



Activity 2.1 - Spatial data exploration across geographic scales

Andy Nelson, CIAT

2000

Executive Summary

Natural resource management problems and economic problems related to agriculture often transcend field or farm boundaries, and can only be understood and corrected through adopting broader perspectives. Progress, however toward the goal of more productive, sustainable and healthy hillside environments has been hindered by; a lack of clear objectives, a failure to quantify variables, and a lack of precision in defining physical areas of interest. All of which are indispensable for arriving at negotiated agreements for community action as well as reproducing results achieved. These shortcomings can be largely attributed to a confusing range of temporal and spatial perspectives among stakeholders. The temporal and spatial interdependence that so characterises many resource management problems therefore necessitates some form of collective action among landscape users. This is the rationale for the objective of this project.

Project objectives

To develop and document principles and procedures for building a scale consistent database and for performing multiscale characterisation of agroecosystems using Honduran hillsides as an example. The objective will be reached through four project outputs:

1. Creation of a quality controlled, multiscale spatial database for Honduras with associated methodology training workbooks.
2. A multiscale characterisation of Honduran agroecosystems for targeting problems, priority areas and beneficiaries.
3. The institutional capacity to supervise and guide change using multiscale spatial analysis.
4. Quality project administration, management and monitoring.

This report focuses on the second objective of multiscale characterisation. The aim was to research methodologies and propose realistic solutions to the following sub-objective:

2.1 Compare and contrast a subset of socio-economic and biophysical variables from output 1 in terms of frequency distribution, redundancy, noise and extreme values at different scales of aggregation and disaggregation.



Introduction

The relationships among pattern, process and scale have long been recognised as a central issue in geography, economy, ecology, and many other sciences. Concerns about the detection of scale dependent phenomena and the prediction of system dynamics across scales have increased significantly in recent years.

Our inability to determine scale dependent effects results in conflicting interpretations of data and an inability to reliably extrapolate results. The failure to account for scale dependent changes in pattern and process has confused and confounded the synthesis of agro-economic data and the extrapolation of system effects from smallholdings to landscapes and beyond. The ability to detect changes in pattern and make predictions at more than one level of organisation would, therefore seem to be fundamental requirements of ecosystem science.

Achieving this goal requires identifying the physical and biological processes of interest and estimating the variables and parameters that affect these processes at different scales. The development of this capability will ultimately allow scale dependent predictions to be tested and hypotheses and relationships verified

Compounding the difficulty of developing general and reliable tools for detecting scale is the fact that agro-ecosystem data rarely produce a single scale that can be regarded as correct or optimal for measurement or prediction

Hierarchical models that take a specific data set, and locate distinct levels (D_1, D_2, \dots, D_N) which are analysed separately, present a potential solution. If these levels can be extracted from the data rather than imposed *a priori* then it is plausible to produce a set of explanatory models, one for each user-specified level. However it is unclear how to model the relationships that might exist between levels.

Theoretically, if we can determine the smooth (spatial trend) and rough (hot-spots or outliers) components of a data set across a very broad range of scales, then we can impose a hierarchical model onto the data set for hypothesis testing and future experiment design. Legendre (1993) has suggested analytical approaches that specifically address the issues of spatial dependence and the problems it presents for statistical testing of hypotheses. Legendre demonstrates that total variance of a table of target variables can be mapped into:

- Non-spatial environmental variation, (perhaps prices and taxes).
- Spatially structured environmental variation (perhaps temperature and soil pH).



- Spatial variation of the target variables that is not shared by the environmental variables (perhaps ethnicity and burning).
- Unexplained non-spatial variation.

This concept can be visualised in Figure 1. This could lead to an improved understanding of the spatial and environmental variance within a system, where a good fit for both the smooth and rough components would minimise the unexplained.

To meet the objective of this activity, this report documents several methods for recreating spatial data at a range of scales, and investigating the effects of scale upon them. The concept of localised indicators of spatial association (LISA) is introduced as a useful visual and analytical method for determining scale dependent behaviour. This method allows comparisons to be made across scales for:

- Determining localised scale effects.
- Comparing local statistics to 'global' whole map statistics.
- Spotting outliers and their effects on the distribution.
- Visualising and assessing scale dependence.
- Suggesting scales of further analysis.

Several traditional 'global' statistics are included, and their value assessed in combination with the 'local' descriptive statistics.

Exploring Scale Patterns in Spatial Data

Point Data Sets

The finest level of resolution currently available for the Honduran censuses is the "aldea" (village) or 4th level of administration. For the 1988 Population census there were 3730 aldeas (shown in figure 2) represented by points, which can be thought of as town or village centres. There are several options for designing a sampling procedure for such data sets:

1. Nearest Neighbour sampling
2. Buffer zone sampling
3. Intensive gliding window sampling

Nearest Neighbour Sampling.

The first method expands upon the traditional semivariogram, whereby each point in the distribution is visited in turn and a comparison made between it, and its nth nearest neighbour



(figure 3). However by allowing any number of nearest neighbours to be considered from between 1 and N we can build a cross scale profile for each point. Another distinction from the semivariogram is that almost any statistic can be generated from the neighbouring points (mean, variance, difference, join-count, correlation, etc.) and the distribution of the statistic across scale can then be explored.

Finally the concept of nearest neighbour, need not be based solely on Euclidean distance, using the methodology from activity 2.5 we can use travel time between points as a distance measure

Buffer Zone Sampling

Buffer zones can also be generated by determining the number of neighbours that fall within a defined radius of each point in the distribution (figure 3). This method is usually faster to compute than nearest neighbour sampling, although the same statistics can be generated. Again, this model can be adopted to use time bands in place of distance bands. Figure 4, clearly highlights the differences between a pure distance metric and a time based measurement.

