AREAL DATA ANALYSIS

1. Overview of Areal Data Analysis

The key difference between *areal* data and *continuous* data is basically in terms of the *form* of the data itself. While continuous data involves *point samples* from a continuous spatial distribution (such as temperature readings at various point locations), areal data involves *aggregated* quantities for each *areal unit* within some relevant spatial partition of a given region (such as census tracts within a city, or counties within a state). Such differences are illustrated in Figures 1.1 and 1.2 below.

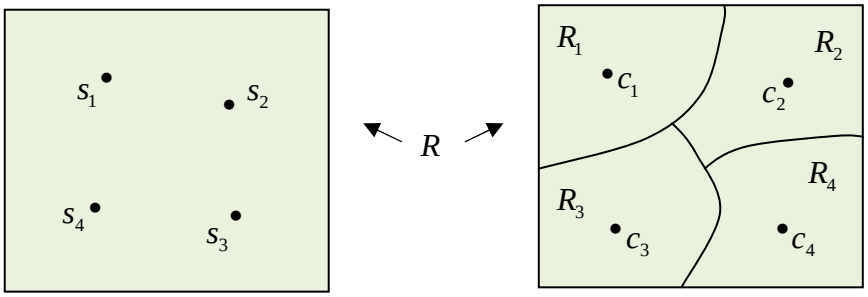


Figure 1.1. Point Samples

Figure 1.2. Areal Units

Here Figure 1.1 shows four sample points, s_i, in region, *R*, (which is qualitatively the same as Figure 1.1 of Part II), and Figure 1.2 represents a partition of region, *R*, into four areal units $\{R_1, R_2, R_3, R_4\}$. Such areal units, R_i, are often represented by appropriate *central locations*, $c_i \in R_i$, such as major cities, or geometric "centroids" (to be defined below).[1] But the data values associated with these points represent summary measures for the areal unit as a whole. For example, rather than measuring the temperature at location, c_i, one could assign (an estimate of) the average temperature over all points in areal unit R_i. More importantly, one can represent values that have no particular point locations at all, such as the population of R_i or the average income of all household units in R_i.

The practical significance of areal data for purposes of analysis is that most socio-economic data comes in this form. For example, while individual income data is generally regarded as proprietary in nature, such data is often made publically available in terms of averages (such as per capita income at the state or county level). More generally, most publically available data (such as US Census data) is only of this type.[2]

[1] This type of representation in terms of point locations has led to the alternative description of areal data as "lattice data", as for example in Cressie (1993, Section 6.1).

[2] There are exceptions however, such as the Center for Economic Studies (CES) run by the Census Bureau, which allows restricted access to individual micro data by qualified researchers.

As in Parts I and II above, it is appropriate to illustrate some of the key features of areal data in terms of specific examples (again drawn from [BG, Part D]).

1.1 Extensive versus Intensive Data Representations

Areal data is most easily represented visually in terms of choropleth maps, such as the *child mortality data* for each of the 167 Census Districts in the city of Auckland, New Zealand over the nine year period from 1977 to 1985 (taken from [BG, pp.249, 300-303]). Here we focus only on the populations "at risk", i.e., children under the age of 5 in each district. Two possible representations of this data are shown in Figures 1.3 and 1.4 below.

