ASSIGNMENT 5

Study 1: Irish Blood Group

1 INTRODUCTION

Recall that in previous reports we have discussed ***point pattern analysis*** ([A1], [A2]) and spatial ***continuous data analysis*** ([A3], [A4]), both of which focus on data in the form of point. In this study, we are going to introduce ***areal data***, whose primary departure from continuous data is basically its own ***form***. Specifically, compared to continuous data designated as ***realizations (point samples)*** from a continuous spatial distribution (e.g., observed Cobalt measurements at each location), areal data is in terms of ***aggregated*** quantities as the overall representatives for each areal unit within some relevant spatial division of a study region (e.g., adult blood group A percentages in each country of Eire shown in the left of Figure 1 below, which is the exact object we attempt to examine in this study).

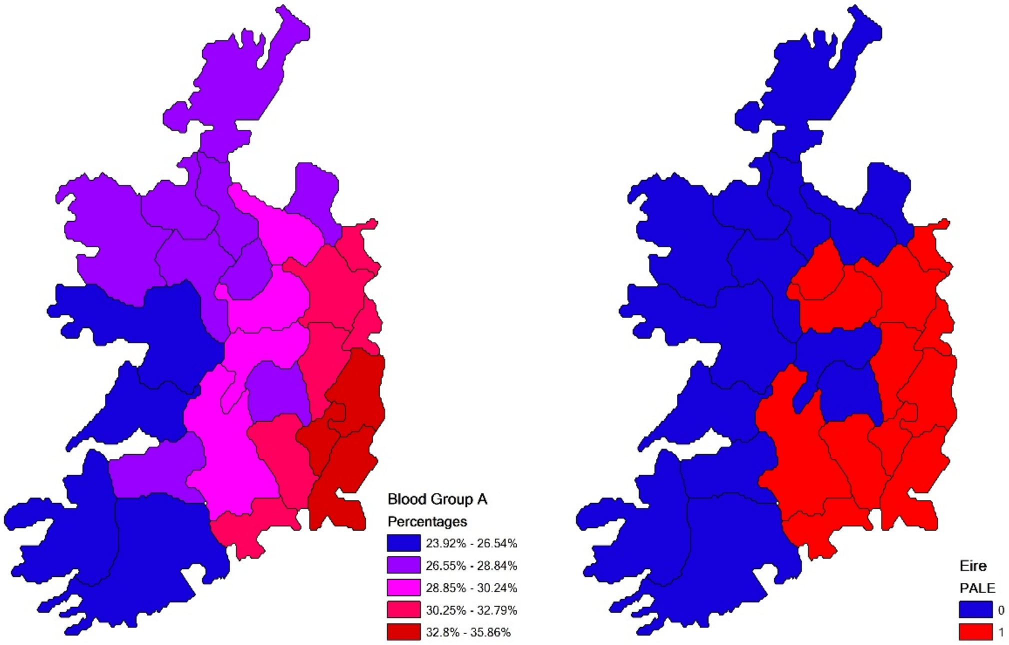

Figure 1. Blood Group A Percentages for Each County of Eire vs. Irish Pale

Visually, we can find some degree of ***concentration*** signs of blood group A percentages in the east Eire (i.e., counties with the highest percentages cluster there), associated with a gradient westbound decline of percentages. Also, notice that areas with relatively higher blood group A percentages to some extent corresponds to the Pale, which used to be under direct control of England in the Late Middle Ages. Such correspondence turns out not to occur by chance, but be a reasonably assumed historical result that the impact of Anglo-Norman colonization of Ireland in the 12th Century was maintained till the present and could be reflected in population composition

It's of interest to find some interpretations of such separation of blood group A from anthropologic researches: Relethford mentioned that the west-east gradient could be explained by historical waves of immigrants that originated from the east and southeast Eire and pushed previous populations further westward, including Anglo-Norman invasion in 1169 and subsequent settlements. Moreover, geographic distance acted to isolate specific population subdivisions in space and hence lead to genetic similarity, of which ABO blood group percentages turn out to be one feature.

Hence, from the perspective of spatial analysis and statistics, the objective of this study is to apply local ***G*-statistics*** (developed by Getis and Ord) to this Irish Blood Group data obtained from [BG, p.253] for the 26 counties of Eire, in order to analyze whether blood group A percentages are ***centered*** (i.e., ***concentration*** pattern) in some specific areas or not, whose magnitude is determined by the G*-statistics ***values*** and associated ***p-values***.

