ELSEVIER

Agriculture, Ecosystems and Environment 85 (2001) 107–131

Agriculture Ecosystems & Environment

www.elsevier.com/locate/agee

Analysing data across geographic scales in Honduras: detecting levels of organisation within systems

Andrew Nelson*

Centro Internacional de Agricultura Tropical (CIAT), A.A. 6713, Cali, Colombia
School of Geography, University of Leeds, Leeds LS2 9JT, UK

Abstract

There is a growing disparity between the disciplines that aim to understand and explain the phenomena and relationships that characterise the environment. Social physics leads towards finding universal relationships in data on the basis of well-understood and statistically robust methodologies. Recent trends in quantitative geography have focused on empirical, visual, exploratory and local methods to reveal patterns that can subsequently form part of a model specification to design experiments and test hypotheses. The information technology data explosion during the last two decades has increased differences between the social physics and the quantitative geographical paradigms.

The view expressed here is that problems in spatial analysis and modelling have been deliberately ignored to date or treated as a special case of aspatial modelling. To continue such a trend would imply that location is not relevant for spatial analysis.

This paper reviews misconceptions surrounding the use of spatial data and describes a set of spatial analysis methodologies to permit scale-sensitive and location specific analyses of socio-economic and biophysical data from a range of sources using examples that demonstrate the need for more geographical approaches to inherent geographical problems.

The examples illustrate that scale and aggregation artefacts can be observed, accounted for and even used to advantage, new relevant areal units can be designed within GIS environments, and that system boundaries of complex agro-ecosystems can be automatically derived from the combination of a spatial regression model and a neural classifier.

Keywords: Scale effects; Spatial processes; GIS; MAUP; Exploratory data analysis

1. Introduction

1.1. The problem

There is an increasing awareness that past assumptions relating to scale of analysis and extrapolation or results may have been flawed (O'Neill and King, 1998, provide an excellent review). This awareness is not only confined to geography, but is also evident in other spatial sciences such as ecology, anthropology and economics. It is possible that some hierarchical models, that dictate the use of one modelling method over another, do not represent the system under study with a great deal of accuracy (Hobbs, 1998). Indeed, when a systems view is taken to its logical extreme there are no parts at all and what is called a part is merely a pattern in an inseparable web of relationships.

A common theme in Ecological Scale (Peterson and Parker, 1998) is that it is becoming more and more evident that levels of organisation are not scalar but rather definitional — in that they come solely from the observer — and at each user defined level, phenomena exhibit properties that do not exist at

* Present address: The World Bank, Development Research Group, Room MC2-560, 1818 H Street, N.W., Washington, DC 20433, USA. Tel.: +1-202-4736400; fax: +1-202-5223230.
E-mail address: anelson@worldbank.org (A. Nelson).

0167-8809/01/$ – see front matter
PII: S0167-8809(01)00191-8

other levels. This concept of emerging properties marks a dramatic shift in thinking from simplification and reduction (modelling discrete entities) to an acceptance that complexity in science is the norm (modelling patterns and relationships).

One of the most pressing methodological issues deals with cross-scale analysis of data through farm, village, regional and country levels. While the actual models used will depend upon data availability, some general rules and methodologies can be developed. An outstanding issue relates to incorporating space or location as an explanatory variable, equivalent perhaps to more traditional variables. Efforts using an approach based on multivariate statistics and factor analyses suggest that dynamic spatial modelling across geographic scale can be incorporated into more traditional dynamic temporal analysis (Veldkamp and Fresco, 1996). Given that structural heterogeneity is widely accepted as the norm within ecosystems, applications of system theory will require explicit identification of the role of spatial structure of socio-economic as well as traditional biophysical factors in examining future impacts of alternative interventions.

Theoretically, if the smooth (spatial trend) and rough (hot-spots or outliers) components of a data set can be determined across a very broad range of scales, then a hierarchical model can be imposed onto the data set for hypothesis testing and future experiment design (Haining, 1990). Furthermore, if these hierarchical levels can be related to management levels, a powerful methodology can be developed for relating decision-making and policy design to potential impact and for improving predictive analyses (Hobbs, 1998). Unfortunately the levels of management (and at which data are most often available) and the levels at which a system appears to be self-organising are rarely congruent, and efforts must be made to create suitable analytical techniques to extract, rather than impose, levels of analysis and to improve the relationships between the two (Fig. 1).

1.2. Project objectives

The CIAT project 'Methodologies for integrating data across geographic scales in a data rich environment: examples from Honduras' (hereafter referred to as the project) developed and documented principles and procedures for building a scale consistent database and for performing multiscale characterisations of agro-ecosystems.

This paper will focus on the multiscale characterisation of Honduran agro-ecosystems for targeting problems, priority areas and beneficiaries (Fig. 2). This output was achieved through:

- Comparing and contrasting a subset of socio-economic and biophysical variables in terms of frequency distribution, redundancy, noise and extreme values at different scales of aggregation and disaggregation.
- Performing data reduction and low-dimensional representation of multiple scales for site sampling and hypothesis generation.
- Testing the significance of spatial structure in hillsides agro-ecosystem characterisation.

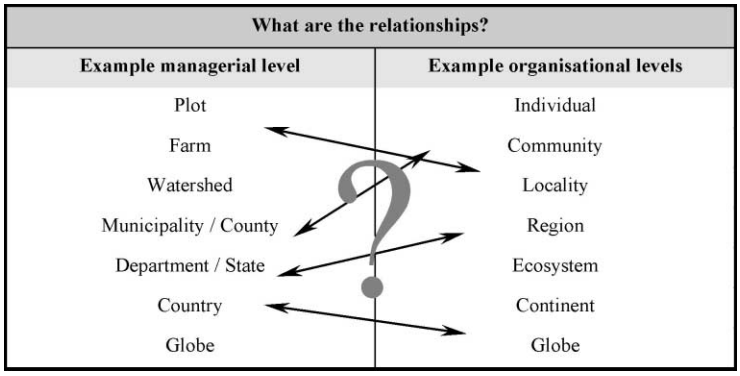


Fig. 1. The common non-relationship between units of measurement and units of management.

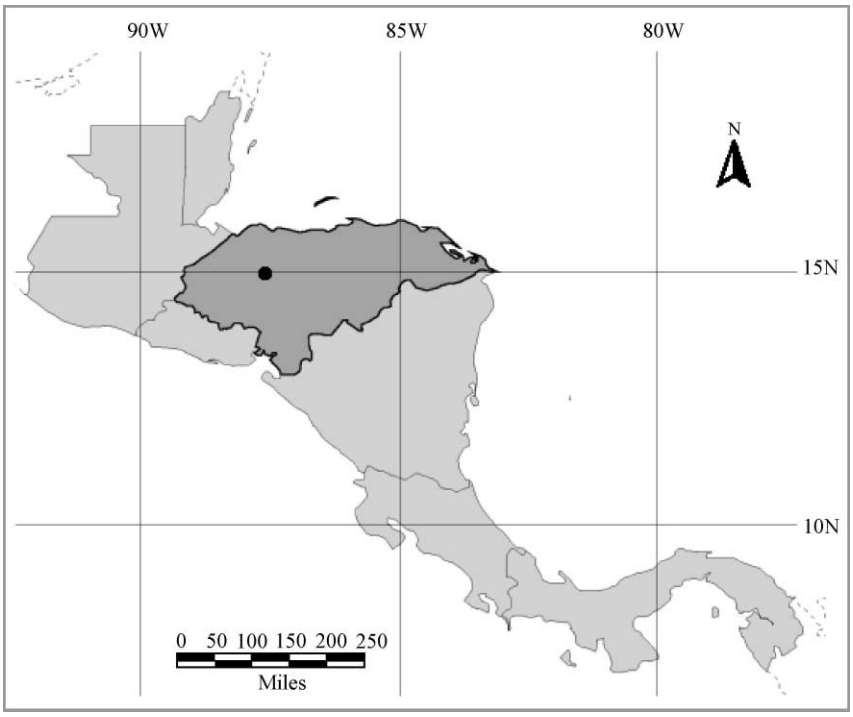


Fig. 2. Central America with Honduras in dark grey (the dot indicates the Tascalapa watershed).

The aim being to assess the importance of scale, location, and organisation within agro-ecosystems by integrating available spatial data sets across a range of geographic scales. This research has aimed towards turning data-rich environments into information-rich environments, to develop common knowledge bases in order to facilitate collective action among stakeholders whose responsibilities require them to view the world at distinct geographical scales.

Section 2 of this paper reviews pertinent spatial data issues and desirable modelling criteria that leads to the list of methodologies presented in Section 3. Example applications are presented in Section 4 and are followed by discussions and conclusions.

2. Geographic information science

2.1. GIS issues

The project objective of modelling geographic information across scales in a data-rich environment brings together two of the most difficult aspects of scale in a geographic sense:

- Aggregation and scale effects.
- Local analysis of spatial data and determining the sampling scale.

