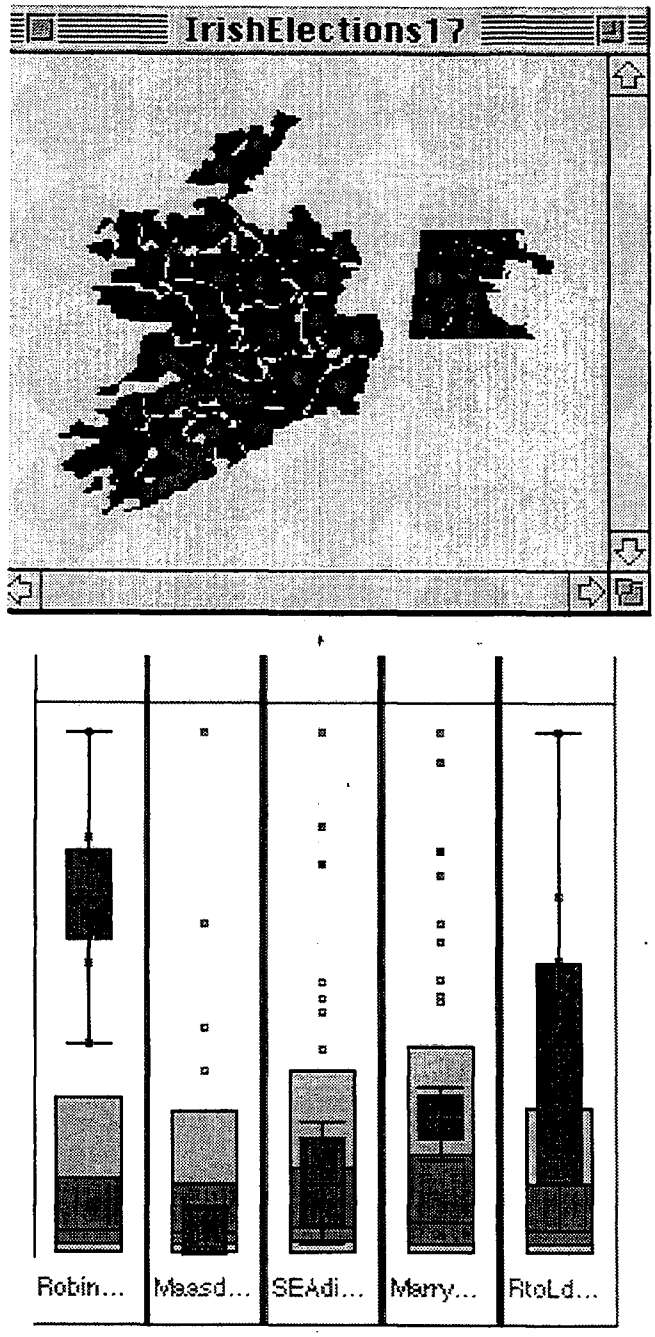

Figure 4: *Large Differences in Support for Mary Robinson in 1990 between Adjacent Constituencies. (The boxplots show from left to right the distributions of first preference percentages for Mary Robinson and the yes vote percentages in the four referenda: Maastricht, the Single European Act, Divorce, Right to Life.)*

to deal successfully in any general way with the wider range of problems within spatial data and most have unnecessarily limited themselves to univariate data. In this sense, Geary's c is as useful a global statistic as anything that has been developed subsequently. The main analytic advance of note has been the introduction of local statistical measures.

Interactive graphical tools, such as are available for spatial data in REGARD,[1] improve analysts' ability to understand and interpret spatial statistics enormously. Their emphasis on the multivariate nature of the data represents a major advance, but there is always the risk of unwarranted conclusions being drawn due to the seductive nature of attractive graphical displays. Statistical tools are essential for assessing results obtained visually and Geary's contiguity ratio, c, is a valuable tool for this purpose. With the increasing application of the new tools of interactive graphics, c may yet gain in importance.

REFERENCES

ANSELIN, L., 1993. *The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association — LISA*, Research Paper No. 9330, West Virginia University.

ANSELIN, L., 1994. *Local Indicators of Spatial Association — LISA*, Research Paper No. 9331, West Virginia University.

CLIFF, A.D., and J.K. ORD, 1981. *Spatial Processes*, London: Pion.

deJONG, P., C. SPRENGER, F. VAN VEEN, 1984. "On Extreme Values of Moran's I and Geary's c", *Geographical Analysis*, Vol. 16, pp. 17-24.

DING, Y., and A.S. FOTHERINGHAM, 1992. "The Integration of Spatial Analysis and GIS", *Computers, Environment and Urban Systems*, Vol. 16, No. 1, pp. 3-19.

GEARY, R.C., 1930. "The Mortality from Tuberculosis in Saorstát Éireann — A Statistical Study", *Journal of the Statistical and Social Inquiry Society of Ireland*, Vol. 83, pp. 67-103.

GEARY, R.C., 1954. "The Contiguity Ratio and Statistical Mapping", *The Incorporated Statistician*, Vol. 5, No. 3, pp. 115-145.

GETIS, A., J.K. ORD, 1992. "The Analysis of Spatial Association by Use of Distance Statistics", *Geographical Analysis*, Vol. 24, No. 3, pp. 189-206.

HAINING, R., 1990. *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge: Cambridge University Press.

HASLETT, J., R. BRADLEY, P. CRAIG, A.R. UNWIN, and G. WILLS, 1991. "Dynamic Graphics for Exploring Spatial Data, with Application to Locating Global and Local Anomalies", *American Statistician*, Vol. 45, No. 3, pp. 234-242.

HEARNSHAW, H.M., D.J. UNWIN (eds.), 1994. *Visualization in GIS*, Chichester: Wiley.

1 *Software*
REGARD was developed in Trinity College Dublin by Graham Wills, Antony Unwin and John Haslett as a tool for the exploratory analysis of spatial data. It runs only on Macintosh computers. Development of and support for REGARD is now based in Augsburg. Further details may be obtained from the author.

HUBERT, L.J., R.G. GOLLEDGE, C.M. COSTANZO, 1981. "Generalised Procedures for Evaluating Spatial Autocorrelation", *Geographical Analysis*, Vol. 13, pp. 224-233.

MORAN, P.A.P., 1948. "The Interpretation of Statistical Maps", *Journal of the Royal Statistical Society Bulletin*, Vol. 10, No. 2, pp. 243-251.

O'LOUGHLIN, J., C. FLINT, AND L. ANSELIN, 1994. "The Geography of the Nazi Vote", *Annals of the Association of American Geographers*, Vol. 84, No. 3, pp. 351-380.

OPENSHAW, S., P.J. TAYLOR, 1981. "The Modifiable Areal Unit Problem", in N. Wrigley and R.J. Bennett (eds.), *Quantitative Geography*, London: Routledge and Kegan Paul.

UNWIN, A.R., 1994. "REGARDing Geographic Data", in P. Dirschedl and R. Ostermann (eds.), *Computational Statistics*, pp. 315-326, Heidelberg: Physica.

VELLEMAN, P.F., 1992. *Data Desk*, Ithaca New York: Data Description.