5. Tests of Spatial Concentration

The above testing procedures are all motivated by the spatial autoregressive model of residual errors. So before moving on to *spatial regression analyses* of areal data, it is appropriate to consider certain alternative measures of spatial association that are also based on spatial weights matrices. By far the most important of these for our purposes are the so-called *G-statistics*, developed by Getis and Ord (1992,1995).[1] These statistics focus on direct associations among (nonnegative) spatial attributes rather than spatial residuals from some underlying explanatory model. For any given set of *nonnegative data,* $x = (x_1,..,x_n)'$, associated with n areal units, together with an appropriate spatial weights matrix, $W = (w_{ij} : i, j = 1,..,n)$, the G^* *statistic* for x is defined to be:[2]

$$(5.1) \qquad G_W^*(x) \;=\; = \frac{\sum_{i=1}^n \sum_{j=1}^n x_i w_{ij} x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i x_j} \;=\; \frac{x'Wx}{(1_n' x)^2}$$

As discussed further below, the diagonal elements of *W* are allowed to be nonzero (since no autoregressive-type relations are involved). However, if one is only interested in relations between distinct areal units, $i \neq j$, so that the diagonal elements of *W* are treated as zeros, then the resulting statistic is called simply the *G statistic*, and is given by:

$$(5.2) \qquad G_W(x) \;=\; \frac{\sum_{i=1}^n \sum_{j\neq 1}^n x_i w_{ij} x_j}{\sum_{i=1}^n \sum_{j\neq 1}^n x_i x_j} \;=\; \frac{x'W^0 x}{(1_n' x)^2 - x'x}$$

where $W^0 = W - diag(W)$. However, our focus will be almost entirely on G^* statistics.[3]

5.1 A Probabilistic Interpretation of *G**

While the definitions in (5.1) and (5.2) serve to clarify the formal similarities between these indices and those of the previous section, there is an alternative representation which suggests a more meaningful interpretation of these indices. Here we focus on G^*. First observe that since $x_i \geq 0$, if we let

$$(5.1.1) \qquad p_i \;=\; \frac{x_i}{\sum_{j=1}^n x_i} \;=\; \frac{x_i}{1_n' x}$$

[1] The 1992 paper is Reference 7 in the class Reference Materials.

[2] While our present focus is on areal units, it should be noted that these *G*-statistics are equally applicable to sets of point locations, such as hospitals or supermarkets within a given urban area.

[3] It should be clear from these definitions that a better choice of notation would have been to use *G* with *W* and G^0 with W^0. But at this point, it is best to stay with the standard notation in the literature.

denote the proportion (or fraction) of x in unit i, and let $p = (p_1,..,p_n)'$ denote the corresponding vector of proportions, then G^* can be rewritten as

(5.1.2) $$G_W^* = \frac{\sum_{ij} x_i w_{ij} x_j}{(1_n' x)^2} = \sum_{ij} \left(\frac{x_i}{1_n' x}\right) w_{ij} \left(\frac{x_j}{1_n' x}\right) = \sum_{ij} p_i p_j w_{ij}$$

Next observe (from the title of their 1992 paper) that Getis and Ord are primarily interested in *distance-based* measures of proximity or accessibility. In particular, if we let d_{ij} denote some appropriate notion of *distance* between units i and j, and let $a(d)$ denote an appropriate (nonincreasing) *accessibility function* of distance [such as $a(d) = d^{-\theta}$ or $a(d) = \exp(-\theta d)$], then we may now interpret each spatial weight as an *accessibility measure*

(5.1.3) $$w_{ij} = a(d_{ij}) \ , \ i, j = 1,..,n$$

and write

(5.1.4) $$G_a^* = \sum_{ij} (p_i p_j) a(d_{ij})$$

To give a concrete interpretation to G_a^*, let us assume for the moment that x_i represents the *population* in areal unit i, so that p_i is the fraction of population in i, and $p = (p_1,..,p_n)'$ is the population distribution among areal units. In this context one may ask: *What is the expected accessibility between two randomly sampled individuals from this distribution*? To answer this question, observe that since p_i is by definition the probability that a randomly sampled individual is from unit i, it follows by independence that $p_i p_j$ must be the joint probability that these two random samples are from units i and j, respectively. So if accessibility is treated as a random variable with values, $a(d_{ij})$, for each pair of areal units, then it follows from (5.1.4) that G_a^* must be the *expected value* of this random variable, i.e.,

(5.1.5) $$G_a^* = E(a)$$

Thus the value of G_a^* is precisely the answer to the question above, i.e., the *expected accessibility* between two randomly sampled individuals in this population.