Scale effects are prevalent in all data sets and subsequently in all spatial analysis. After a thorough review of the landscape ecology and geographical literature, Jelinski and Wu (1996) concluded that there was no suitable encompassing theory for indicating how sensitive results are to the scale of the analysis and to variations in the way in which data are represented. Openshaw and Clarke (1996), Fotheringham (1998), Fischer et al. (1996) and many others have also voiced concern over this lack of theory behind data representation and a lack of theory behind GIS in general.

It is of paramount importance to realise that when analysing spatial data, it may be incorrect to assume that the results obtained from a study region apply equally to all individuals within that region

(Fotheringham, 1997). Adopting techniques that can identify local rather than global spatial relationships, can help in avoiding this type of ecological fallacy, as can improved access to data. When spatial data form part of an analysis, it would seem prudent that the results should be in the form of a map or image as well as just tabulated results and general statistics.

2.2. Scale effects

The digital terrain model (DTM) of the Tascalapa region in Honduras, shown in Fig. 3, was classified following a scheme proposed by Wood (1996) into geomorphological features such as peaks, ridges and channels. Analysing the surface at scales from 50 m to 5 km (Fig. 4a), even intuitive concepts such as 'what

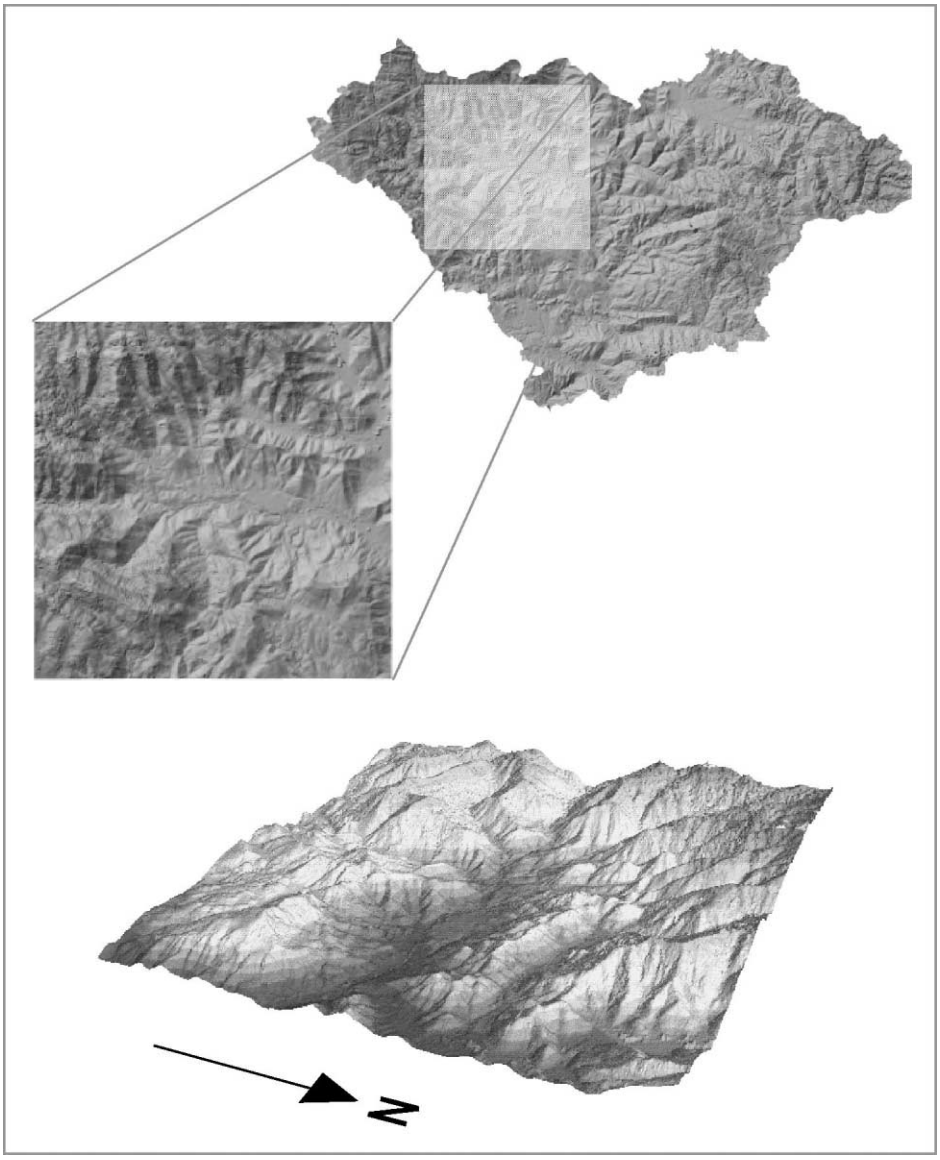


Fig. 3. DTM of the Tascalapa watershed, Honduras (location shown in Fig. 2).

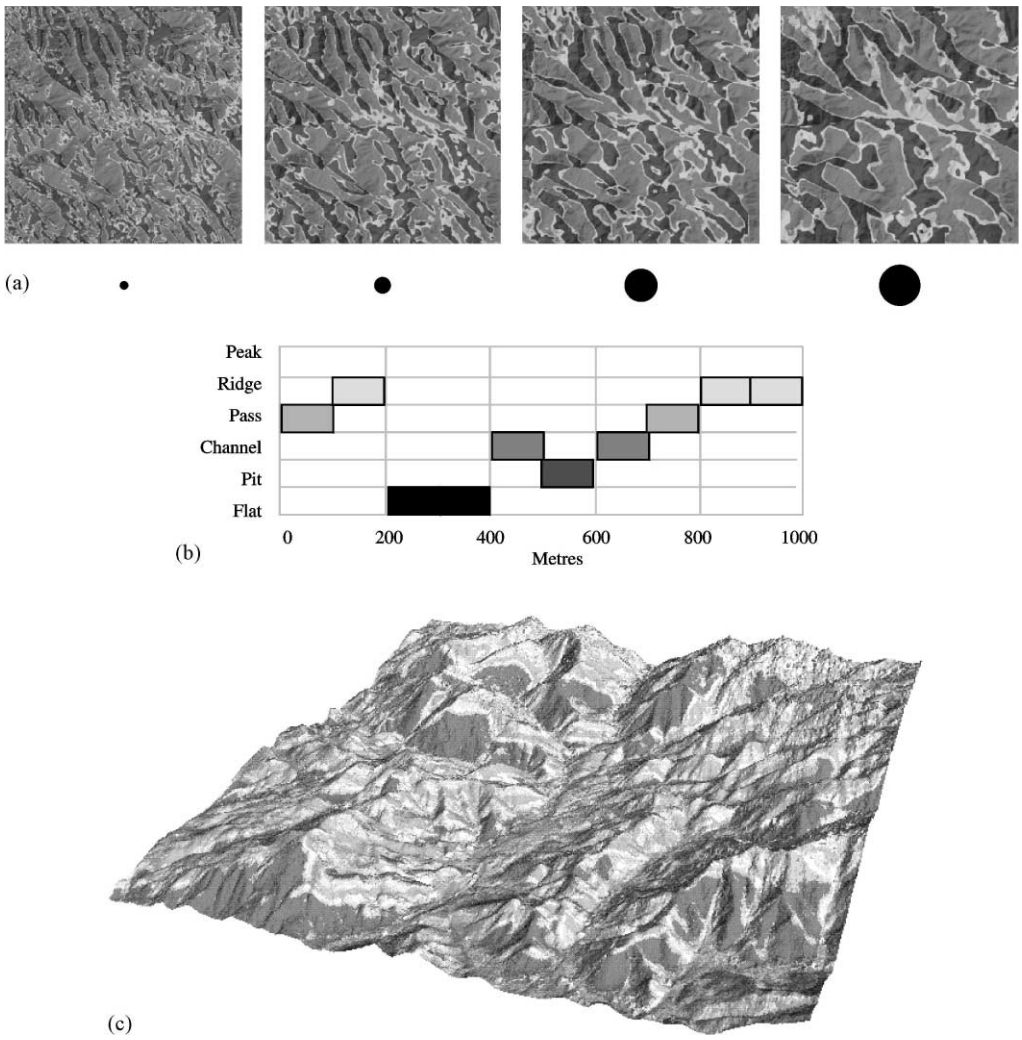


Fig. 4. Various visualisations of a cross-scale analysis of surface features.

is a peak' and 'what is a channel' are scale dependent. Surface features are produced by many processes ranging from soil processes at the finest scale to plate activity at the global scale and Fig. 4b charts this change in feature type for one location across scale. Fig. 4c summarises the entire data set with a measure of entropy, indicating regions of high/low scale dependence.

What is seen is entirely dependent on both how it is measured and the scale at which it is measured. Choosing an inappropriate scale, and the results, will be more likely than unsuitable for the purpose at hand.

2.3. The modifiable areal unit problem (MAUP)

Until very recently, almost all spatial data was only available in some aggregated form, and was often aggregated further to suit the users needs. The results of aggregate analysis are not only related to the degree of aggregation but also strongly dependent on the choice of reporting zones. Areal units, such as administrative boundaries or even image pixels, are usually determined arbitrarily and are modifiable in that they can be spatially aggregated in an infinite number of ways.