Intensive Gliding Window Sampling.

A somewhat different approach is to overlay a grid system on the distribution and place a circular window on each grid intersection in turn (figure 5). Each circle overlaps its neighbours, ensuring a complete sampling of the distribution. After each intersection has been visited, the grid size and circle size is increased and the process repeated. Since this method is not point centred (i.e. each circle is not centred upon a point in the distribution), the statistics which can be generated are limited to simple momental statistics, cluster analysis and pattern spotting, however this is by far the most intensive and complete sampling method.

To speed up the calculation of these statistics, a distance matrix is computed to record the distance between every location, which in this case requires a 3730 x 3730 matrix to be generated. This matrix need only be generated once for each data set. Similarly from activity 2.5, a cost-distance matrix is generated containing travel times between each aldea.

Areal Data Sets

In many GIS there exists a basic implementation of topological relationships, such that it is easy to determine neighbouring areas by connectivity (figure 6) where the area under study is shaded green and it's 1st, 2nd, and 3rd level neighbourhoods are in gold, yellow, and light grey respectively. Using this basic concept it is possible to calculate a connectivity matrix, populated by 1's (adjacent areas) and 0's (non-adjacent areas) for the region. However as discussed in report 2.2, sometimes areal units such as municipios and departments do not make much



geographic sense for representing the underlying geographic distribution of the population, the landscape and any other socio-economic or agricultural data. The zone design algorithms documented for activity 2.2. allow us to generate zoning arrangements to suit our requirements, and from these we can generate cross scale profiles, using the same statistics as the nearest neighbour method for point data sets.

Image Data Sets

Intensive sampling is not new for remote sensing imagery. Convolution filters have been used in RS packages for many years. However there are many other statistics that can be incorporated into this technique. The raster data structure permits easy sampling of local values by using a square (or circular) 'floating' window (figure 7) to calculate statistics and create new raster imagery of the same resolution. Maintaining the original resolution is key to later analysing the 'stack' of cross scale images with other visualisation software or summary statistics. The statistics that can be generated are similar to the nearest neighbour method, but the list also includes some specific image analysis tools.

Local and global statistics

The use of LISA type statistics, allows the results of the cross scale analysis to be visualised as a map rather than a single statistic or table. Each location in the data set is considered in turn, and some statistic is generated based on its local neighbourhood. For each neighbourhood size, a map is generated and can be interpreted on its own, or in combination with the original data and other neighbourhood sizes. Finally, if N neighbourhood sizes were investigated, it is possible to create several summary images where the variance in the statistic across the N scales is visualised. These visualisations are presented in the examples at the end of this report.

Table 1 lists the statistics that have been used in this activity. For each spatial data type (point, areal and image) we considered 3 variable types (binary, qualitative and quantitative), and statistics were chosen accordingly. In addition to many common exploratory statistics, there are several that are not so well known, which will be expanded upon in the next section. Also many of these statistics can be calculated as either as local measures or for the entire map. Table 2 lists all the local statistics with their global counterparts.

Local measures of spatial association

Nomenclature

For local statistics, the point, areal unit or image pixel under analysis is termed P_0 , and its surrounding n local values are termed P_i where i runs from 0 to n . The total number of observations in the data set is N .

**Local Momental statistics.**

The output is assigned the 1st, 2nd, 3rd, or 4th momental value of its local spatial neighbourhood. For example the mean variance and standard deviations are calculated below (Equations 1 to 3). Similar calculations are performed for the skewness and kurtosis values.

$$P_{mean} = \frac{1}{n} \sum_{i=0}^n P_i \quad \text{Eq. 1.}$$

$$P_{var} = \frac{1}{n} \sum_{i=0}^n (P_i - P_{mean})^2 \quad \text{Eq. 2.}$$

$$P_s = \sqrt{\frac{1}{n} \sum_{i=0}^n (P_i - P_{mean})^2} \quad \text{Eq. 3.}$$

For qualitative values the mean is replaced by the modal (most frequent) value, and standard deviation can be replaced by a measure of entropy (Eq. 4), where $P(f_i)$ is the proportion of measurements classified as feature type i and m is the number of classes.

$$P_s = \frac{\sum_{i=0}^m P_{(f_i)} \cdot \log[P_{(f_i)}]}{-\log\left(\frac{1}{n}\right)} \quad \text{Eq. 4.}$$

By visualising the change in these statistics for varying neighbourhood sizes, the degree of sensitivity to scale for each location is revealed. Areas of high/low scale dependence can be delineated and the influence of outliers can be determined.

Local Spatial Lag

Spatial lag is a measure of the difference in value between the location under study and the sum of its neighbouring values (Eq. 5). It is a measure of regional similarity.



$$P_{lag} = P_0 - \left(\frac{1}{n} \sum_{i=1}^n P_i \right) \quad \text{Eq. 5.}$$

By repeating the statistic for various lags, the influence of outliers and extreme values is clearly visible.