In terms of this particular example, there are several additional features that should be noted. First it should be clear that two individuals in the same areal unit are by definition maximally accessible to one another. So any measure of overall accessibility will surely be distorted if these relations are omitted – as in *G* statistics. It is for this reason that our focus is almost exclusively on G^* statistics. Notice also from the definitions of a and p

that G_a^* must achieve its maximum value when all population is concentrated in the smallest of these *n* areal units. This suggests that G_a^* is more accurately described as a measure of *spatial concentration* than association.

More generally, these interpretations carry over to essentially any nonnegative data. For example, if x_i denotes income or crime levels, then G_a^* represents the spatial concentration of income or crime. But here one must be careful to distinguish between *extensive* and *intensive* quantities. For example, while proportion of total income (dollars) in areal unit i is straightforward, the "proportion" of per capita income is less clear. Hence one must treat such intensive quantities in terms of density units that can be added. So for example, if per capita income is twice as high in i as in j, this would here be taken to mean that the income density in i is twice that in j. So a better interpretation of G_a^* in this case would be in terms of the spatial concentration of income density. In any case it is certainly meaningful to ask whether certain spatial patterns of per capita income are more concentrated than others

Finally, we should add that even for spatial weights matrices, *W*, that are not distance based (such as spatial contiguity matrices), such weights can still be viewed as measures of "closeness" in an appropriate sense. So in the analyses to follow, we shall continue to interpret G_W^* in (5.1.2) as measuring the degree of spatial concentration of quantities, $x = (x_1,..,x_n)'$.

5.2 Global Tests of Spatial Concentration

To test whether population (income, crime, etc.) is "significantly concentrated" in space, it is natural to again consider permutation tests involving G_W^*, where w_{ij} is implicitly interpreted as a measure of accessibility, *a*, as in (5.1.3) above. The details of such a testing procedure are essentially identical to the *sac_perm test* above. The only difference is that the relevant test statistic, *S*, in Section 4.3.1 above is now G_W^* rather than say the Moran statistic, I_W. This procedure in operationalized in the MATLAB program, **g_perm.m**.

As one application of this testing procedure, we again consider the *English Mortality* data in Figure 1.9 above (p.III.1-5). For purposes of illustration, we here consider a new type of spatial weights matrices, namely *exponential-distance weights* [expression (2.1.13)] which is also constructed by using the MATLAB program, **dist_wts.m**. Starting with exponential-distance weights, say

(5.2.1) $$w_{ij} = a(d_{ij}) = \exp(-\theta\, d_{ij})$$

we first note that since the negative exponential function approaches zero very rapidly, it is often advisable to normalize distance data to the unit interval to avoid vanishingly small values.[4] To do so we first identify the largest possible centroid distance, $d_{\max}$, between all pairs of Health Districts, and then convert centroid distances, d_{ij}, to the unit interval by setting

$$(5.2.2) \qquad d_{ij}^{*} = d_{ij} / d_{\max} \quad , \quad i, j = 1,..,n\,(=199)$$

so that $0 \le d_{ij}^{*} \le 1$. Using this normalization, we can then design exponential distance weights to yield some appropriate "effective band width" by simply plotting the function $\exp(-\theta d), 0 \le d \le 1$, for various choices of θ. For our present purposes, the value $\theta = 10$ yields the plot shown in Figure 5.1 below, [5] which is seen to yield an effective bandwidth of about $d = 1/2$ (shown by the red arrow). In terms of our normalization in (5.2.2) this yields the familiar value, $d_{\max}/2$:

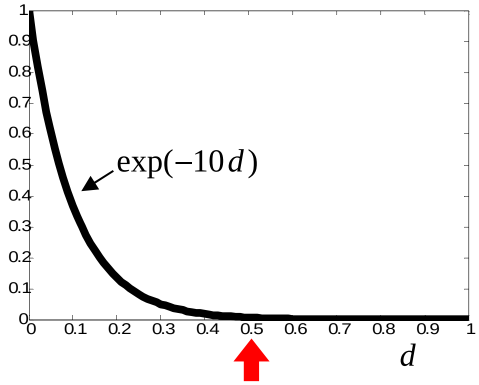


Figure 5.1. Negative Exponential Function

Using the workspace, **eng_mort.mat**, the corresponding spatial weights matrix, **W1**, is constructed by using **dist_wts.m** with the commands:

```
>> info.type = [4,10,1];
>> W1 = dist_wts(L,info);
```