In Openshaw's study of the MAUP (1984), he states: 'The process of defining or creating areal units would be quite acceptable if it were performed using a fixed set of rules, or so that there was some explicit geographically meaningful basis for them. However there are no rules for areal aggregation, no standards, and no international conventions to guide the spatial aggregation process. Quite simply, the areal units used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating'. This was in 1985, and to date the situation has not improved much. More disturbingly, Gehlke and Biehl (1934) first explored the concept almost 70 years ago, and there has been little change since.

Openshaw (1996) points out that in order to compare zonal objects it must be certain, from a geographical point of view, that they are indeed comparable objects, otherwise there is a realm of comparing chalk with cheese. For example, correlating the percentage of elderly voters with Republican voters in Iowa counties, Openshaw and Taylor (1979) produced almost any result from perfect positive to perfect negative correlation by manipulating the reporting zone boundaries. Fotheringham and Wong (1991) have produced similar disturbing results when applying the same techniques to site selection and resource allocation models.

2.4. Implications

Progress in GIS has been mainly in data handling and user-friendly software rather than spatial analysis. The potential for misleading or inappropriate analysis and error is magnified by the availability of more data, and users who are unaware of the problem. Aggregation of data must be a controlled and well-defined process. There is a grave danger of obtaining biased, misleading, or poor results when data for possibly inappropriate areal units are studied.

2.5. Solutions?

2.5.1. Greater access to finer-level data

This would not only enable scale and MAUP effects to be further explored, but it would also allow analysts to produce data sets that are closer to their needs, rather than relying on predefined aggregated data.

2.5.2. Improve awareness of these problems among those working with spatial data

Recent texts relating to spatial issues, spatial data integration and multi-disciplinary GIS applications such as Ecological Scale (Peterson and Parker, 1998), People and Pixels (Liverman et al., 1998) and GIS solutions in Natural Resource Management (Morain, 1999) make no reference to the MAUP, suggesting that either the problem is unknown, considered unimportant, or that it has been deliberately ignored. Yet as Openshaw and Clarke (1996) adroitly put it 'It is unacceptable to assume that the MAUP does not exist'.

2.5.3. Provide analysts with tools to investigate and minimise scale and MAUP effects

Much progress and understanding can be gained if the analyst is able to not only demonstrate the effects of scale and the MAUP for themselves, but can also control and define the aggregation process thus creating suitable areal units for the data (Haining, 1990; Fotheringham and Wong, 1991; Openshaw and Rao, 1995; Martin, 1998). Another option is to define new units of measurement. As part of this research methods for delineating new areal boundaries have been investigated using a combination of existing GIS methodologies, new spatial analysis techniques and neural networks. This has led to two new approaches to the problem described in Section 3.

2.6. How can location, scale and the MAUP be accounted for in spatial analyses?

Scale effects complicate any straightforward understanding of spatial data and there is a need to explore and quantify their nature. Adopting an approach that furthers understanding about scale effects should enable greater focus on the scales that relate to the process under study. The first stage of the analysis is both exploratory and empirical. Emphasis is placed on visualisation of the data at a range of scales, to detect the variation in variables and relationships with respect to scale. The second stage suggests improved scale-sensitive process orientated representations and models.

3. Methodologies

Here we briefly review the range of methodologies. Firstly, a spatial characterisation toolkit was created using a combination of traditional and new techniques. Secondly, two methods were adopted for accommodating the MAUP, one previously published, one a novel approach of the project. Finally, a new method of estimating spatial structure and defining system boundaries was investigated by combining two new methodologies for multivariate data analysis. Colour illustrations and further information are available online (http://www.ciat.cgiar.org).

3.1. Developing an exploratory analysis toolkit for assessing scale effects

It is clear that statistics, tables, graphs and traditional exploratory analysis are not enough to describe the complex, scale-dependent relationships that exist in geographic data. From that viewpoint, it becomes imperative that any analysis pertaining to be spatially explicit must have results that are mappable. For this reason, primarily maps represent the majority of the examples in the paper.

3.1.1. Characterising spatial data across scales and across locations

The most commonly applied technique for resampling spatial data is known as convolution filtering, sometimes called moving window sampling or kernel filtering. It is a flexible technique that can be applied to point, areal and image data types.

A (circular) window moves over the study region and for any given location, all data falling within that window are filtered to produce a new data value for that location. This method of intensive sampling does not aggregate the data set spatially (i.e. if the input data has 100 points, so will the output; if the input image has 1000 pixels, so will the output) but rather it extracts certain characteristics from the data. The window can contain any number of functions, from a simple mean filter to a diversity index, to a measure of spatial autocorrelation, to a multivariate regression. Table 1 lists the functions that have been applied. By repeating the process with a wide range of window sizes, the effects of scale can be visualised and assessed for every location in the region.

Table 1
The local statistics and their global counterparts that have been applied in the project

Statistic	Local	Global
Momental measures	Yes	Yes
Quartile measures	Yes	Yes
Majority/minority filters	Yes	Yes
Autocorrelation (standard and anisotropic)	Yes	Yes
Correlation coefficients	Yes	Yes
Semivariance (standard and anisotropic)	Yes	Yes
Spatial lag operators	Yes	Yes
Getis and Ord G^* statistic	Yes	Yes
Lacunarity measure	Yes	Yes
Diversity indices	Yes	Yes
Join count analysis	No	Yes
Clustering algorithms (GAM)	Yes	No
Texture analysis	Yes	No
Morphometric analysis	Yes	No

Traditionally, the type of data being filtered will define the window size (pixels, distance to points, nearest neighbours, degree of connectivity between regions).

3.1.2. Extending the concept of a spatial neighbourhood

All these measures are based on Euclidean distance. Other factors such as time, cost, energy or accessibility might be more appropriate, dependent on the data, the scale and the purpose of the analysis. Two locations (A and B) might well be equidistant from a third location C, yet the time or cost required to travel from A to C could be twice as much to travel from B to C. Fig. 5 highlights the differences between a distance buffer and a time buffer around two locations in Honduras.

3.2. Deriving geographically sensible regions to represent the data

There is little doubt amongst geographers that the MAUP is one of the biggest bottlenecks for performing sensible spatial analysis, but it appears that the message has been slow to spread to other spatial disciplines. Any study region is characterised by a spectrum of social, cultural, physical, economic and agricultural factors and constraints, and traditional spatial units (watersheds, political units, pixels) cannot claim to represent all four of these dimensions, and hence other spatial units are required.

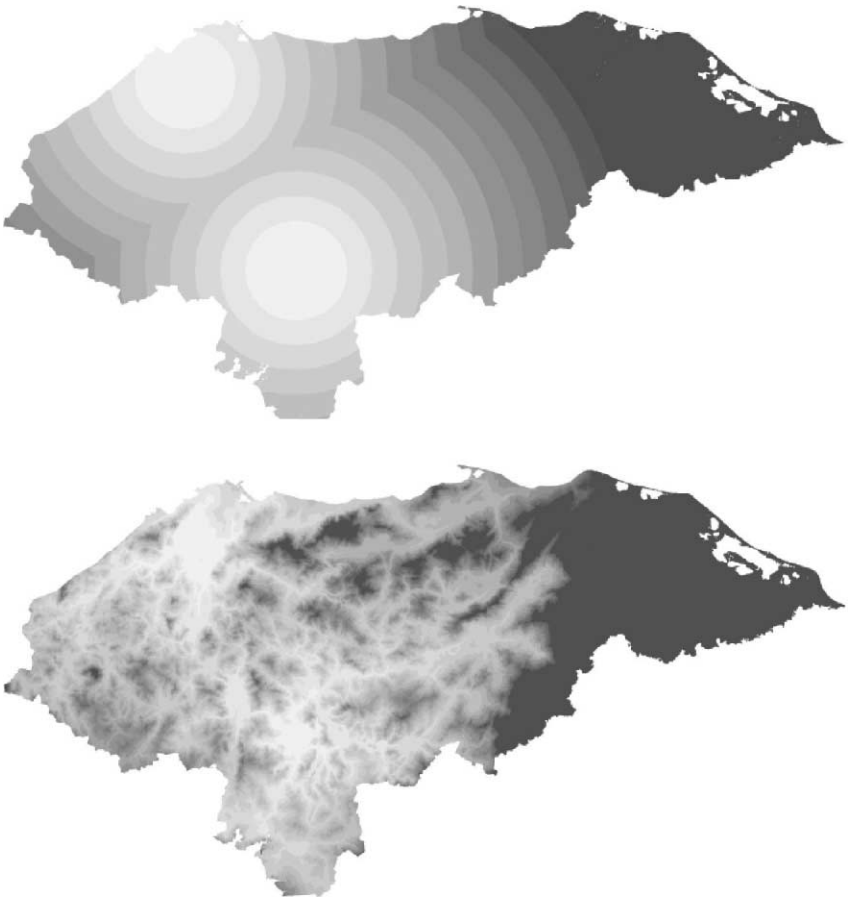


Fig. 5. Distance (top) and travel time (bottom) from Tegucigalpa and San Pedro Sula in Honduras.