2 METHODS & RESULTS

2.1 Point Representations of Areal Units - Spatial Centroids

The analysis starts with the interpretation of another major difference between continuous data and areal data, designated as the representation of ***spatial structure*** itself. Unlike the unambiguity of "distance between points", it's relatively difficult to identify "distance between areal units", which is conventionally solved by defining such measurement as distance between ***representative points*** of respective areal units. Usually, a computationally convenient choice for such representatives is the spatial ***centroid*** of areal unit, *R*. If we denote the area of *R* as *a*(*R*), and the centroid as *c*, the centroid can be calculated by minimizing the average squared distance to all points in *R*:

$$(2.1.1) \quad \min_c \frac{1}{a(R)} \int_R \| x - c \|^2 \, dx$$

Hence, each areal unit (county in this context) can be represented by corresponding centroid with aggregated quantities (blood group A percentages) of the whole unit (as shown in Figure 2 below) for the following analysis.

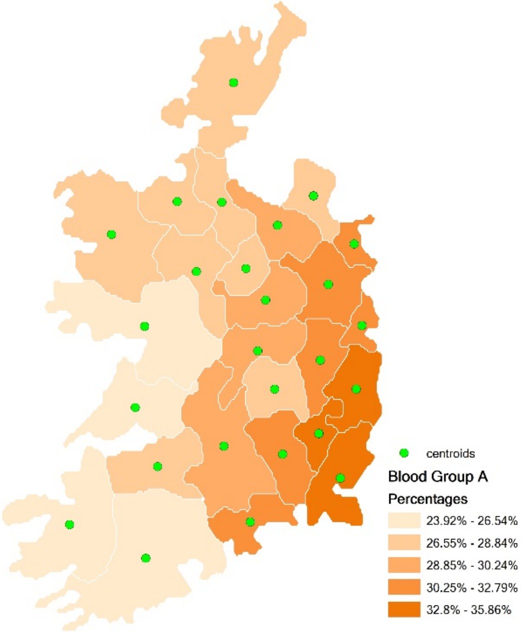

Figure 2. Centroids of Each County of Eire

2.2 Spatial Weights Matrices

Before continuing the interpretation and construction of the G*-statistics, recall that this index serves to measure the concentration, which is a typical case of ***spatial relations***. Hence, in order to model spatial relations between areal units, another concept to be defined here is ***spatial weight matrices***. In the context of the Irish blood group study, the study region, R = Eire, is divided into n = 26 counties, $R = \{R_i : i = 1,...,n\}$, as shown in Figure 2 above. It's basically hypothesized that the "spatial influence", "proximity" or any other reasonable spatial relations of R_j on R_i can be described as a numerical weight, $w_{ij} \geq 0$, where higher values indicate higher levels of such relations. Thus, such spatial relations between each pairwise units R_j and R_i within the full set can be represented as a single nonnegative ***weight matrix***:

(2.2.1) $$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}$$

There are a set of different spatial weights based on centroids, among which distance usually acts as an important factor of spatial influence. That is, for a target areal unit (e.g., a county), its closer neighbor counties in space (measured as distance between centroids) are assumed to make more impacts on it. Such a spatial matrix determined by centroid distance we are going to use in this study is ***Exponential-Distance Weights***:

(2.2.2) $$w_{ij} = \exp(-\alpha d_{ij})$$

Given the definitions of centroid and weight matrix, then we use MATLAB to compute each centroid location for these 26 counties of Eire, based on which an ***exponential*** matrix is created. Notice that the matrix we make here is a ***row-normalized*** one by normalizing the weights in each row of the matrix to uniform sum:

(2.2.3) $$\sum\nolimits_{j=1}^{n} w_{ij} = 1 \quad , \quad i = 1,\ldots,n \quad ,$$

for the purpose of removing dependence on extraneous scale factors in exponential weights. Another notable thing here is that all diagonal weights in the matrix is as calculated (i.e., exp(0) = 1, since the distance between a given county and itself is zero).

2.3 Spatial Concentration Index

Based on the exponential weight matrix, w_{ij}, defined above, next we will apply local ***G*-statistics*** to analyze and measure the degree of ***spatial concentration***between blood group A percentages of these 26 counties (observed in Figure 2 above). Generally, for the given data set of blood group A percentages, denoted as $\{x = (x_1,\ldots,x_n)', n = 26\}$, and associated exponential weight matrix, w_{ij}, written in the form of vector, $W = (w_{ij} : i, j = 1,\ldots,26)$, the G*-statistic for *x* is defined as:

(2.3.1) $$G^*{}_W(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i w_{ij} x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} = \frac{x'Wx}{(1_n' x)^2}$$