Here **L** is the 199x2 matrix of centroid coordinates, '**4**' indicates that exponential-distance weights are option 4 in **dist_wts.m**, '**10**' denotes the exponent value, and (most importantly) '**1**' denotes the option to leave all diagonal elements as calculated [in this case, $\exp(0) = 1$]. Note also that since these weights are already guaranteed to lie in the unit interval (as in Figure 5.1), there is no need to consider any additional normalizations (as provided by the **info.norm** option). Finally, denoting the myocardial infarction rates

[4] For example, if distance were in meters, then while a distance of 800 meters is not very large, you will discover that MATLAB yields the negative exponential value, exp(-800) = 0. Moreover, this is not "rounded" to zero, but is actually so small a number that it is beyond the limits of double precision arithmetic to detect.

[5] This plot is obtained with the commands: **x = [0:.01:1]; y = exp(-10*x); plot(x,y,'k','Linewidth',5);**

by **z = mort(:,3),** the test of spatial concentration using **g_perm.m** is performed with the command:

```
>> g_perm(z,W1,999);
```

The results of this test (with **999** random permutations of Health Districts) is shown below:

```
SPATIAL CONCENTRATION RESULTS

INDEX     VALUE      PROB
G         0.0055     0.0010
G*        0.0054     0.0010
```

Notice first that both G and G^* values are reported, even though G^* is of primary interest for our purposes. Next observe that, not surprisingly, these myocardial infarction rates are maximally significant given 999 permutations, and that in this case there is very little disagreement between G and G^*.

For purposes of comparison, we also try the more local spatial weights matrix, **Wnn_5**, already employed in Section 4.3.2 above to test for spatial autocorrelation in the regression residuals for this same data. Here the results of using

```
>> g_perm(z,Wnn_5,999);
```

are seen to be practically the same:

```
SPATIAL CONCENTRATION RESULTS

INDEX     VALUE      PROB
G         0.0057     0.0010
G*        0.0056     0.0010
```

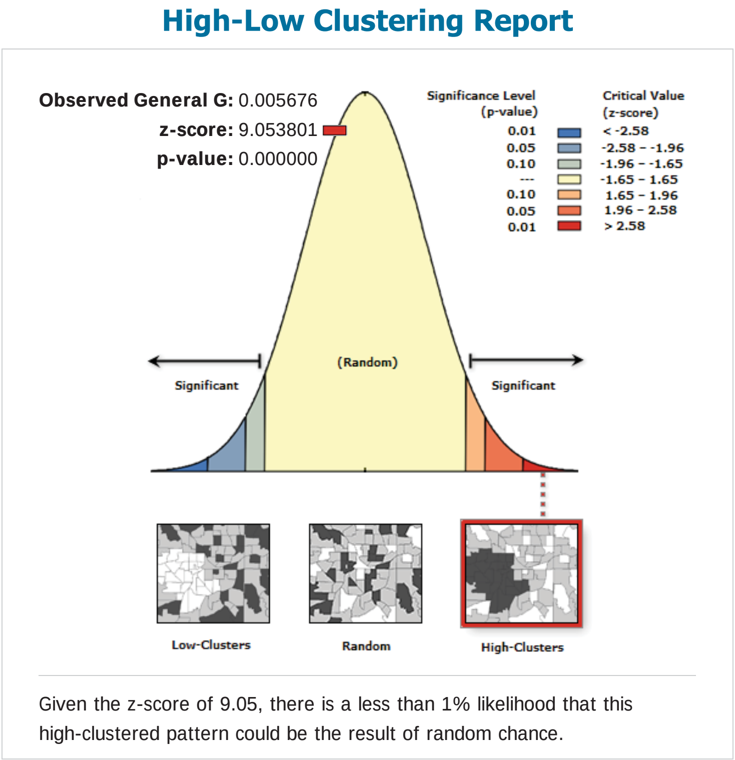
As with spatial autocorrelation, it is always a good idea to use several spatial weight matrices to check the robustness of the results. Here it is clear from the very different (implicit) bandwidths used in these two examples that the significance of spatial concentration in this case is firmly established.