3.2.1. Regionalisation and zone design algorithms

Of the methodologies reviewed in this report, this is possibly the most conflicting and controversial. Once it is accepted that the results of studying zonal data depends on the particular zoning system that is being used, then it is no longer possible to continue using the normal science paradigm. The data are not fixed; therefore the results depend, at least in part, on the areal units that are being studied; units that are essentially arbitrary and modifiable. The selection of areal units, or zoning systems, cannot therefore be separate from, or independent of, the purpose and process of a particular spatial analysis; indeed it must be an integral part of it (Openshaw, 1984).

Zone design or regionalisation is based on a spatial optimisation process, where n original zones are reduced to m output zones, with the constraints that each of the original zone belongs to only one m zone and that all members of the output zone must be connected. Essentially it is a constrained, non-linear, integer, optimising problem that can only be solved via heuristic methods such as simulated annealing.

Regionalisation algorithms not only provide evidence of the potential for misinterpretation of spatial data, but also are a means to explore and develop new meaningful data representations. Since the zone simplification is a controlled process, there is greater confidence that the new areal units are indeed comparable, unlike many other geographical analyses where the data are represented by units that are neither similar, comparable nor relevant.

A software tool called ZDES (Zone DEsign System) developed by Openshaw and Rao (1995) has been applied in this project to perform this type of user-controlled data reduction.

3.2.2. Delineating new study regions with a cost–distance algorithm

Instead of using existing areal units as the basis for regionalisation, it is also possible to define completely new areal units using some common GIS algorithms. One possibility is a cost–distance function, which is available in GIS packages such as ArcInfo and IDRISI. Specifically the algorithm takes a target coverage, such as a map of markets, hospitals, schools or any other point of interest, and a friction coverage, where friction is some measure of the cost required to travel from any location in the study region to the nearest target, and computes catchment areas around each of the targets. The friction surface can be a combination of roads, rivers, slopes, land use, international boundaries, indeed anything that is relevant to the study at hand. In this way, the analyst can explicitly define the cost required to gain access to each location. Then the cost–distance algorithm is applied and a set of explicitly defined the catchments are generated around each target. These sheds can be applied to many other types of target (health care and schools are two possibilities) to generate *ecoregional-sheds*.

These generated units are unique in that:

- Each catchment is focused on a resource or market, commonly referred to as a target.
- They are defined by local physical, agricultural and economic factors.
- The units are dynamic since they adapt with temporal changes in the input data.
- Accessibility is an inherently scaleable framework.
- They can be applied to a range of issues.
- They are entirely user and purpose defined.

The final point is critical. It is possible to create exactly the areal units or boundaries that are required by explicitly stating the focal point or resource that the local population need to access, and their potential to gain access to it. This cost–distance algorithm is only one possible approach to the MAU problem within a GIS environment.

3.3. Defining spatial structure and system boundaries

The increasing availability of large and complex spatial data sets has led to a greater awareness that conventional statistical methods are of limited use, and that there is a need to understand local variations in more complex relationships. In response, several attempts have been made to produce localised versions of traditionally global multivariate techniques, with perhaps the greatest challenge being to produce local versions of regression analysis (Fotheringham et al., 1997).

Regression is the most commonly applied method for multivariate analysis, and is the focus for this section of the paper. However, before regression can be considered as an analytical framework for discovering the spatial structure within complex agro-ecosystems, it is useful to reiterate the seven classical assumptions that must be met in order for the OLS estimates to be the best available.

1. The regression model is linear in the coefficients and has an additive error term.
2. The error term has a zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are uncorrelated with each other (no serial correlation).
5. The error term has a constant variance (no heteroskedasticity).
6. No explanatory variable is a perfect linear function of any other explanatory variable.
7. The error term is normally distributed.

This places fairly restrictive limitations on what can and cannot be reasonably expected of an OLS regression in a spatial context. With that in mind, the assumptions can be compared with a list of some of the characteristics of spatial data and the consequences for regression modelling (Table 2).

Some of these characteristics are always present, which all too often means that the majority of traditional statistics or modelling techniques are inappropriate, invalid or too general. Since this realisation (in the 1950s), two distinct branches of research have appeared.

- The continued application of these models either in ignorance of the autocorrelation case or on the

Table 2
Spatial data characteristics and the problems for regression modelling. Sources (Openshaw and Openshaw, 1997; Haining, 1990)

The characteristics of spatial data	Potential consequences
The presence of spatial dependencies and autocorrelation	Inferential tests invalid, inflated R^2
Non-linear relationships	Non-independent residuals
Presence of discontinuities	Distorted model fit
Non-stationarities	Error estimates are biased
Non-normal frequency distributions	Inferential tests invalid
Scale and aggregation dependencies (MAUP)	Meaningless results
Mixtures of measurement types	Need a GLM (logistic model)
Surrogate data and proxy variables	Inferences are suspect
Noisy, incomplete and often ill-suited	Inferences are suspect
Data outliers	Model is distorted
Varying degrees of data reliability and understanding	Meaningless results
Huge amount of variables, some of which are redundant	Inefficient, need a stepwise model
Non-numeric data	Need a GLM (logistic model)

grounds that biased-coefficients problem refers only to the use of the general linear model in forecasting and prediction, and does not affect the procedure as long as it is used purely for descriptive purposes.
- The development of procedures for spatial forecasting, with a focus on the patterns rather than on their generating processes, i.e. deducing spatial processes from mapped patterns (Johnston, 1991).

Of the problems stated in Table 2, there are three that require direct intervention in the design of a multivariate analysis for spatial data. In order of importance (Griffith and Amrhein, 1997), they are:

- The variations in relationships and processes over space (spatial non-stationarity).
- The spatial dependencies across space and across variables (spatial autocorrelation).
- That geographical data rarely have a linear distribution.

If non-stationarity in a regression model can be dealt with, then it can be further adapted to address the two lesser difficulties of non-linearity and autocorrelation. One method for accommodating spatial variation within a regression analysis framework is geographically weighted regression (GWR), which has been adapted and extended in this research as a method for detecting system boundaries within a multivariate analysis.

3.3.1. Geographically weighted regression

GWR accounts for the spatial 'drift' in linear relationships (Fotheringham et al., 1997), by localising the regression (placing it inside a kernel) and accepting that linearity does exist, but only over limited spatial scales. This concept is valuable for several reasons:

- It allows greater insights into the nature and accuracy of the data under scrutiny.
- It provides a detailed understanding of the relationships and their spatial variation.
- It demonstrates the possible naiveté of conventional approaches to data analysis that often ignore spatial non-stationarity.
- It allows a more detailed comparison of the relative performances of different types of analysis or different models (Fotheringham et al., 1997).

It is possible to test whether hypotheses such as does the GWR model describe the data better than a global regression, and do the regression coefficients vary significantly over space for a given scale? The ability to not only model spatial variation but also to map the regression parameters and goodness-of-fit measures makes GWR a very attractive option for analysing spatial structure. By mapping the coefficients, regions where a coefficient 'flips' between a positive and negative contribution are revealed. This method produces a huge amount of output information — each coefficient is now represented by a map rather than a solitary number. The problem of how to condense this information into system boundaries is dealt with in the following section.

3.3.2. Self-organising maps for data characterisation

The surfaces that are generated by GWR can be viewed as a set of system parameters that can be

combined to define system boundaries. These boundaries can be discovered by applying some form of classification or clustering to the parameters. Any spatial structure or organisation would be visible if there is some degree of clustering in the parameters and in their spatial distribution.

The data set must also exhibit some clustering tendencies in order that the use of clustering algorithms would be sensible at all (Kaski and Kohonen, 1996). Where there is no relevant understanding or no function to optimise, a potential solution for deriving simplified data sets is the unsupervised or competitive learning neural net. The network discovers for itself the features of the data that matter most by developing its own featural representation that captures the salient characteristics of the input data. The only assumption is that some structure exists.

The self-organising map algorithm (SOM) (Kohonen, 1997) is a unique method in that it combines the goals of the projection and clustering algorithms. It can be used at the same time to visualise the clusters in a data set, and to represent the set on a two-dimensional map of neurons (or processing elements) in a manner that preserves the non-linear relations of the data items. That is to say that nearby items are located close to each other on the map. The SOM has several interesting properties, namely:

- The mapping represents data in an ordered form, whereby mutual similarities of data samples will be visualised as geometric relations of the images that form the map.
- The natural order inherent in the mapping enables the map to be used as a groundwork on which the individual inputs can be visualised as grey levels.
- The structures in the data can be automatically visualised on the map whereby the degree of clustering is represented by colour coding or shades of grey.
- The commonplace issue of missing data can be treated elegantly (Kaski and Kohonen, 1996).

The neurons have a specified spatial structure to them, usually arranged on a two-dimensional grid. Data sets are 'dropped' into the net one by one as inputs and the neurons compete amongst themselves for the right to represent the input. The neurons that is most similar to the input is declared the winner. Once a winner is found then updating takes place not only of the winning neuron but also of those within a certain critical neighbourhood of it, whilst those further away are inhibited. This neighbourhood is gradually reduced during training. Once the network has converged to a solution, the weights associated with neuron define an efficient partitioning of the multidimensional data. The SOM provides a non-parametric pattern classification (Openshaw and Openshaw, 1997).