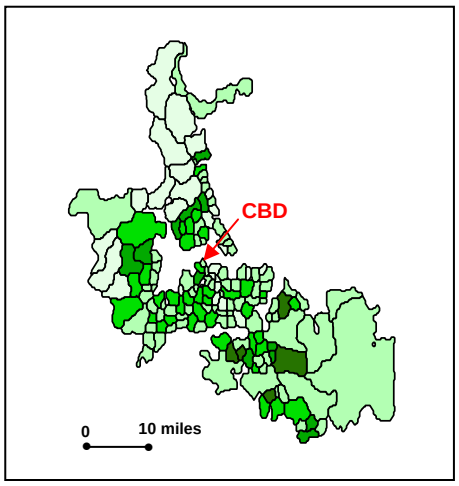


Figure 1.3. Raw Population Data

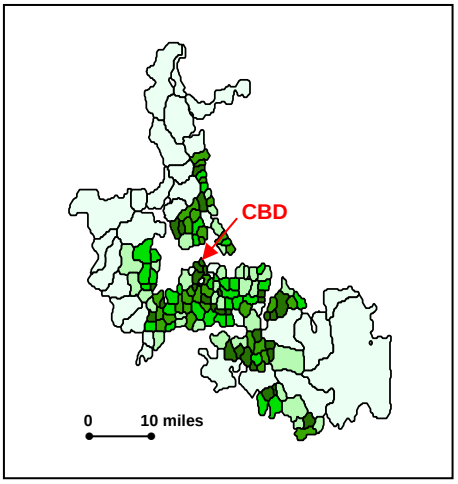


Figure 1.4. Population Density Data

The representation in Figure 1.3, which shows the actual number of children under 5 in each district, appears to suggest that the most substantial concentration of these children lies in districts to the southeast of the Central Business District (CBD). But it is important to note that census districts are specifically designed to include roughly the same population totals in each district. So the smaller districts around the CBD indicate that population densities are much higher in this area.

An alternative representation of this population in given in Figure 1.4, which displays the density of such children in each district, i.e., the number of children per square mile (approx.) Here it is clear that the most dense concentrations are precisely in the smallest districts, including the CBD. So this representation suggests (not surprisingly) that children under five are in fact quite evenly spread throughout the population as a whole.

This example serves to underscore the fact that the distribution of areal data is usually more accurately represented in terms of density values. More generally, representations in terms of actual data totals (such a population counts) are designated as *extensive* representations of areal data, and representations in terms of densities (such as population densities) are designated as *intensive* representations of areal data. The key difference is that intensive representations allow more direct comparisons between values in each areal unit. For example, "population per square mile" has the same meaning everywhere, and is *independent of the size of each areal unit*.[3]

1.2 Spatial Pattern Analysis

The above example also demonstrates that when intensive data representations are used, choropleth maps can serve to reveal meaningful patterns in areal data. In this case, children under five (and indeed all people) are more concentrated around the CBD than in outlying areas. But there are many more interesting pattern examples that this.

One example of *comparative pattern analysis* is provided the Chinese socio-economic data in [BG, p.249-250]. Figures 1.5 and 1.6 below show the shift in per capita gross domestic product (pcGDP) in the provinces of China from 1984 to 1994. (Note again that this data is in *intensive* form, where "per capita" indicates that the relevant units of comparison here are "individuals" rather than "square miles".)[4]

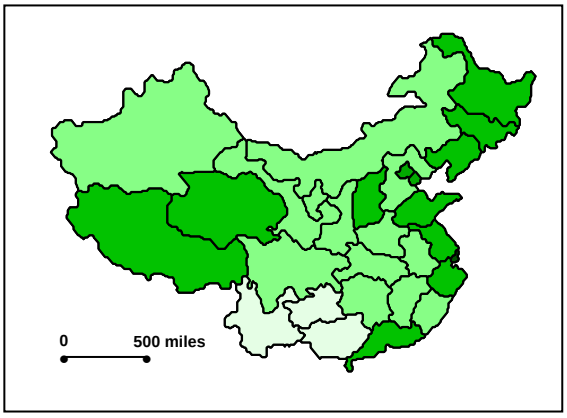


Figure 1.5. 1984 Per Capita GDP

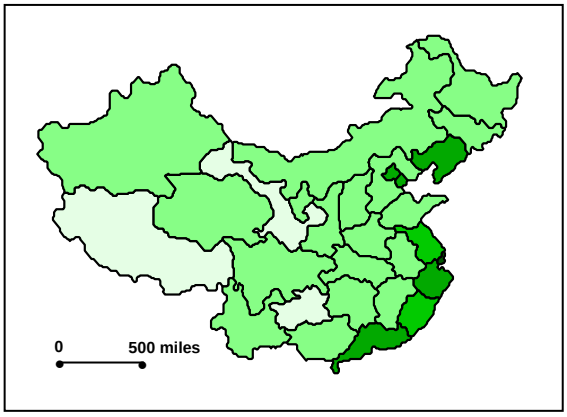


Figure 1.6. 1994 Per Capita GDP

Here it is clear at a glance that coastal region of China has been the *high growth area*. Statistical analysis can of course be applied to confirm this. But the key point here is that *visual pattern analysis* is a powerful *heuristic* tool for discerning relations that may not be immediately evident in the data itself.