Before moving on to the more interesting local tests of spatial concentration, it is of interest to note that such tests can also be done in ARCMAP. Here ARCMAP has for some reason chosen to use only G-statistics rather than G^*-statistics.[6] But in the more important case of local spatial concentration below, they do use G^*-statistics. So we shall not spend much time on this particular application, other than to note that it can be accessed by

[6] To see this, simply Google "How High/Low Clustering (Getis-Ord General G) works".

ArcToolbox > Spatial Statistics Tools
> Analyzing Patterns
> High/Low Clustering (Getis-Ord General G)

For sake of comparison with the MATLAB results above, we have used exactly the same procedure developed in Section 4.2.2 above for testing spatial autocorrelation in terms of **Wnn_5**. Here the only difference is that **General G** is used rather than **Moran's I**. The graphical output for this application is shown in Figure 5.2 below:



High-Low Clustering Report

Observed General G: 0.005676
z-score: 9.053801
p-value: 0.000000

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	🟩	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 9.05, there is a less than 1% likelihood that this high-clustered pattern could be the result of random chance.

Figure 5.2. Application of the G Statistic

Notice from the value of G = 0.005676 that this is the same value (when rounded) as that obtained in MATLAB above. Notice also that the result here is in terms of the asymptotic normal approximation of this G statistic (obtained by Getis-Ord, 1992, under the same random permutation hypothesis as above), and is thus reported as a z-score (9.0538) with extremely small p-value. This again suggests that the MATLAB results would continue to obtain maximal significance for many more permutations than 999.

5.3 Local Tests of Spatial Concentration

Observe that both G_W^* and G_W are decomposable into *local* measures of concentration about each location i as follows. Let the *local* G_W^* value at i be defined by

$$(5.3.1)\qquad G_W^*(i) = \frac{\sum_{j=1}^n w_{ij}x_j}{\sum_{j=1}^n x_j} = \sum_{j=1}^n p_j w_{ij}$$

and similarly, let the *local* G_W value at i be defined by

$$(5.3.2)\qquad G_W(i) = \frac{\sum_{j\neq i} w_{ij}x_j}{\sum_{j\neq i} x_j}$$

where, again, our interest focuses almost entirely on $G_W^*(i)$. Note in particular from (5.1.2) that these local measures of concentration are related to G_W^* by the identity,[7]

$$(5.3.3)\qquad G_W^* = \sum_{i=1}^n p_i \left(\sum_{j=1}^n p_j w_{ij}\right) = \sum_{i=1}^n p_i\, G_W^*(i)$$

Thus G_W^* can be viewed as a weighted average of these local concentration measures, where the weights, p_i, are simply the proportions of x in each areal unit i. In terms of the probability interpretation above, if we again consider accessibility weights of the form, $w_{ij} = a(d_{ij})$, then $G_a^*(i)$ is precisely the expected accessibility from a randomly sampled unit of x in i to any other randomly sampled unit, i.e., the *conditional* expected accessibility

$$(5.3.4)\qquad G_a^*(i) = \sum_{j=1}^n p_j\, a(d_{ij}) = E(a\,|\,i)$$

In these terms, it follows from (5.1.5) together with (5.3.4) that the decomposition in (5.3.3) is simply an instance of the standard conditional-expectation identity:

$$(5.3.5)\qquad E(a) = \sum_i p_i\, E(a\,|\,i)$$

But the real interest in these local measures is that they provide information about *where* concentration is and is not occurring.[8] In particular, by assigning p-values indicating the significance of local concentration at each areal unit, one can map the results and visualize the pattern of these significance levels. Those areas of high concentration are generally referred to as “hot spots” (in a manner completely analogous to strong clusters in point patterns).

[7] It is of interest to note that this decomposition is an instance of what Anselin (1995) has called *Local Indicators of Spatial Association* (LISA).

[8] Indeed, the original paper by Getis and Ord (1992) *starts* with these local indices, and only groups them into a “General G” statistic a later section of the paper.