By using the parameter maps as inputs, any strong spatial pattern in the regression parameters can be extracted by the SOM and used as a base map for delineating system boundaries.

4. Example applications

4.1. Bivariate correlation: local analysis versus global analysis

The DTM example has shown that cross-scale information can be easily extracted from spatial data, and how this information affects interpretation. Here this process is taken further by indicating how it might be used for suggesting the direction of further modelling with the data. Essentially this example is a comparison of a local analysis of bivariate correlations with national level correlation coefficients using key agro-economic variables in Honduras, namely the number of dry months, population density per *aldea* and the percentage of farms using improved varieties.

Local variations in the relationships were compared to administrative boundaries to 'eye-ball' regions where the *municipios* do or do not adequately represent the underlying patterns in the data. Fig. 6a shows the agricultural productivity per *aldea* (white is low, black is high), and Fig. 6b represents the three variables to be correlated with productivity; improved varieties, rainfall and population density, respectively, where again darker shades indicate higher values.

The first stage is to conduct the correlation analysis at a huge range of scales, to give an indication of the rate of change of the measurement with scale. Fig. 7 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

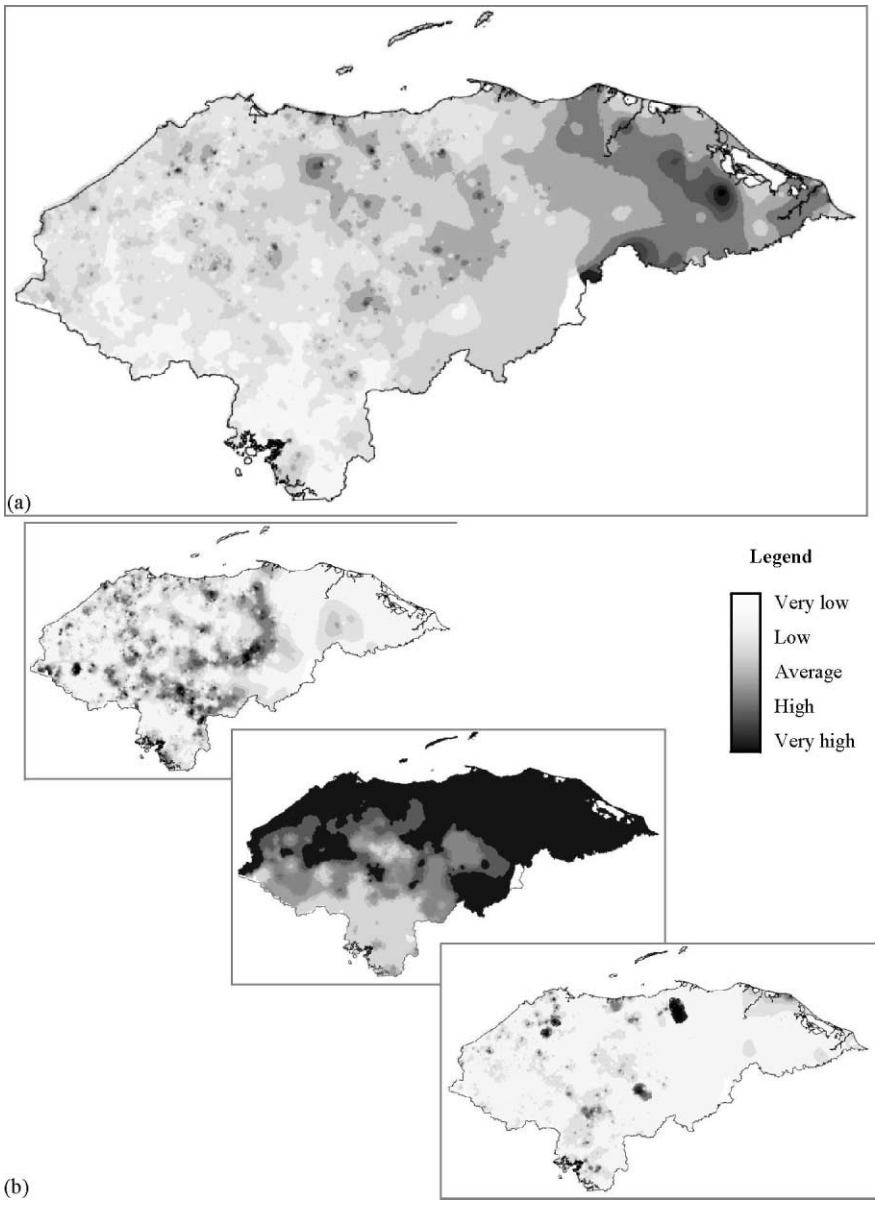


Fig. 6. Agricultural income and three correlate variables (seed varieties, rainfall and population).

of *aldeas* in the data set) 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.

Presenting maps for each scale would be very laborious, so one scale (24 nearest neighbours) has been chosen that corresponds to the number of *aldeas* in the majority of *municipios*. For visualisation purposes, the correlation coefficient is mapped as an interpolated surface. Fig. 8a is for improved varieties, Fig. 8b for monthly rainfall and Fig. 8c for population

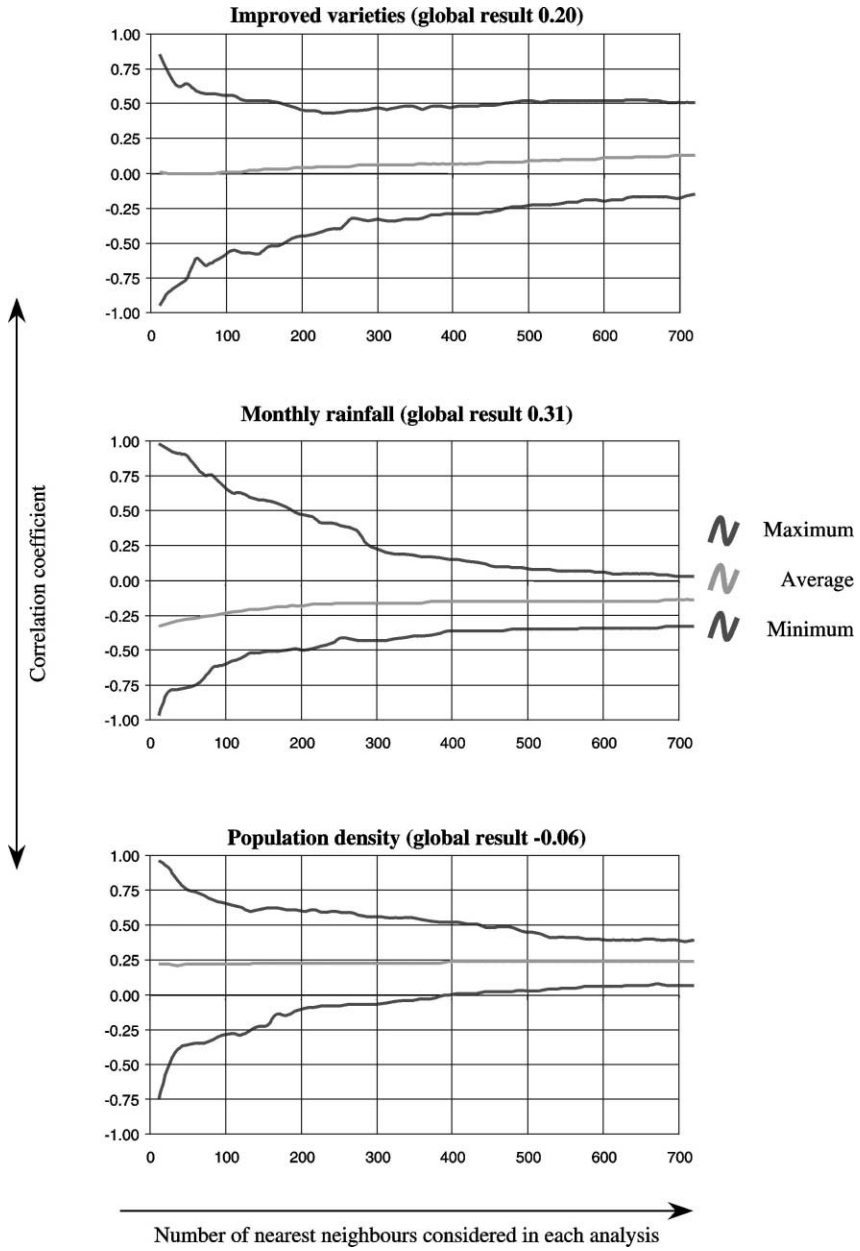


Fig. 7. Summarised correlation across-scale between agricultural income and each variable.

density. 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 data set. These patterns possibly relate to local processes, which could improve the 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 data set. A localised regression