Local Semivariance

Semivariance (Eq. 6) is a geostatistical technique used to estimate the spatial scales over which patterns are dependent. Unlike the spatial lag statistic, semivariance is a vectoral measure and thus has direction as well as value. Hence if the data is suspected to be anisotropic, then the semivariance can be measured in several directions (see Two Dimensional Moran Index)

$$P_g = \frac{1}{2n} \sum_{i=1}^n (P_i - P_0)^2 \quad \text{Eq. 6.}$$

This measure, along with the indices of autocorrelation, highlights locations of similarity and strong spatial structure within the data. By repeating the measure for various scales, the changes in these patterns indicate the scales at which these patterns occur, and the implications that autocorrelation will have on multivariate analysis (regression for example) at these scales.

Local Moran index of spatial autocorrelation

The degree to which close neighbours share similar properties is most commonly measured by the following two indices of spatial autocorrelation, The Moran index (Eq. 7) and the Geary index (Eq. 8)

$$P_I = \frac{1}{n \cdot P_{\text{var}}} \sum_{i=1}^n (P_0 - P_{\text{mean}})(P_i - P_{\text{mean}}) \quad \text{Eq. 7.}$$

Local Geary index of spatial autocorrelation

The output is assigned the Geary index value of its local spatial neighbourhood.



$$P_c = \frac{\sum_{i=1}^n (n-1) \cdot (P_i - P_0)}{\sum_{i=1}^n n \cdot (P_0 - P_{mean})} \quad \text{Eq. 8.}$$

Local Getis and Ord G statistic

A recently developed statistic to measure the spatial clustering within a dataset is given below (Eq 9).

$$P_{G^*} = \frac{\left(\sum_{i=1}^n P_i \right) - n \cdot P_{mean}}{\sqrt{\frac{P_s \cdot n \cdot (N - n)}{N - 1}}} \quad \text{Eq. 9.}$$

Diversity Indices

There are a plethora of diversity indices that have been proposed in the biological and ecological literature. Whilst the academic debate on what *diversity* actually means rages on, it is useful to apply a few of the most common indices to geographic data. Traditionally such indices have been calculated either on a whole map basis, or by quadrat analysis, where a region is arbitrarily divided up into large quadrants and the index calculated per quadrant. Activity 2.2 deals with the implications of such arbitrary division (aggregation), and we have again adapted the methodology such that the indices can be calculated for local neighbourhoods. Two indices are described here, the Simpson and Shannon Weaver heterogeneity indices, although any one of a dozen familiar diversity indices could have been used.

The Simpson index (Equation 10) measures the probability that two records selected at random from a sample will have the same class, where i is the number of classes and n_i is the number of records in class i .

$$P_L = \frac{\sum_{i=0}^n [n_i(n_i - 1)]}{[N(N - 1)]} \quad \text{Eq. 10.}$$



The Shannon Weaver index (Equation 11) represents the amount of uncertainty that exists regarding the class of a randomly selected record from the data set.

$$P_H' = \frac{-\sum_{i=0}^n (n_i / N)}{\log(n_i / N)} \quad \text{Eq. 11.}$$

Lacunarity

This is a measure based on fractal analysis and first proposed by Mandelbrot in 1983. Lacunarity can be thought as the distribution of holes or gaps in the data. Although it is a whole map statistic, the technique analyses the data set at a range of scales and the output is a graph of lacunarity versus scale for the entire region, indicating the scale and form of patterns within the data. Using the gliding window technique, at each window location the number of occurrences s of a value are counted. Once all areas have been sampled for window size r , the counts are summarised as a frequency distribution (Eq 12).

$$n(s, r).N(r) \quad \text{Eq. 12.}$$

The frequency distribution is converted to a probability distribution $q(s,r)$ by dividing each frequency by the number of gliding windows that were used. The first and second moments are calculated as follows (Eq's 13 and 14), and lacunarity is estimated by the ratio in Eq 15.

$$Z_{(1)} = \sum_{s=1}^r sq \quad \text{Eq. 13.}$$

$$Z_{(2)} = \sum_{s=1}^r s^2 q \quad \text{Eq. 14.}$$

$$L = \frac{(Z_{(2)} - Z_{(1)})^2}{Z_{(1)}} \quad \text{Eq. 15.}$$

The window size is increased and the process is repeated until a window size threshold is met. The Lacunarity per scale can be normalised by dividing the L value for each scale by the L value for a window size of 1. These values are then plotted versus the window size and compared with



plots for random and other controlled data sets. The plot reveals the degree of similarity across scales, and changes of slope in the graph indicate scales where different patterns in the data are revealed, although it says nothing about their geographic location in the region.

Join-count statistic

Join-count is a measure of clustering for qualitative data. For each location the nearest neighbour is determined and their attributes compared and the type of join between them is recorded (like to like, or like to unlike). The number of each possible type of join is recorded for each pair of neighbours in the data set. This observed value is compared to the number of expected join types in a random distribution. An excess of any type of join suggests the presence of clustering.

We have expanded upon the test by considering larger neighbourhoods (2nd nearest neighbour and so forth) to indicate the scale of the clustering, and again the statistic can be plotted against neighbourhood size. The main criticism of this method is the test versus randomness, which is regarded as a trivial and meaningless test for geographic data since they are nearly always non-random. However by comparing the join count for different neighbourhoods, the degree of clustering relative to neighbourhood size can reveal scale dependence within the data.

Coefficient of correlation

A simple measure of the correlation between two variables in the local neighbourhood

$$\mathbf{r}_{x,y} = \frac{Cov(X,Y)}{\mathbf{s}_x \cdot \mathbf{s}_y} \quad \text{Eq. 16.}$$

Where

$$-1 \leq \mathbf{r}_{x,y} \leq 1$$

and

$$Cov(X,Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mathbf{m}_x)(y_i - \mathbf{m}_y) \quad \text{Eq. 17.}$$

This statistic, is useful for highlighting localities of strong positive or negative correlation, which represent spatial structure within the region, especially when these localities possess wildly



different correlation values from the global (whole map) correlation. It is these “outliers” and hot spots that invalidate the General Linear Model assumption of non-correlated variables.