5.3.1 Random Permutation Test

In this setting, one may *test* for the presence of such "hot spots" with respect to data set, $(x_i : i = 1,..,n)$ by employing essentially the same random permutation test as above. In particular, for any random permutation, $\pi = (\pi_1,..,\pi_n)$, of the areal unit indices $(1,..,n)$, one may compute for each unit i the associated statistic, $G_W^*(i)$, and compare this observed value with the distribution of values, $G_W^*(i,k)$ for N random permutations, $\pi^k = (\pi_1^k,..,\pi_n^k)$, $k = 1,..,N$. Here it is important to note that the index i is itself included in this permutation. For if the value of x_i is relatively large, then to reflect the significance of this local concentration at i it is important to allow smaller values to appear at i in other random permutations.

If the observed value of $G_W^*(i)$ has rank k_i among all values $[G_W^*(i), G_W^*(i,1),.., G_W^*(i,N)]$ (with rank 1 denoting the highest value), then the *significance of concentration* at i is again represented by the *p-value,*

$$P_i = \frac{k_i}{N+1} \ , \quad i = 1,..,n \,. \tag{5.3.6}$$

It is these values that are plotted to reveal visual patterns of concentration.

5.3.2 English Mortality Example

This testing procedure is implemented for local G^*-statistics in the MATLAB program, **g_perm_loc.m**. Here it is assumed that tests for *all* areal units, $i = 1,..,n$, are to be done. Hence the outputs contain the local G^*-statistic and *P*-value for each areal unit. To illustrate the use of this local-testing procedure, it is convenient to continue with the *English Mortality* example above. For the *exponential-distance weights matrix*, **W1**, constructed above, together with the myocardial infarction data, **z**, the command:

```
>> GP1 = g_perm_loc(z,W1,999);
```

yields a (190 x 2) output matrix **GP1** $= [(G_i^*, P_i) : i = 1,..,190]$ containing the local G^*-statistic, $G_i^*\ [= G_{W1}^*(i)]$ and *P*-value, P_i, for each of the 190 districts, based on 999 random permutations. These values were imported to ARCMAP and displayed in the map document, **Eng_mort.mxd**, as shown in Figure 5.3 and 5.4 below. Figure 5.3 plots the actual values of G_i^* in each areal unit, i, with darker green areas denoting higher values. The corresponding P-values are shown in Figure 5.4, where darker red shows the area of most significance (and where only the legend for P-values is shown). As expected, there is seen to be a rough correspondence between high local G^* values and more significant areas of concentration.

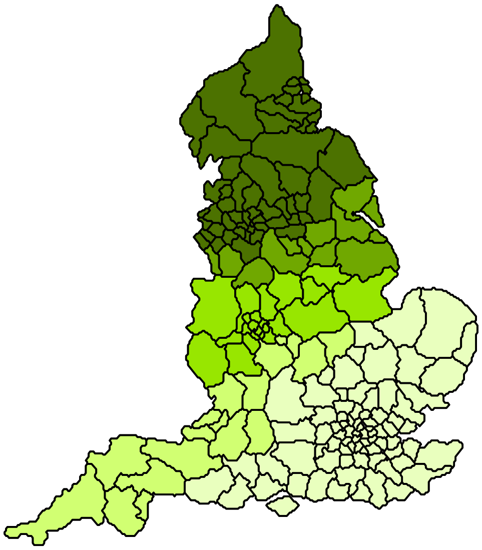


Fig.5.3. Exponential *G-Values**

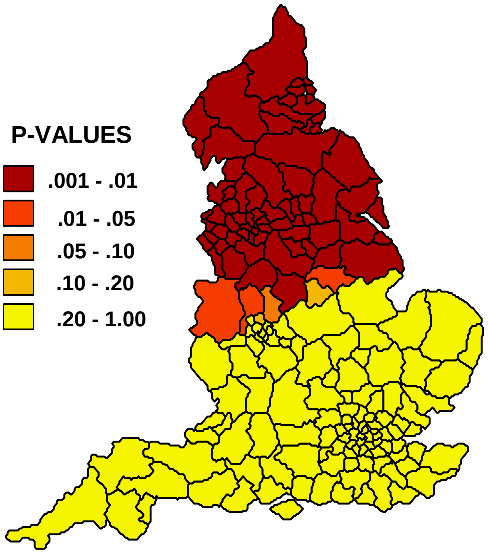


Fig.5.4. Exponential *P*-values

Notice in particular that the local G^*-values reflect the general concentration of myocardial infarction rates in the north that is seen in the original data set [Figure 1.9 (p.III.1-5)], but now are smoothed by the exponentially weighted averages in the local G^* statistics. However this "north-south" divide ([B-G], p.279) is seen to be much more dramatic in the associated *P*-values, where the darkest region, denoting *P*-values less than .01, now covers all of Northern England.