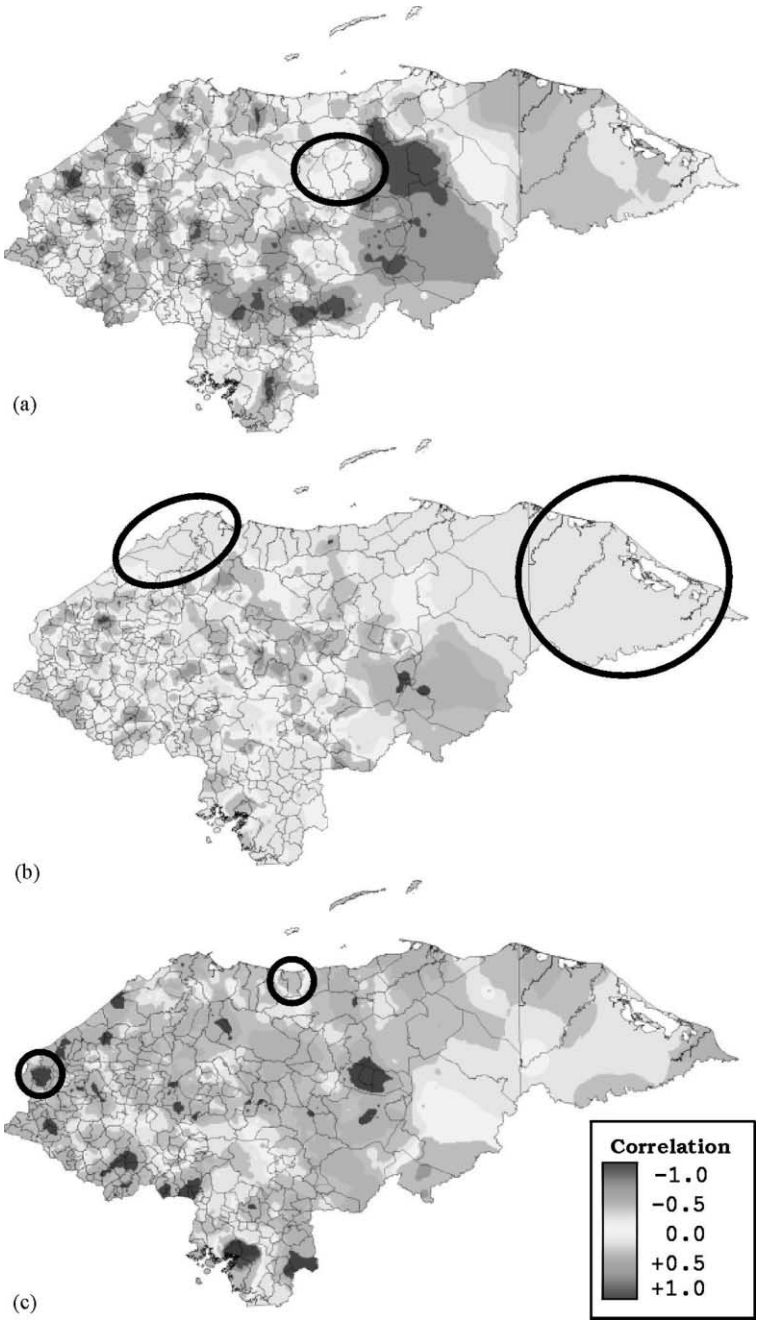


Fig. 8. Surfaces of local correlation for the three variables (seed varieties, rainfall and population).

method based on a similar neighbourhood function is presented in Section 4.3 and this same data set 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 there are very good fits between the boundary and the underlying data (e.g. the highlighted areas in Fig. 8a and b and large areas of Fig. 8c), but in most cases they are not. With the exception of monthly rainfall (since monthly rainfall patterns change little over such small regions as

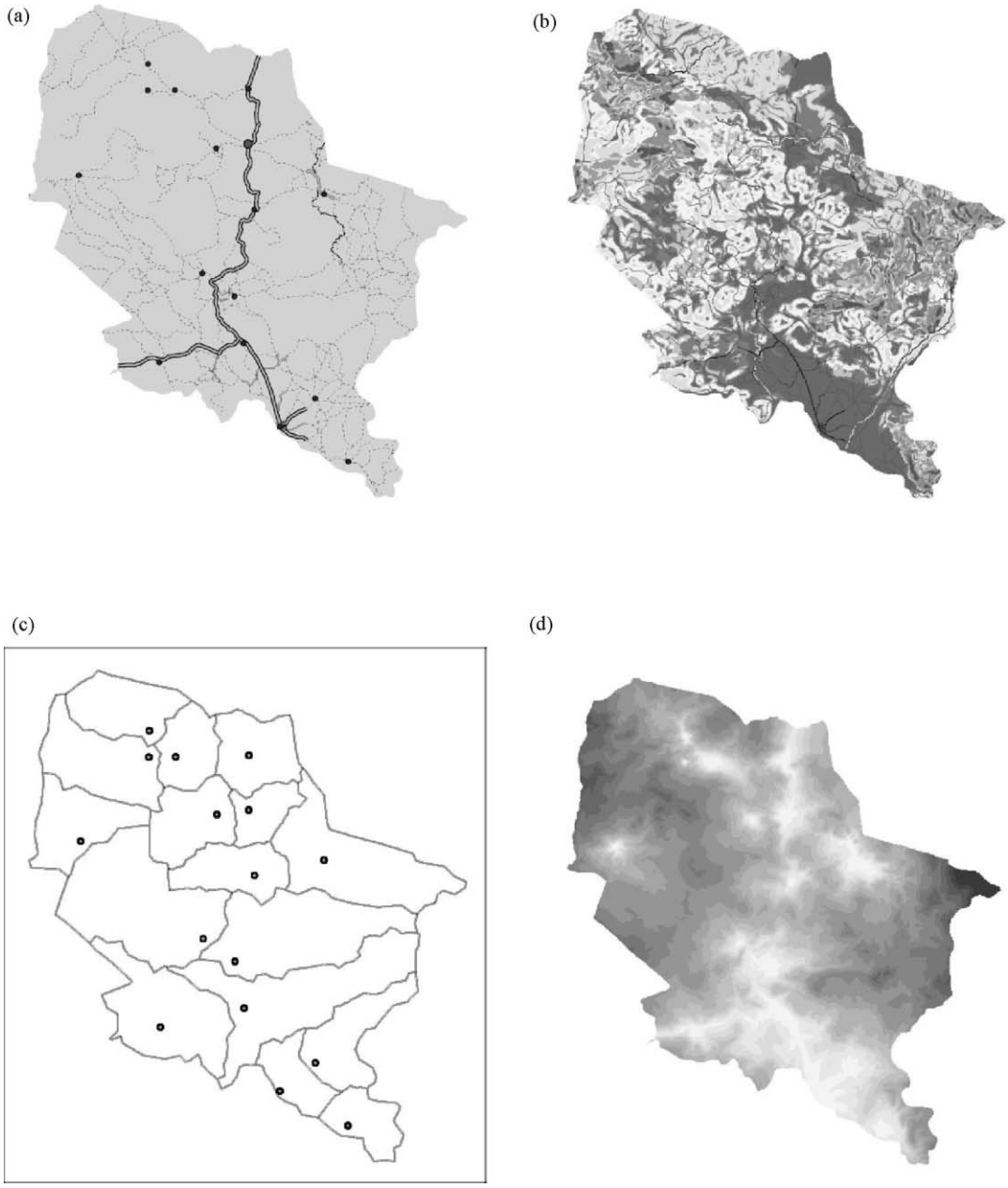


Fig. 9. Infrastructure and travel cost (upper), generated boundaries and travel time (lower).

municipios) aggregation to municipal boundaries will distort and may destroy these relationships.

4.2. Generating ecoregional-sheds

This study compares accessibility to markets with the Von Thünen model. The Von Thünen model defines concentric rings of land use surrounding an (isolated) market based on Euclidean distance from the market. With accessibility, bands of land used are defined based on time to market, changing the symmetrical nature of the model. This can be seen as a simple adaptation of the Von Thünen model where time replaces distance. It is expected that the amount of agricultural land will decrease steadily with increased distance from the market, and likewise the amount of forested land should increase. Additionally, urban or degraded land should decrease with distance from market.

The test site for this study surrounds the CIAT benchmark site of Tascalapa, in the *municipios* of Yorito and Sulaco. The main road runs north–south through the two *municipios*, passing through the town of Yorito and other *aldeas* (Fig. 9a). The road network and slope map were combined to create an ease of travel surface (Fig. 9b) and along with the *aldea* centres were used to compute the boundaries (Fig. 9c) and accessibility (Fig. 9d) around each *aldea*. Fig. 10 shows the land-use/land-cover images (1994, TM image, 30 m resolution) that was available for this region. This land-cover image was combined with the access map to determine the percentage of land use within each time band. For ease of visualisation the land classes have been reduced to farmland, forested and *other* (urban areas, degraded land or bare soil). The distance map and the land-cover percentages per distance band are represented in Fig. 11a and b, respectively. As expected there are definite trends for each class in each image:

- The percentage of the farmland class decreases with distance.
- The percentage of the other class decreases with distance.