Local morphometric feature extraction (adapted from Wood 96)

The output is assigned 1 of 6 possible morphometric features based on its local spatial neighbourhood. The most widely used set of characteristics, is the subdivision of all points on a surface into one of pits, peaks, channels, ridges, passes and planes. The names of these features suggest a geomorphological interpretation, but they may be unambiguously described in terms of rates of change of three orthogonal components (see Table 3). Note that the components x and y are not necessarily parallel to the axes of the lattice, but are in the direction of maximum and minimum profile convexity.

At each scale, the resulting surface is divided up into the 6 possible classes. By summarising across scales the most common classification (modal value) can be determined along with the variance in classification. These summary images describe the areas of high/low scale dependence in the surface as well as a degree of certainty as to whether a region is indeed (for example) a valley.

Clustering (Geographical Analysis Machine, Openshaw 81)

GAM works by examining a large number of circles of varying sizes that completely cover the region of interest. This is important to ensure that no locations are missed out. The circles overlap to a large degree to allow for edge effects and to provide a degree of sensitivity analysis. Data are retrieved for each circle and some statistical assessment made of whether or not the incidence rate in each circle is unusually high (or small).

The GAM algorithm is based on a gliding window and involves the following steps:

Step 1. Read in X,Y data for population at risk and a variable of interest from a GIS.

Step 2 Identify the rectangle containing the data, identify starting circle radius, and degree of overlap.

Step 3 Generate a grid covering this rectangle so that circles of current radius overlap by the desired amount.

Step 4 For each grid-intersection generate a circle of radius r .

Step 5 Retrieve two counts for the population at risk and the variable of interest.

Step 6 Apply some 'significance' test procedure.

Step 7 Keep the result if significant.



- Step 8 Repeat Steps 5 to 7 until all circles have been processed.
- Step 9 Increase circle radius and return to Step 3 else go to Step 10.
- Step 10 Create smoothed density surface of excess incidence for the significant circles using a kernel smoothing procedure and aggregating the results for all circles.
- Step 11 Map this surface.

The surface smoothing in step 10 involves the following:

- Step 1. Initialize a raster map grid.
- Step 2. Read circle_X, circle_Y, circle_RADIUS.
- Step 3. Compute 'excess' (observed-expected cases).
- Step 4. Apply Epanechnikov kernel to create density surface using circle_RADIUS as bandwidth.
- Step 5. Surface summed over raster over map grip.
- Step 6. Repeat for all 'significant' circles and map surface.
- Step 7. Apply cluster peak detector or eye-ball surface looking for large peaks.

Two dimensional Moran index (adapted from Wood 96)

A distinctly different approach to measuring scale dependency and anisotropy in a surface is the construction of a two dimensional variogram. The variogram is a representation of all possible lag distances and all possible directions of lag. This information is visualised by mapping each result back onto the image where the lag distance and magnitude are indicated by their location relative to the centre of the image. These images are initially very hard to interpret as (unlike all the previous measures) they do not directly represent the original data, and require a huge amount of processing time. However since the 2D plots strikingly reveal the spatial structure within the data set, and were considered useful for this purpose as well as for activity 2.9

Standardising local values

To standardise a data set, i.e. transform it so that it has a mean value of 0 and a standard deviation of 1, we apply the following equation (18) to all values of P .

$$P'_i = \frac{P_i - P_{global_mean}}{P_{global_s}} \quad \text{Eq. 18.}$$



Example 1. A cross scale comparison of topography.

This example examines a Digital Elevation Model, possibly one of the most common GIS coverages, and also one of the most important for deriving, slope estimates (surface runoff), aspect (agricultural productivity), watershed boundaries and other information vital to ecoregional studies.

This dataset is taken from our test site in Yorito. The original dataset was derived from several ortho-photos, from which a DTM at 10m resolution was produced. For this example a 5000m x 5000m section has been extracted (Figure 8).

Four measures will be applied to the original 10m resolution DTM

1. Elevation
2. Slope
3. Feature classification
4. Two dimensional index of spatial autocorrelation

Each measure will be applied at scales of 100m, 200m, 300m etc. up to 1000m, creating 10 cross-scale datasets for each measure (except the index of correlation). This example will show

- An analysis of elevation and how its sensitivity to scale varies over the region.
- How a derived data set (slope) has different characteristics across the same scales.
- A feature analysis, which will highlight possible classification errors across scale.
- A 2 dimensional Moran index of spatial autocorrelation, to highlight the autocorrelation of regions topography.

For visualisation purposes, we will use two representations of the region. One is a traditional 2D view of the imagery; the other is a pseudo 3D view of the region, created by draping the various results onto a 3D representation of the elevation data (Figure 9). The 3D image has been rotated somewhat to provide a more revealing view of the landscape.

In figure 10a, the images show the DTM at 4 different scales (100m, 300m, 700m, 1000m) although each image maintains its resolution of 10m. Below each image, the circle indicates the size of the smoothing filter. The small cross indicates where a sample has been taken to measure the change across scale at that location (figure 10b). The variance of the data across all scales is shown in figure 10c. Areas that are scale dependent (red) and which are not, are clearly visible. Essentially we are highlighting the amount of information that has been "lost" as



we generalise the data. The outliers (peaks and valleys) in the data that are greatly effected by scale change, however the effects are not consistent over the region.