Turning next to the *nearest-neighbor weights matrix*, **Wnn_5**, the test results are now obtained with the command,

```
>> GP2 = g_perm_loc(z,Wnn_5,999);
```

which again yields a (190 x 2) output matrix **GP2** $=[(G_i^*,P_i):i=1,..,190]$ containing the local G^*-statistics and *P*-value for this case. By again importing these values to ARCMAP, we obtain the comparable displays shown in Figures 5.5 and 5.6 below. Notice that key difference between these two sets of results is the additional local variation in values created by the smaller numbers of neighbors used by **Wnn_5**. For example, while each areal unit has only 5 neighbors in **Wnn_5**, if we approximate the bandwidth in exponential matrix, **W1**, by counting only weights, $w_{ij} > .01$, then some areal units i still have more than 70 neighbors. So the degree of smoothing is much greater in the associated G_i^* values. But still, the highest values of both G_i^* and P_i continue to be in the north, and in fact are seen to agree more closely with those concentrations of myocardial infarction rates seen in the original data, such as the concentration seen around Lancashire county [compare Figure 1.6 (p.I.1-3) with Figure 1.9 (p.III.1-5)]. So it would appear that 5 nearest neighbor yields a more appropriate scale for this analysis.

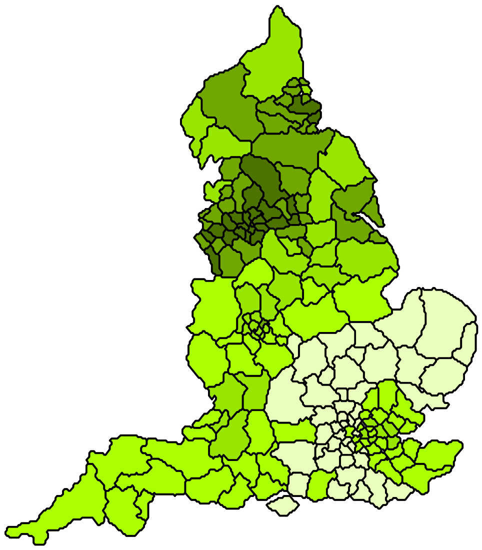


Fig.5.5. Nearest Neighbor *G-Values**

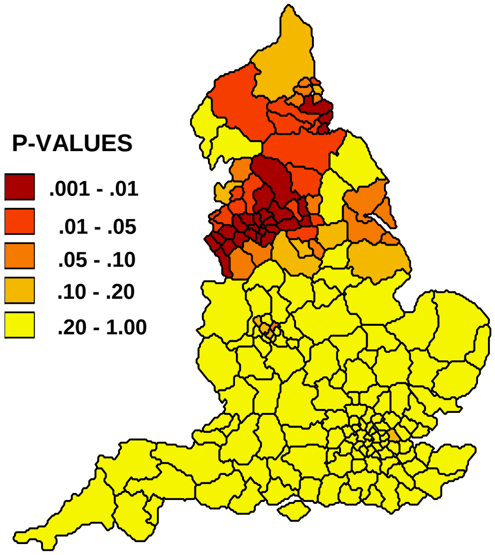


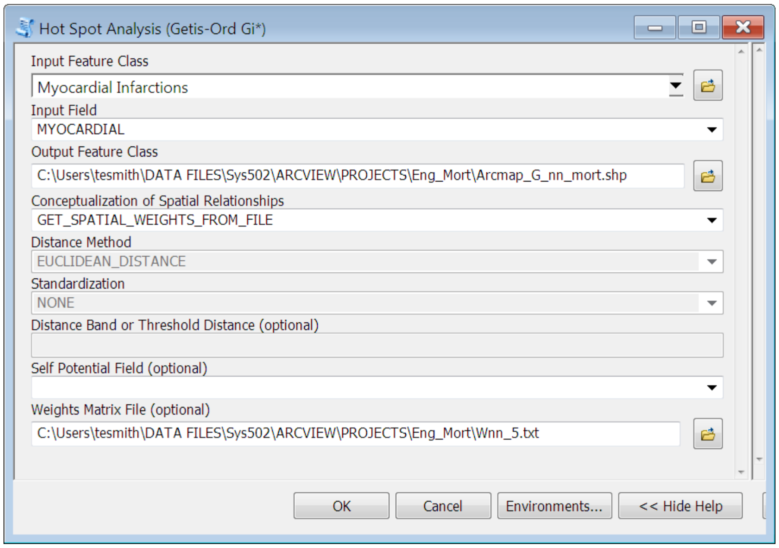
Fig.5.6. Nearest Neighbor *P*-Values

5.3.3 Asymptotic *G** Test in ARCMAP

An alternative test using G^* is available in ARCMAP. This procedure can be found at:

ArcToolbox > Spatial Statistics Tool
> Mapping Clusters
> Hot Spot Analysis (Getis-Ord *G)**

To employ this procedure, we will again use the English Mortality data with the nearest-neighbor spatial weights matrix, **Wnn_5**, already constructed for ARCMAP in Section 4.3.2. In the Hot Spot window that opens, type:



Hot Spot Analysis (Getis-Ord Gi*)

Input Feature Class: Myocardial Infarctions