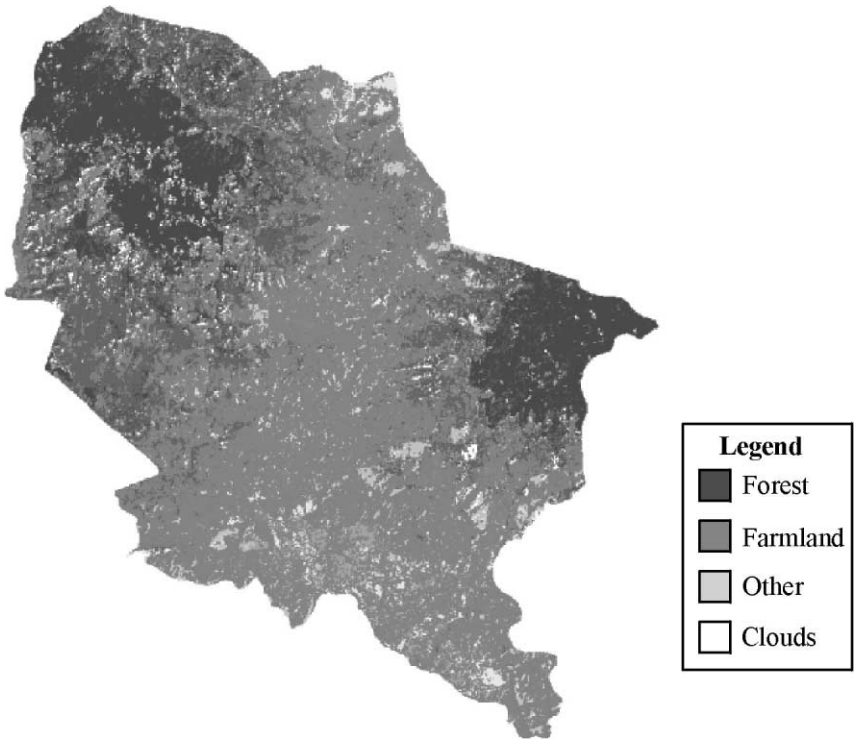


Fig. 10. Land-cover classes.

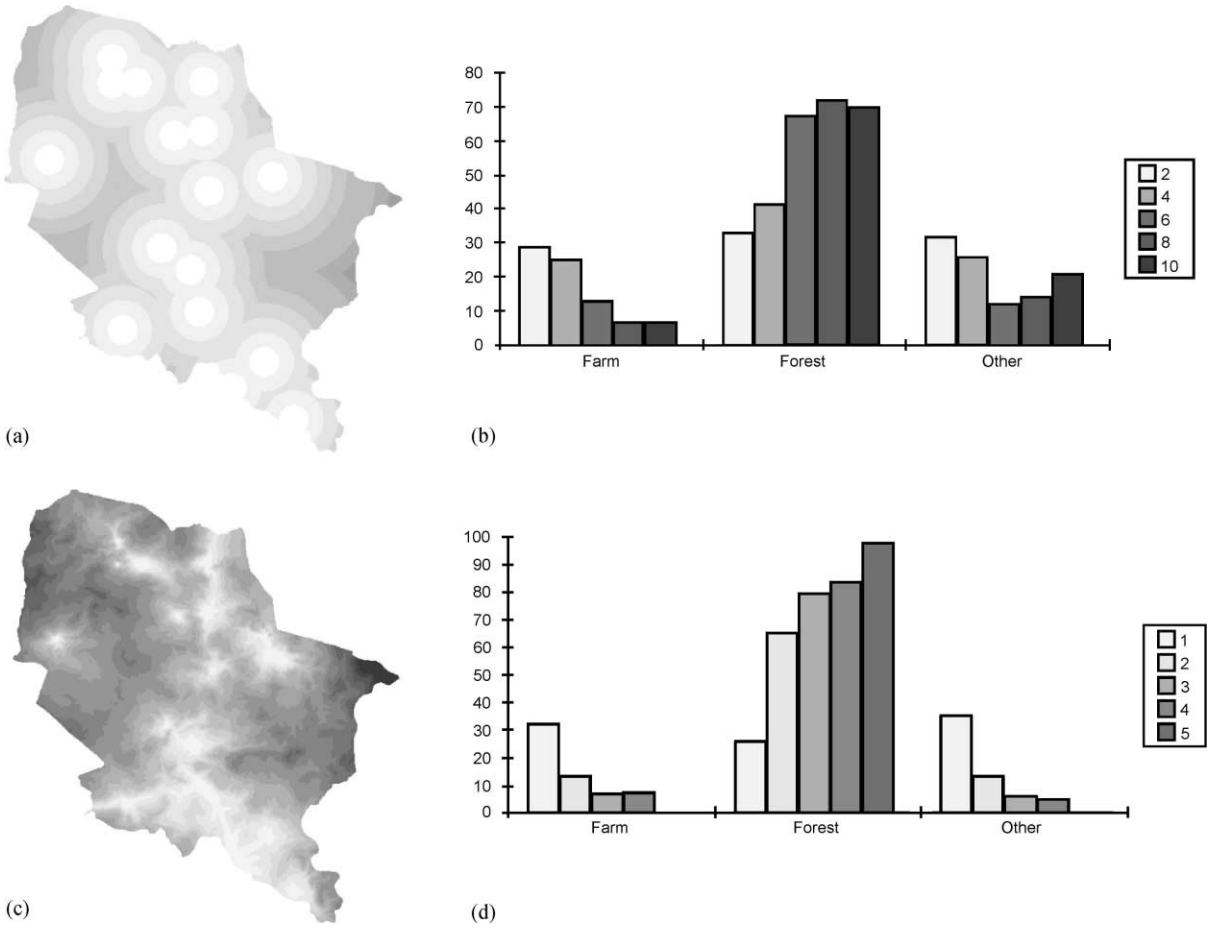


Fig. 11. Distance (upper) and time (lower) compared to land cover for all *aldeas*.

- The percentage of the forested class increases with distance.

Similarly, travel time from each *aldea* (Fig. 11c) was combined with the land-cover image, and percentage land cover per time band was calculated (Fig. 11d). The same strong trends are evident for all three land classes in both time periods.

The town of Yorito is recognised as being the major local market for this area, and so the experiment was repeated for Yorito only. Does the major local market have an influence on the surrounding land-use patterns? Fig. 12a shows the distance in kilometres around Yorito, and Fig. 12b the corresponding land-cover percentages per distance band. The trends have completely changed in the distance model, the majority of forested land appears to be nearer to Yorito than the majority of agricultural land, which exhibits a U shaped histogram. But when travel time is considered (Fig. 12c and d) the same trends appear for Yorito as in all the *aldeas*.

The relationship between land use and distance/time is both clear and intuitive at the local level. Each *aldea* has an influence on its surrounding land use. The relationship between land use and distance has been lost in the 'up-scaling' from local population centres to the major local market. However the accessibility model has retained the overall relationship, suggesting that boundaries and surfaces of accessibility are a better framework in which to study market linkages and the relationships that exist in agro-ecosystems at one or more scales. The project has also performed analyses

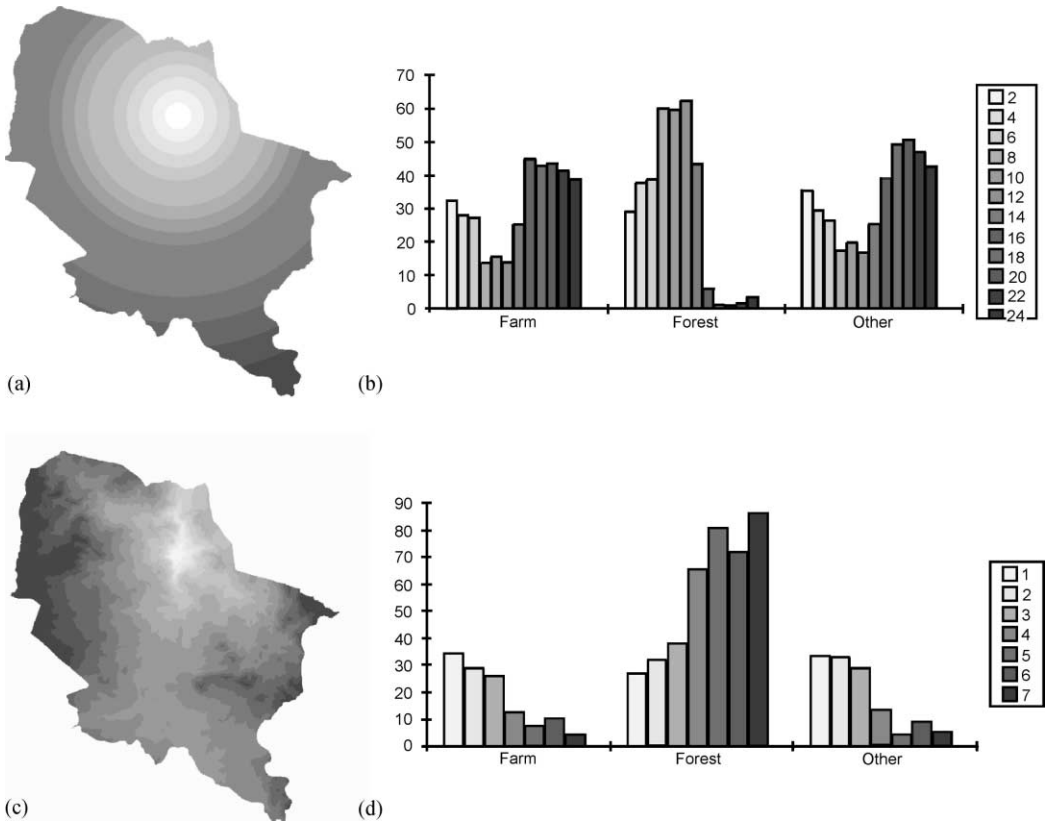


Fig. 12. Distance (upper) and time (lower) compared to land cover for Yorito.

for two land-cover dates across three test regions in Honduras with similar results.