Similarly, the ***local*** measures of concentration for each county *i* can be defined as:

$$(2.3.2) \qquad G^*_W(i) = \frac{\sum_{j=1}^{n} w_{ij}x_j}{\sum_{j=1}^{n} x_j} = \sum_{j=1}^{n} p_j w_{ij} \quad ,$$

where p_j is a probabilistic form denoted as:

$$(2.3.3) \qquad p_j = \frac{x_j}{\sum_{i=1}^{n} x_j} = \frac{x_j}{1_n' x_j} \quad ,$$

In MATLAB, we can easily calculate G*-statistics of blood group A percentages for each county of Eire. The evaluated results are shown in Figure 3 below (after being joined to "Blood Group A" map in Figure 1), from which we can find out that similar concentration patterns among G*-statistics values that higher values denoted as darker red also concentrate in east Eire (Pale areas). Notice that we can't determine whether the levels of spatial concentration are high or low based on the G*-statistics values alone. The method to identify the relative magnitude of a specific county's observed G*-statistics value is to compare it with those values under ***null hypothesis***, H_0, of no spatial concentration that can be simulated by random permutations.

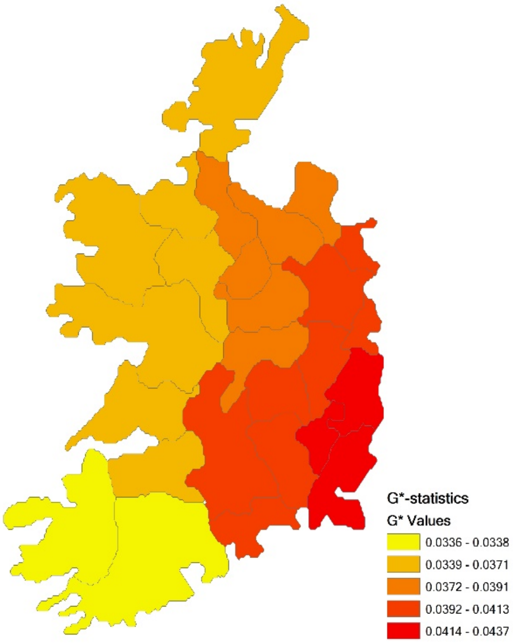

Figure 3. Exponential G*-Values

2.4 A Random Permutation Test of Spatial Concentration

Such random permutation test is based on Monte Carlo methods that have been used in random rebelling of [A2]. To start with, we postulate that the particular spatial arrangement of samples doesn't matter and hence the labeling of these samples does not, either. Specifically, it shouldn't matter which G*-statistics is labeled as "G^*_1", "G^*_2", or "G^*_n". With this overview, taking the Blood Group data as an example, the permutation test steps for each given county *i* are described as followed (similar process for significance calculation is applied for three spatial autocorrelation statistics in study 2 followed):

- Compute G*-statistics of the ***original*** blood group A percentages for this county.
- ***Randomly simulate (permute)*** the percentages of 26 counties by ***relabeling*** to create a new realization, and calculate the respective G*-statistics for this target county in the ***newly*** permuted map.
- ***Repeat*** the simulation process and associated calculation of G*-statistics a large enough number of times (here, 9,999 times).
- ***Rank*** all the 10,000 G*-statistics (one for the original and 9,999 for the random permutations) in ***descending*** order (i.e., ***from high to low***), so that the rank of a realization is designated as *k* if it yields the *k*th highest value.
- ***Compare*** the G*-statistics of original sample data with the ***distribution*** of simulated statistics under H_0. If the observed G*-statistics value for county *i*, $G^*(i)$, has rank k_i among all values $[G^*(i), G^*(i,1), ..., G^*(i,N), N = 9999]$ (with rank denoting the highest value), the ***significance of concentration*** at county *i* is represent by the ***p-value***:

$P_i = \dfrac{k_i}{N+1}$, $i = 1,...,n$ (here, n = 26 counties, N = 9,999 permutations)

In MATLAB, the p-values are also calculated associated with corresponding G*-statistics at each county, which can be easily imported to ArcMap and displayed for the subsequent discussion. We then present the Blood Group A map and P-values map side by side for comparison below.