Next, we perform the same mean filter to the first derivative of elevation; slope. From figure 11b, it is immediately clear that scale effects on slope calculation are values are entirely different to scale effects on the original elevation data. In the previous example, the scale effects were very predictable, high or low, geographically isolated areas were subject to the greatest loss. Following this reasoning we should expect small steeply sloping areas to suffer the greatest changes across scale, but clearly there are huge changes in relatively flat valley areas (referring to the valleys closest to the point of observation). We can attribute this to the fact that at the larger scale, the filter has incorporated the valley walls into the slope calculation of the valley floors, rapidly changing their values over a small range of scales. Again this effect is not consistent over the region. Figure 11a shows the relative change of slope for 2 samples, the green is from a low-lying valley and orange is from a steep slope.

Using the feature extraction filter (Wood 96), we see that even an intuitive concept such as "what is a peak?" or "what is a ridge" are scale dependent. Figure 12a shows the change in surface features across 4 scales and 12b marks the change in feature classification across-scale for a given location. In figure 12c it is evident that feature classification has a different pattern of scale dependence than either elevation or slope. We might imagine standing on a boulder (peak) situated in the middle of a valley (channel) which is part of a larger plane (flat), and we start to see that even simple and intuitive classifications of a surface reveals the complex interactions between the scale of observation and the result of the observation.

Here we take a different approach to measuring scale effects. Figure 12 shows a two dimensional variogram of Morans I index of autocorrelation. Such information is usually represented as a graph of variation against lag distance, but here all directions and lags have been calculated with the results mapped in such a way that distance from the centre represents the lag measured and direction from the centre is the direction in which the correlation was measured. The concentration of high values (red) along the leading diagonal indicate strong positive correlation, over a very small distance, and that the correlation is asymmetric.

Whilst the 2D variogram gives a strong visual impression of the spatial relationship between neighbouring values, it is still very difficult to interpret due to its rather abstract nature (abstract both in the sense of the concept of autocorrelation, and the nature of 2D mapping).



Example 2. A cross scale comparison of productivity.

Using data from the agricultural census of 1993, Barbier et al (1999) estimated the average production per worker per village as a proxy for labour productivity.

The yields for each crop and livestock production were combined with national commodity price data, to determine household production. The total agricultural production per aldea has been divided by the number of permanent workers. This ratio represents labour productivity.

This example will compare 'traditional' results from the previous research; relationships between productivity and other variables (natural resources, socio-economic and farming system variables), with cross scale results. This will highlight spatial 'pockets' where the relationships contradict the global model and also reveals regions of high spatial autocorrelation that would necessitate a more regional approach to further modelling.

The variables are:

- Number of dry months (natural resource variable).
- Population density per aldea (socio-economic variable).
- Percentage of farms using improved varieties per aldea (farming systems variable).

We will compare local and global correlation and visualise the range of local correlation that can occur across scale. These local variations in the relationships will be compared to administrative boundaries to 'eye-ball' regions where the Municipios do or do not adequately represent the underlying patterns in the data.

Figure 14 shows the agricultural productivity per aldea (red is low, green is high), and figure 15 represents the three variables we will correlate with productivity; rainfall, population density and improved varieties respectively.

The first stage is to run the correlation analysis at a huge range of scales, to give an indication of the rate of change of the measurement with scale. Figure 16 shows three graphs, one for each variable, comparing the range of correlation values that were generated for each scale. These values were computed by measuring the correlation at 3730 locations (the number of aldeas in the dataset) at each scale, and computing the maximum, minimum and average values. Each variable clearly responds differently to changes of scale although the general trend is the same; great variation at fine scales with coarser scales tending towards the global result. The global correlation result is included with each graph.



Taking these results, and presenting the maps for each scale would be very laborious, so for this example, we have chosen one scale (24 nearest neighbours) which roughly corresponds to the number of aldeas in the majority of municipios. The average value is 12, biased by a substantial number of very sparsely populated municipios.

For each variable we have mapped the correlation coefficient by aldea (points), and for easier visualisation with municipio boundaries, by an interpolated surface. Figures 17 and 18 are for population density, 19 and 20 for improved varieties and 21 and 22 for monthly rainfall. There are several points to note from these images.

The variations from the global correlation value are not random in their location, and there are strong spatial patterns in each correlation dataset, that would not be visible in any other analysis. These patterns relate to local processes, that could improve our understanding of the dynamics within the agricultural system.

Any further multivariate analysis is likely to miss these variations and will probably not generate a good fit to the dataset. A localised regression method, based on a similar neighbourhood function is presented in report 2.9, and this same dataset is used to investigate the significance of spatial structure.

The patterns generated do not correspond in any regular way to the municipio boundaries. In some instances we can see that there are very good fits between the boundary and the underlying data (for example the highlighted areas in figures 18, 20 and large areas of figure 22), but in most cases there are not. With the exception of monthly rainfall (since monthly rainfall patterns change little over such small regions as municipios) aggregation to municipal boundaries will distort and maybe destroy these relationships. In this example, they are not suitable for measuring even the socio-economic variable.