[3] For a more detailed discussion of intensive versus extensive data representations, see the classic paper by Goodchild and Lam (1980).
[4] To allow direct comparison, data on both maps has been normalized to have unit maximum values.

A second example is provided by the Irish blood group data from [BG, p.253] for the 26 counties of Eire. From an historical perspective, there is strong reason to believe that the Anglo-Norman colonization of Ireland in the 12th Century had a lasting effect on the population composition. Figure 1.7 below shows the estimated proportion of adults in each county with blood group A in 1958 (where values increase from blue to red). Figure 1.8 shows the original colonized area of Eire, known as the "Pale". Since blood group A is much more common among individuals with Anglo-Norman heritage, a visual comparison of these two figures strongly suggests a continued pattern of Anglo-Norman influence in the region around the Pale. We shall later confirm these findings with spatial regression.

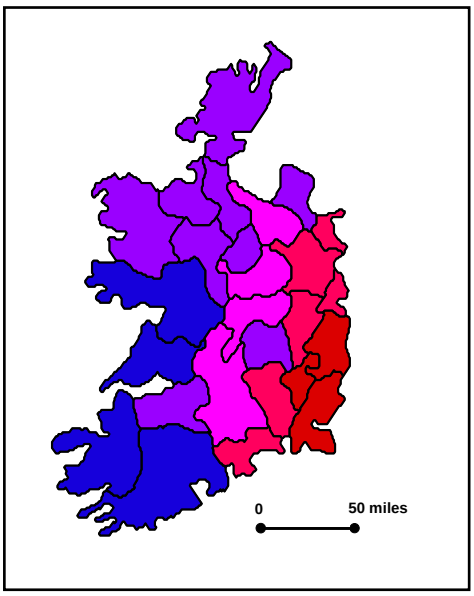


Figure 1.7. Blood Group A Percentages

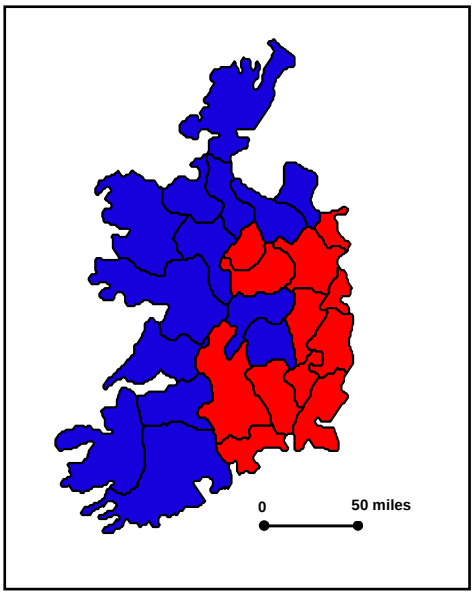


Figure 1.8. Counties in the Pale

1.3 Spatial Regression Analysis

A final example of areal data is provided by the English Mortality data from [BG, pp.252-253]. Here the areal units are the 190 Health Authority Districts throughout England, and the data used involve deaths from *myocardial infarctions* among males (35-64), as shown in Figure 1.9 below. This data is in *standardized rates*, defined here to be the number of deaths in the period 1984-1989 divided by the expected number of deaths during that period based on national averages. Such standardized rates are quite typical for medical data, and help to identify those areas where death rates are much higher than expected relative to national averages. In particular, the darkest areas on the map indicate rates well above average. While such higher rates may be influenced by many factors, the present analysis focuses on aspects of "social deprivation" as summarized by the "Jarman underprivileged areas score", or *Jarman score* (which is a weighted average of factors including levels of unemployment and overcrowding). This measure for each Health

Authority District is shown in Figure 1.10, where darker areas here show higher levels of "social deprivation".