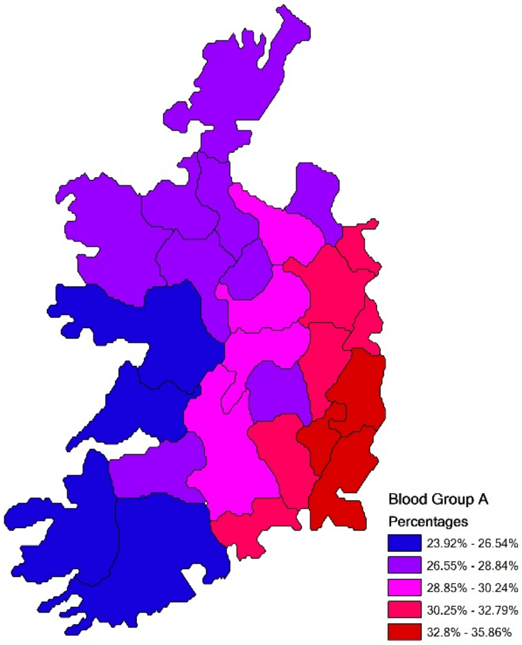

Figure 4. Blood Group A Percentages

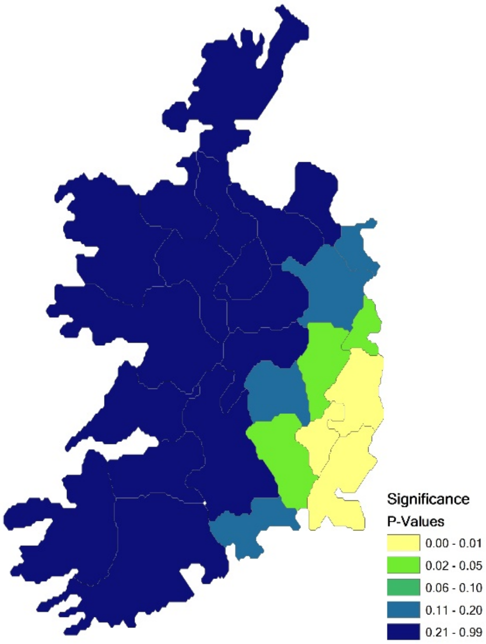

Figure 5. Exponential P-Values

Figure 4 plots the adult blood group A percentages of each county in Eire, with darker red areas denoting higher values. The corresponding p-values are presented in Figure 5, where lighter yellow corresponds to areas of most significance. It's expected to observe such more concentration in the East of Eire (indicated by stronger statistical significances) that agrees with observations and the history.

Hence, it's of interest to look inside some specific counties. For example, the blood group A percentage of Waterford (denoted in grey in Figure 6) is 31.04% (ranking 6th highest), a little bit higher than the mean value of Eire designated as 29.53%, which can be considered as somewhat slightly more concentrated than other areas. However, the associated p-value for G*-statistics designated as 0.1654 is ***not statistically significant*** (ranking 10th lowest). More specifically, the evaluated chance of observing this large a G*-statistics value here is 1654/10000 based on permutation test. This ***discrepancy*** can be accounted by its location ***at the middle***. Recall that the gradient decline of blood group A percentages turns out to be a result of historical population movements ***from east coast*** and hence geographic separation. Therefore, the ***closer to the east coast*** where Anglo-Norman invaders landed and settled, the more concentrated blood group A percentages have become due to ***stronger anthropologic influence***, while Waterford is located at the middle rather than the east Eire. Additionally, mathematically, the computed values of G*-statistics for each county are more relied on neighboring counties in the vicinity due to exponential weight matrix that assigns more weights on closer neighbors. Compared to counties in the east that are surrounded by counties with higher percentages, one of Waterford's three contiguous neighbors (outlined in black in Figure 6) is with a percentage below the mean designated as 23.92%, and would definitely result in relatively lower G*-statistics, which is not large enough to rank in the top of permutations.