Discussion

Exploratory spatial data analysis

It is clear that statistics, tables, graphs and traditional exploratory analysis are not enough to describe let alone explain the complex, scale-dependent relationships that exist in geographic data. For example, figure 23 shows histograms of slope values, taken at 4 different scales, with each column representing the number of pixels found in each slope class. There are obvious changes visible across all 4, but even if the analyst were armed with a barrage of other statistics and charts, they would not be able to tell you *where* the information was lost. If we are to assume



that geography does matter, then *where*, is exactly the kind of information that we need to know. Performing any kind of spatial analysis, from simple overlay through regression to complex spatial interaction models, without any information on scale effects and without knowing the limitations of the data is dangerous. Given that perfect and complete information never exists, we must always be expected to manage with the data at hand, in which case it is vital to understand what can and cannot be reasonably expected from the analysis, and what degree of confidence we can have when:

Dataset X

is analysed at

scale Y

for

purpose Z.

Such 'error checking' or 'fitness for use' metrics, are almost unheard of in GIS systems, for two reasons.

The scale issue is not well understood nor thoroughly researched. When this project was initially funded in 1996, there were few if any recognised texts or papers dealing with the issues of complexity and scale, and those that did, were not spatially explicit. Over the last 4 years, books such as 'Ecological Scale' and 'Scale in Remote Sensing and GIS' have appeared, but even in books such as these, the predominant theme seems to involve defining scale and talking about scale, rather than actually doing anything about scale.

GIS vendors develop commercial GIS rather than scientific GIS, and the commercial market has little need for spatial analysis. It can be argued that the further development of commercial GIS threatens their viability as an analytical tool. But it is also clear that many research projects have their own self written spatial analysis toolkits which contain simple visual and data handling mechanisms, with many novel and powerful analysis methods incorporated. How many of these methods will be incorporated into future GIS remains to be seen.



It is disturbing to think that as we integrate ever-increasing amounts of digital data into (ever-increasing amounts of) available desktop GIS, that there are no safeguards or safety nets to prevent the (ever-increasing amounts of) uneducated users from producing practically meaningless results. Whilst moves are afoot to improve the quality of metadata (data about data); thus informing the user of the errors within the dataset, there are no scale aware models (to prevent for example, incompatible overlays) on the horizon.

There are a range of possible statistics and visualisations that can be applied, of which we are surely only scratching the surface in this project. They allow all types of spatial data (binary, qualitative, quantitative, point, areal and image) to be compared across scales, highlighting extreme values, distribution changes and trends. Some of the methods are unusual and perhaps difficult to interpret, they are certainly not traditional nor can they be found in many statistical or data analysis books or software packages.

Modelling with geographic data

Geographic data possesses special characteristics (spatial dependence and spatial heterogeneity amongst others) that have traditionally been ignored in modelling work. There is a great danger that many existing models can be applied to spatial data using unrealistic assumptions thus creating misleading and potentially dangerous results. In order to ameliorate these problems, the GIS data revolution, which was much heralded in the 1980's, must now be accompanied by an appropriate methods revolution. Given that scale effects make any single or multivariate analysis of aggregated spatial data highly suspect, one way of assessing the importance of scale effects is to document the effects by reporting results at different levels of data manipulation. However, great care must be taken to ensure that these levels are context specific and not imposed on the data *a-priori*. Such context specific reporting can be made easier by the increased use of techniques such as those espoused in this paper.

Here we have argued in favour of a series of techniques and tools that allow spatial data sets to be constructed and de-constructed in a generalised yet context sensitive manner. We state that the outputs from such techniques can be explored and described through various user-defined levels, thus revealing spatial patterns and processes that are arguably more useful than raw data or standard representations. Potentially, complex hypotheses and models can be developed based on the improved understanding that such mapping techniques provide. Additionally, the opportunity to re-express the data at different levels - levels appropriate to different decision-makers - enables conflicts to be rapidly highlighted and the effects of a decision at one level to be visualised at other levels of organisation.



The functionality of GIS can be classified into three levels:

The use of GIS to do simple things that we have always done.

The use of GIS to do complex things that we seldom or never do.

The use of GIS to do new things that revolutionise our thinking and create new hypotheses.

(Arnold and Appelbaum 1996)

Here we have described a range of methodologies ('level 2' functionality) that can be applied to the huge, multivariate, and very complex databases that are fast becoming the de-facto standard in many projects that have a geographical nature. GIS have often languished as the role of data mapper ('level 1') in many interdisciplinary projects, and while such use is often valuable and necessary, we must not overlook the potential for GIS to play a far more interesting role that can lead to new possibilities and the heady heights of 'level 3' type contributions.

Scale effects

Scaling up is generally understood to refer to the translation of information from smaller to larger scales, but discussions about scaling are often more accurately characterised as translations between levels. The increase in spatial scale – from square meters through hectares to square kilometres - results in new interactions and relationships as changes in organisation are encountered. A change in scale almost always necessitates consideration of new levels of organisation (O'Neill and King 1998).

Scale also refers to the scale of observation, if we do not observe a process at the correct scale, i.e. the scale at which the process is occurring, then we are unlikely to gain insight into the process and its contribution to the system as a whole. Observing our data across a continuum of scales is currently one of the few ways of spotting these processes. Since processes and patterns change with scale, aggregating data or averaging results does not seem to be the answer to understanding or extrapolating results to other scales. The whole is almost never equal to the sum of its parts, and insensitive aggregation and averaging to arbitrary scales is a sure-fire way to destroy just those patterns we have been looking for. Here we have defined the 'correct' scale is the one that permits the 'best' prediction of a process based on a statistical or mechanistic model of that process. But how can one select the scale of the study to maximise predictability?