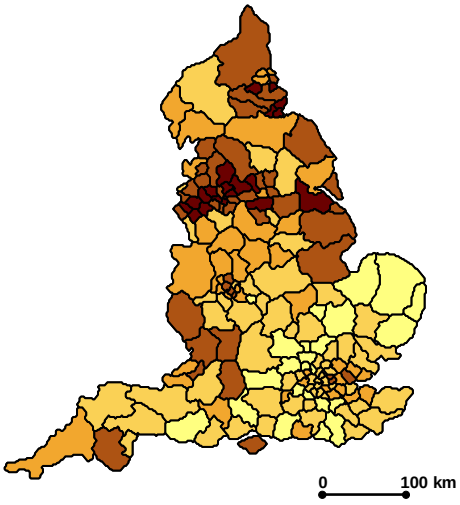


Figure 1.9. Myocardial Infarctions

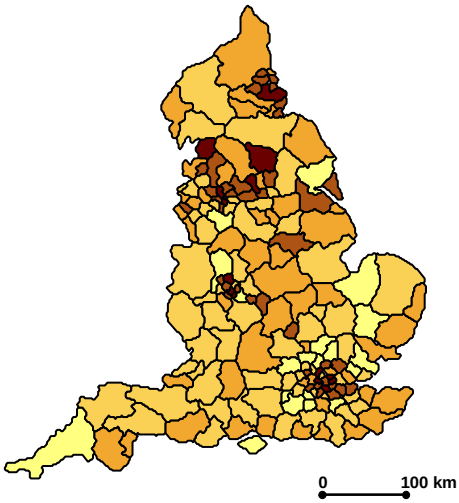
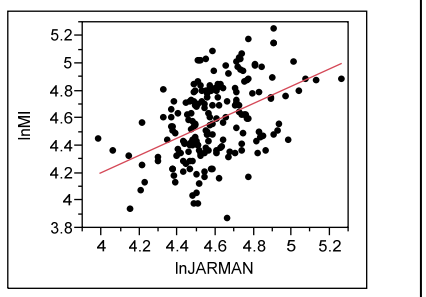


Figure 1.10. Jarman Scores

A visual comparison of these two maps suggests that there may indeed be some positive correlation between these two patterns, especially in Northern England where the highest levels of both death rates and social deprivation seem to occur.

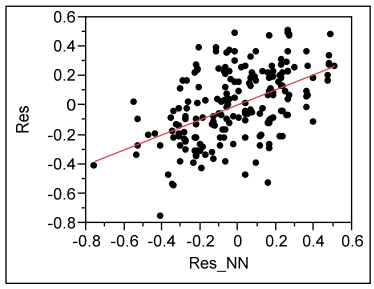
This relation can be readily confirmed by a simple regression of log Myocardial Infarction (**lnMI**) rates on log Jarman scores (**lnJARMAN**). The results in Figure 1.11 below show that there is indeed a very strong relation between the two.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.6897938	0.424182	3.98	<.0001*
lnJARMAN	0.6290693	0.09235	6.81	<.0001*

Figure 1.11. Regression Results



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0061292	0.015545	0.39	0.6938
Res_NN	0.5135203	0.063817	8.05	<.0001*

Figure 1.12. NN-Residual Analysis

However, it is also clear from Figures 1.9 and 1.10 above that there is a strong correlation between both MI rates and Jarman scores in neighboring districts. Moreover, since it is highly unlikely that the correlations among Jarman scores could completely account for those among MI rates, one can expect there to be a strong spatial autocorrelation among the regression residuals. This is confirmed by the simple nearest-neighbor analysis of these regression residuals shown in Figure 1.12 (where nearest neighbors are here defined with respect to centroid distances between districts). In fact the correlation among these residuals is even stronger than that between **lnMI** and **lnJARMAN** (as can be seen by comparing the t-ratios of **Res_NN** versus **lnJARMON**). While much of this residual correlation could in principle be removed by including a range of other relevant explanatory variables, it is quite apparent from Figure 1.9 that significant autocorrelation will remain.

With these observations, our ultimate objective is to extend this simple nearest-neighbor analysis to a broader and more rigorous framework for spatial autocorrelation analyses of areal data. But to do so, we must first address the difficult issue of defining appropriate measures of "distance" between areal units.