Input Field: MYOCARDIAL

Output Feature Class: C:\Users\tesmith\DATA FILES\Sys502\ARCVIEW\PROJECTS\Eng_Mort\Arcmap_G_nn_mort.shp

Conceptualization of Spatial Relationships: GET_SPATIAL_WEIGHTS_FROM_FILE

Distance Method: EUCLIDEAN_DISTANCE

Standardization: NONE

Distance Band or Threshold Distance (optional):

Self Potential Field (optional):

Weights Matrix File (optional): C:\Users\tesmith\DATA FILES\Sys502\ARCVIEW\PROJECTS\Eng_Mort\Wnn_5.txt where the specific path names will of course vary. Click **OK**, and a shapefile will be constructed and added to the Table of Contents in your map document. The result displayed is shown in Figure 5.7 below (where the legend from the Table of Contents has been added).

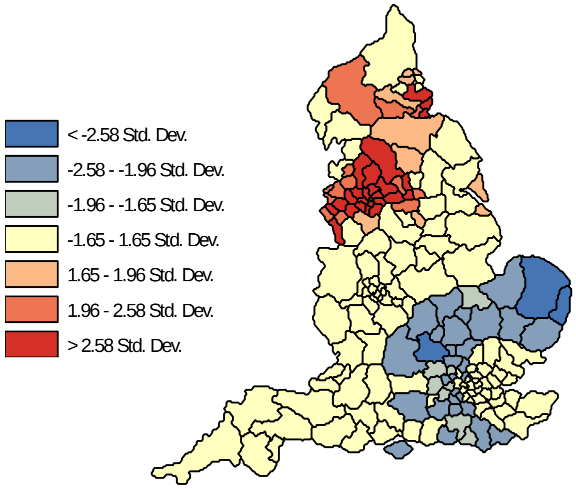


Figure 5.7. Asymptotic *G Test Output**

As with the **General G** test in Figure 5.2 above, this test is based on the asymptotic normal approximation of the local G^*-statistics under the same random permutation hypothesis as above. So the values shown in the legend above are actually in terms of the z-scores obtained for each test. For example, the familiar "1.96-2.58" valued in the second to last red entry indicates that myocardial infarction rates for districts with this color are significantly concentrated at between the .05 and .01 level. (The actual p-values are listed in the Attribute Table for this map). Here it is important to note that *two-sided tests* are being performed. So for a corresponding *one-sided test* (as done above), these values are actually twice as significant (i.e., with one-sided p-values between .025 and .005). So even though the red areas look slightly "smaller" than those in Figure 5.6, the results are actually *more* significant than those of MATLAB, in a manner consistent with all of the asymptotic tests we have seen so far. Notice also that because two-sided tests are being done, it is also appropriate show areas with significantly *less* concentration than would be expected under the null hypothesis. These districts are shown in blue.