Increase in predictability with scale is not monotonic and local *maxima* coincide with intrinsic scales of organisation. Based on the above description it follows that the correct sampling scale



should coincide with the intrinsic scales determined by the process structure (Goodwin and Fahrig 1998). However this approach requires knowledge of the process structure which, in the data-rich, theory-poor world of GIS is often lacking. The approach throughout this research has been one of:

I initially know nothing, and therefore I will search everywhere for everything at all possible scales to determine the 'correct' scale.

We have seen that each process exhibits patterns that occur at scales specific to that process. Gradually change the scale of analysis and the pattern will also change (predictably, dramatically or maybe very little at all) and there will come a point in the scale continuum where the pattern change is so dramatic that we have to rethink our assumptions of the processes that are occurring.

We argue that these breaks in the continuum are vital for understanding and accommodating scale effects. They define the limits of our chosen modelling techniques, and they indicate where assumptions and hypotheses about the system must be changed for another set. These breaks suggest a change in the level of organisation. Unknowingly crossing this break can render a model inapplicable, as the observation set has changed to a new set of organisation.

Conclusions

Outputs compared with a restatement of the objective

Compare and contrast a subset of socio-economic and biophysical variables from output 1 in terms of frequency distribution, redundancy, noise and extreme values at different scales of aggregation and disaggregation.

Methodologies (incorporating new in-house software for analysis and visualisation) for performing the task stated in the objective have been investigated and presented in this report.

The concept of a local neighbourhood has been introduced as basis of measuring spatial data across scales. This approach has been extended from previous research by the adoption of travel time (in place of distance) to define local neighbourhood.

Statistics and measurements from diverse fields such remote sensing, geostatistics, econometrics, computer visualisation, ecology and biology have been considered as viable approaches for assessing the scale dependent features in a range of datasets. These have been



augmented by the inclusion of the Geographical Analysis Machine (GAM) as a powerful tool for assessing and visualising spatial clustering of key variables, and the new feature classification method for digital elevation data.

All of these statistics have been recreated as local statistics in place of traditional global or whole map measures. Some of these are completely new, such as local diversity analysis.

Key eco-regional variables such as topography and agricultural income have been presented as examples from the *shadow* database, which contains many other datasets with similar cross-scale representations. Spatial datasets have been recreated at various scales to analyse the changes as stated in the objective.

It is clear that these methods allow the data to be analysed and visualised in new (or novel) ways, and that these representations are of great use for determining scale effects, for detecting spatial pattern and for suggesting further avenues of research.

Final remarks

The use of local statistics allows us to visualise the effect of scale and then act up on that knowledge. If we analyse a dataset at a given scale, then we must be confident that the data is representative of the phenomenon we wish to study. By analysing the data set at a range of scales (scales finer and coarser than the study scale), and examining the variability interactively or with summary 'scale dependence' images, it is possible to attach a degree of confidence to the datasets ability to represent the phenomenon at that particular scale. Which measure used to determine this is entirely dependent on the purpose of the analysis. As we have seen in the DTM analysis, slope and landscape metrics give completely different surfaces of scale dependence.

It is also possible to use local statistics to extract potential new levels of analysis from the data. As will be highlighted in Activity 2.2, we are often bound by the limitations of the data, software or methodology, to analyse data at inappropriate 'fixed' levels. Local statistics and high-resolution data allow us to see just how these fixed levels relate to the underlying fine data, and to determine if there are areas where they do fit the data and areas where they don't.

Spatial structure, scale effects, and the complexity that exists within geographic databases are not easy issues to incorporate into analytical methods. Data of higher resolution and greater quantities do not necessarily translate into better data or more information. It often leads to the opposite, as more complex patterns and structures are revealed.



Data aggregation or generalisation must be a controlled and understood process. Local statistics can highlight areas where unacceptable amounts of information will be lost if function X is applied to dataset Y in area Z . It is unacceptable to assume that scale effects are not important or do not exist within a dataset.

There are important theoretical and practical reasons for the further study of scale effects, spatial phenomena, and changes in pattern across scale. If we wish to predict the consequences of both natural and human interactions within a system, then understanding scale dependencies and levels of organisation would seem to be a precondition for this kind of approach. Infant mortality rates, for example, may be related to water quality at a local level, but accessibility to health care might also contribute albeit at a higher regional level. Spatial mapping of these phenomena might be able to provide insight to the structure of these organisational levels and can help in the future conception and design of experiments.

The identification of appropriate scales for analysis and prediction is an interesting problem. Although the factors producing scale dependent patterns may not be clearly understood, we have been able to create accurate and reliable descriptions of scale dependent patterns and processes to design data sampling procedures and test the accuracy and reliability of methods of prediction. There is clearly a some way to go before scale effects can be fully understood and accommodated, but this research has aimed to be the 'next step' in that process.



Appendix A: Software

There are three main software tools in this activity.

Cross scale tool for image and raster data (CIAT)

The raster tool is described in a separate document, it operates under UNIX and Windows and can be integrated into ArcView 3 as an extension or can be run from ArcInfo through AML scripts, or as a stand-alone program. It reads and writes standard ESRI raster data formats. This software will be available from CIAT's webpage <http://www.ciat.cgiar.org> by February 2001.