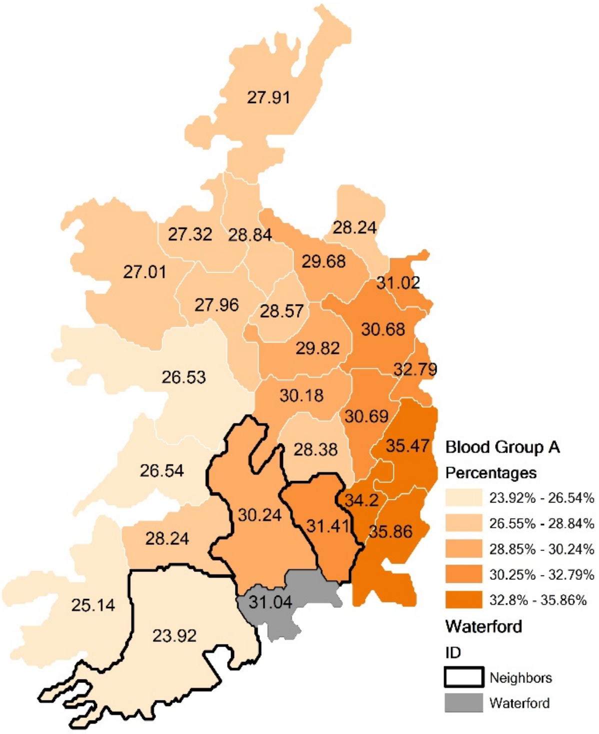

Figure 6. Waterford and Its Contiguous Neighboring Counties

Finally, we move on to take a look at the counties of Kildare and Meath (denoted as dark and light blue in Figure 7), both of which yield similar blood group A percentages designated as 30.69% and 30.68% respectively. However, the discrepancy is found among the significance levels that the p-values of Kildare is designated as 0.0268 (significant), while that of Meath is 0.1158 (insignificant). Similar to the influence of locations and surrounding neighbors we have discussed above, Meath is located at the north whose most-influential nearest neighbors exhibit relatively lower percentages than those of Kildare, which in particular are closer to the three counties with the highest percentages in the southeast Eire, where the anthropologic influences of Anglo-Norman colonization turns out to be the most.

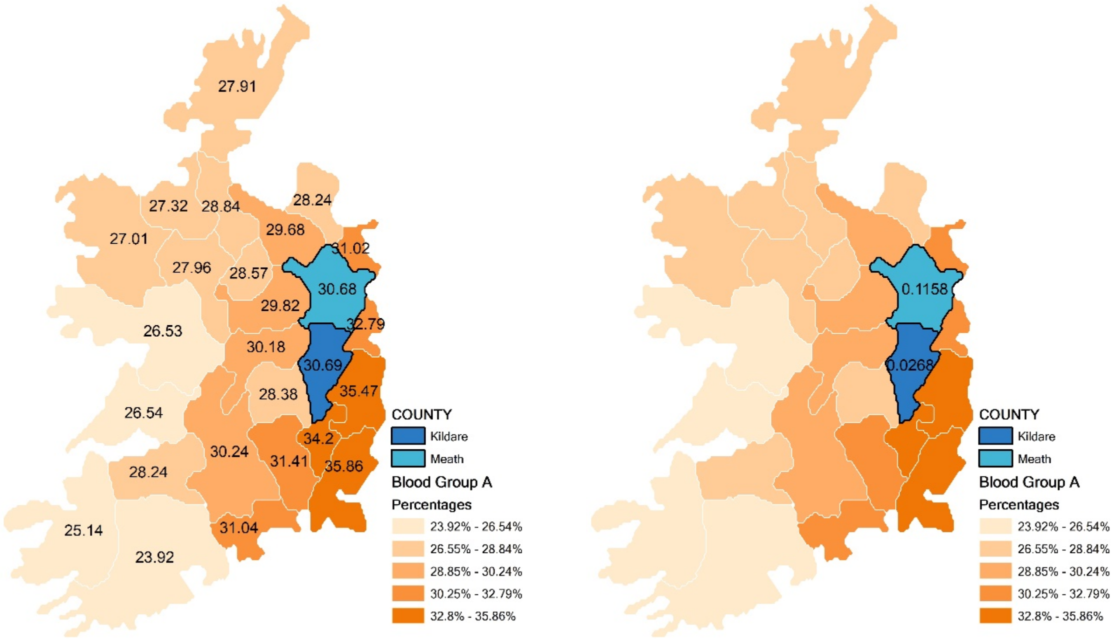

Figure 7. Kildare and Meath

3 DISCUSSION

In this report, we have built up the basic framework for areal unit data analysis, including spatial centroid, spatial weight matrix and a useful tool applied to examine the existence and magnitude of spatial concentration, based on the “Blood Group A” study case. Given the analysis results above, we can conclude that for the counties locating at the southeast Eire, the null hypothesis for blood group A percentages is rejected that they do indeed exhibit statistically significant spatial concentration compared to the others. These results do agree well with findings and conclusions of some anthropologic researches.

Notice that the calculation of G*-statistics and associated p-values take all areal units into consideration, so that the concept of concentration is relative, which means we cannot say the variable value of interest at a specific areal unit is more spatially centered than the others by examining each unit with only a part of its neighbors alone

4 REFERENCE

D. Tills, P. Teesdale and A. E. Mourant, “Blood groups of the Irish”

J. H. Relethford, “Genetic structure and population history of Ireland: a comparison of blood group and anthropometric analyses”

W. E. R. Hackett, G. W. P. Dawson and C. J. Dawsont, “The Pattern of the ABO Blood Group Frequencies in Ireland”

[A1-A4] J.Wu (2019), “Assignment 1 - 4: Studies for ESE 502”