5.3.4. The Advantage of *G** over *G* for Analyzing Spatial Concentration

Before leaving this topic, it is instructive to consider an additional example that illustrates the advantage of local G^*-statistics over G-statistics for the analysis of spatial concentration. Here we construct a fictitious population distribution for the case of Eire in which it is assumed that there is a single major concentration of population in one county (FID 18 = "Offaly" County), as shown in Figure 5.8 below.[9]

[9] In particular, about 25% of the population has been placed in this county, and the rest has been distributed randomly (under the additional condition that no other county containing more than 5% of the population).

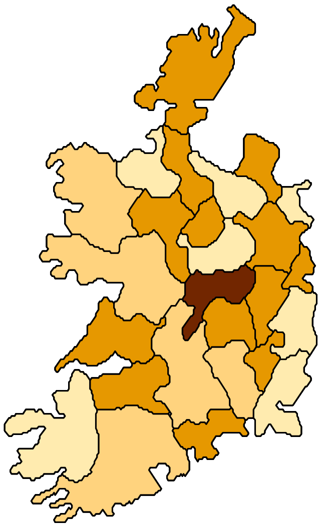


Fig.5.8. Fictitious Data

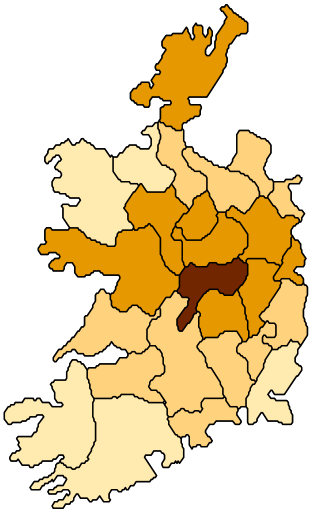


Fig.5.9. Exponential *G-Values**

Here an exponential-distance matrix has been constructed similar to **W1** above (to ensure a smooth representation), and the local G^*-statistics for this case are shown in Figure 5.9. Notice these these G^*-values roughly approximate the concentration of the original data, but are somewhat smoother (as was also seen for the myocardial infarction data above using **W1**). The corresponding P-values (again for 999 simulations) are shown in Figure 5.10 below.

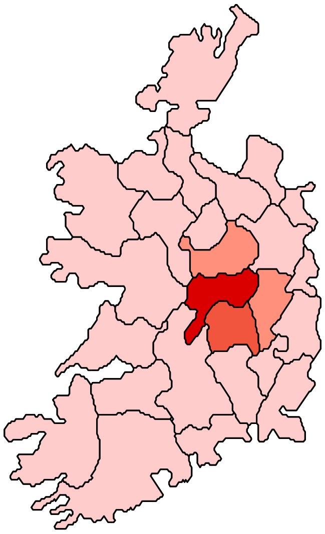


Fig.5.10. P-Values for G*

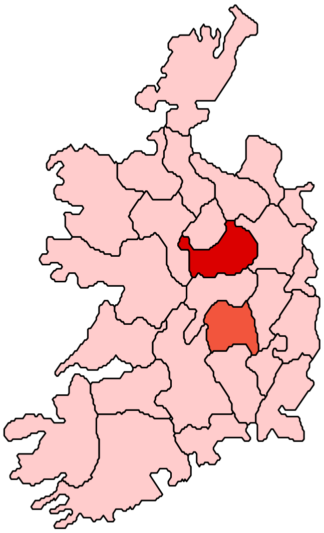


Fig.5.11. P-Values for G

These results confirm that Offaly County is the overwhelmingly most significant concentration of population ($\text{P-Value} \approx .02$), with several of the surrounding counties gaining significance from their proximity to Offaly. However, if one carries out the same test procedure using local G-statistics, then a substantially different picture immerges. Here Offaly County is not in the least significant – but two of its immediate neighbors are. The reason of course is that by setting the matrix diagonal to zero, the population of Offaly itself is ignored in the local G-test for this county. Moreover, since its neighbors do not exhibit unusually high population concentrations, the local G-value for Offaly will not be unusually high compared to the corresponding values for random permutations of county populations. However, its *neighbors* are still likely to exhibit significantly high values, because their proximity to the population concentration in Offaly yields unusually high local G-values compared to those for random permutations. Hence the anticipated result here is something like a "donut hot spot", with the "donut hole" corresponding to Offaly. This is basically what is seen in Figure 5.10, except that some neighbors are closer (in exponential proximities) to Offaly than others. This extreme example serves to underscore the difference between these two local statistics, and shows that local G^*-statistics are far more appropriate for identifying significant local concentrations.