The program requires a command file, with instructions for the analysis. A typical file would look like this

```
INPUT FILE:      infile
OUTPUT FILE:     outfile
NEIGHBOURHOOD:  circle or square
SMALLEST SIZE:  number (window size in pixels)
LARGEST SIZE:   number (window size in pixels)
INCREMENT:      number (increment in pixels)
MEASURE:        name (any of the measures)
NORMALISE:      yes or no
```

The program accepts ASCII or binary floating point grids (from ArcInfo or ArcView) as input, these can be created in ArcInfo as follows.

```
in_ascii.txt = GRIDASCII ( in_grid , int/float )
in_float.bin = GRIDFLOAT ( in_grid )
```

The output is a series of ASCII or floating point files (depending on the input type) and there is one file for each scale measured. These files can be converted to ArcInfo grid format as follows.

```
out_grid1 = ASCIIGRID ( grid.txt , int/float )
out_grid2 = FLOATGRID ( grid.bin )
```



Cross scale tool for point and areal data (CIAT)

The point tool, is more a collection of separate programs for reading a writing Shapefiles and Database files, and for visualising point data. It has not been integrated into ArcView, simply because the methodology is far more interactive and exploratory, and this kind of analysis is very awkward to implement in ArcView. Figures 24, 25, 26 and 27 show screenshots of several point data analyses. The software runs under UNIX (Tcl/Tk scripting language) only, and efforts are underway to port the software to Windows for distribution along with the raster tools. As shown in the screenshots, the interfaces are highly interactive, allowing the user to interrogate and select data sets with ease.

The program requires a command file, with instructions for the analysis. A typical file would look like this

```
INPUT FILE:      infile
OUTPUT FILE:     outfile
DATA COLUMN1:   name (database file column)
DATA COLUMN2:   name (only for correlation measure)
NEIGHBOURHOOD:  buffer OR nearest
DISTANCE METRIC: time OR distance
SMALLEST SIZE:  number (buffer size or neighbours)
LARGEST SIZE:   number (buffer size or neighbours)
INCREMENT:      number (in units of distance)
MEASURE:        name any of the measures
NORMALISE:      yes or no
```

The program accepts ESRI Shapefiles as the input format. It reads the X,Y coordinates from the *.shp* file and the unique location ID, and data from the *.dbf* file.

If necessary the program will create a distance matrix for the data. Once created, the program will search for this matrix in future analyses.

The program outputs a Shapefile (with *.shp*, *.shx* and *.dbf* components), and a comma delimited text file. The output datasets contain the ID, (X,Y) co-ordinates, original data column(s), one column of data for each scale of analysis, as well as mean, and standard deviation values across the measured scales.



Geographical Analysis Machine (Stan Openshaw, University of Leeds)

Homepage.

<http://www.ccg.leeds.ac.uk/smart/intro.html>

<http://www.ccg.leeds.ac.uk/smart/gam/gam.html>

Software.

<http://www.ccg.leeds.ac.uk/smart/gam/gamin.html>

Online References

Openshaw, S., Turton, I., Macgill, J. and Davy, J. Putting the Geographical Analysis Machine on the Internet (**word doc**) (**postscript**) Paper presented at GISRUk'98, Edinburgh.



Appendix B : Point data analysis figures

Figure 24. Local Indicators of Spatial Association(1).

The interface has 6 distinct parts

1. Locational data. The 6 boxes contain information on the location we are studying (it's code, and X/Y location), the value of the original variable at that location and its value when represented at a different scale, and a scale bar
2. Geographic data. The map represents the aldeas in the department of El Paraiso. The colours indicate the value at each location for given scale. The yellow spot is the location we are currently investigating. The colours of the aldeas change as we change scale, reflecting their change in value under different scale s of analysis.
3. Fractal dimension chart. This shows the change in variable across scale for the chosen location. The data has been transformed to mimic a fractal measure. Changes in the slope indicate changes of pattern, which can be related to a process. The chart changes as we move from location to location
4. Moran Scatterplot. This plots the value of each point in the map against the Moran index of spatial autocorrelation. The yellow point indicates the relative location of the se point in the map. The chart changes as we change scale.
5. Graphs of cross scale measure, for each location for each scale we can measure (from right to left) Spatial Lag, Moran Index, Geary Index, Semivariance, and Getis and Ord Statistic. This is done for each point at each scale. Each graph charts the value of its respective statistic for the location a under study, across all scales.
6. Scale bar. This sliding bar changes to scale of the analysis, in this window we are in the lowest scale, 1.

Figure 25. Local Indicators of Spatial Association(2).

This shows the same interface at a far coarser scale, scale 8.

**Figure 26. Local Indicators of Spatial Association(3).**

1. Locational data. The 5 boxes contain information on the location we are studying (it's code, and X/Y location), the value of the original variable at that location and its value when represented at a different scale.
2. Scale bar. This sliding bar changes to scale of the analysis, in this window we are in the lowest scale, 1.
3. Boxplot. This window shows the range of values within the distribution at each scale, the colours indicate deviation from the mean and the yellow line refers to the location under study.
4. Histogram of the cross scale measure. Here we are measuring variance across scale. The graph represents the variance at the location under study across all scales, the colour of each bar indicates that values relative location in the distribution. The yellow block indicates the current scale of analysis.
5. The map. This analysis is of the aldeas within the department of Olancho. Again the colours of the points represent their value at the chosen scale and the yellow point is the location under study. The concentric grey scale circles indicate the scales that have been analysed, giving an indication of the neighbourhood of the point.
6. Movable map elements. The shaded circle is the neighbourhood of the current scale of analysis. It, along with the scale bar and North Arrow can be moved over the map.
7. Map selector. There are 18 Departments in Honduras, these buttons allow you to switch between each one.

Figure 27. Local Indicators of Spatial Association(4).

The same data set at a coarser scale and with more spatial neighbours indicated by the grey scale circles.