4.3. Geographic regression and system boundaries

This example examines the link between agricultural labour productivity and natural resource, socio-economic and farming system variables at the national level in Honduras. An OLS regression was applied to the variables followed by a GWR model, and the results compared.

Each variable has been calculated at the *aldea* level. The spatial distributions of some of the dependent (average production per worker) and independent variables are shown in Fig. 6 as interpolated surfaces. Note that this interpolation is for visualisation purposes and does not play any part in the model. There is a strong spatial trend in the dependent variable with the lowest production values being in the south west near the El Salvadorian border, and the highest values being in the plantation areas of the Caribbean coast and the sparsely populated Mosquita region to the east.

There is little collinearity in the independent variables and an OLS regression model was applied, with the results shown in Table 3. All variables are significant (at the 1% confidence limit) and the parameter estimates have the expected signs (positive for rainfall, education, credit, technical assistance, labour and improved varieties, and negative elsewhere). However the fit (0.38) is not great, although it is not too bad considering the number of independent variables. The local correlation analyses in Section 4.1 suggest that a GWR model could be justified.

A GWR model was applied to the data, and a cross-validation procedure determines the optimum bandwidth of 51 km (shown in Fig. 14). The results of the GWR model are summarised in Table 4, where the summary values for each parameter are presented

Table 3
Tabulated OLS results (root mean square error = 0.371)

Parameter	Estimate	Standard error	*T*-test
Intercept	+3.843	0.0377	+101.5
Months of rainfall	+0.050	0.0020	+25.3
Access to ports	+0.003	0.0002	−14.4
GINI coefficient	−0.400	0.0353	−11.5
Level of education	+0.014	0.0012	+11.6
Technical assistance	+0.001	0.0001	+3.8
Credit	+0.001	0.0001	+4.8
Temporal labour	+0.009	0.0023	+3.9
Improved seed varieties	+0.003	0.0002	+12.1
Population density	−0.014	0.0011	−12.4
Access to market towns	−0.002	0.0004	−4.8

as the model varies across the region. The variations in each local parameter lie outside the range of their global values and standard errors. Every parameter (except the intercept) changes sign across the region, and these variations are visualised in Fig. 13, where dark regions are below the global value and light values are above. Very dark or very light shades indicate greater variation from the global value (in white). The GWR R^2 values are represented in Fig. 14 where dark shades indicate that an improvement over the OLS model was achieved. Table 5 justifies the application of a GWR model.

The contribution of each variable in the regression equation has changed over the study region, including sign change from positive to negative and vice versa. For instance, the access to ports (globally positive) parameter is highest in the areas nearest to the ports, implying that the cost of transport to the ports is prohibitive past a certain threshold. Conversely, access to towns (globally negative) has a positive impact for those regions with little access to ports. There is an inverse relationship between temporal labour and education, suggesting that labour costs are greater where literacy rates are higher. Another revealing pattern is found in the population density parameter (globally negative), which is only negative in isolated inaccessible areas, such as the eastern and western extremes of the country with little road access. The rate of change of each parameter is related to the chosen bandwidth, a smaller bandwidth leads to rapid changes and conversely as the bandwidth with tends to infinity, so the GWR model tends to the OLS result. Here the bandwidth is relatively large, creating surfaces with gradual trends. The next step is to use these parameter maps as inputs to the SOM to extract system boundaries from the model.

The 11 parameters for 3500 locations were normalised and a two-dimensional map of 24 × 12 neurons used to represent the data structure. Training took

Table 4
GWR results (with a bandwidth of 51 km)

Parameter	Minimum	L. quartile	Median	U. quartile	Maximum
Intercept	+0.788	+3.962	+4.363	+4.542	+6.423
Months of rainfall	−0.229	−0.001	+0.011	+0.033	+0.290
Access to ports	−0.007	−0.003	−0.000	+0.003	+0.016
GINI coefficient	−0.675	−0.532	−0.419	−0.314	+1.149
Level of education	−0.059	+0.008	+0.010	+0.012	+0.049
Technical assistance	−0.003	+0.000	+0.001	+0.001	+0.003
Credit	−0.003	+0.000	+0.001	+0.001	+0.004
Temporal labour	−0.048	+0.004	+0.010	+0.016	+0.081
Improved seed varieties	−0.026	+0.002	+0.002	+0.003	+0.073
Population density	−0.118	−0.021	−0.016	−0.013	+0.014
Access to market towns	−0.010	−0.001	+0.001	+0.004	+0.007

Table 5
ANOVA comparison of the OLS and GWR models

Source of variation	SS	MS	Definitions
Ordinary least squares regression	236	0.037	SS (residual sum of squares)
GWR	220	0.034	MS (mean square)

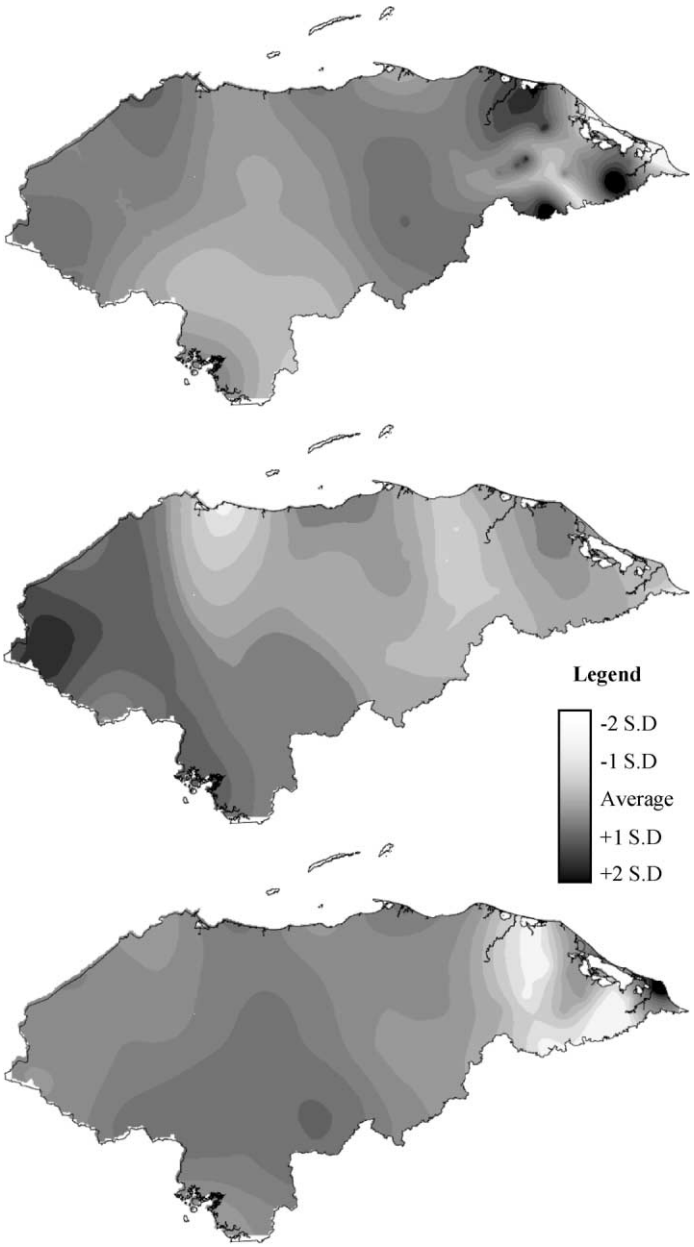


Fig. 13. Three parameter maps: intercept (upper); access to towns (middle); rainfall (lower).

about 1 h on a PC laptop and a very low quantisation error of 0.81 was achieved. The output map is shown in Fig. 15a where similar hues of grey indicate clustering and rapid changes in colour represent the different patterns within the parameters. The shading was applied directly to a map of *aldeas* (Fig. 15b). Not only are there strong patterns within the parameters (from the SOM map), but the patterns also have very strong geographic distributions with definite boundaries (from the geographic map). There are a few isolated *aldeas*, such as the dark region on the eastern coast, but in general, the groups

Fig. 14. R^2 map from the GWR model, darker shades indicate higher R^2 values.

are compact and contiguous. Table 6 lists the seven groups, their locations, and the main agriculture associated with each region. The seven groups relate well to the predominant farming trends and agro-economic trends that are known to exist in Honduras, although the grouping of Atlantida (north coast) and a large swathe of south-western Honduras seems counter intuitive.

5. Discussion

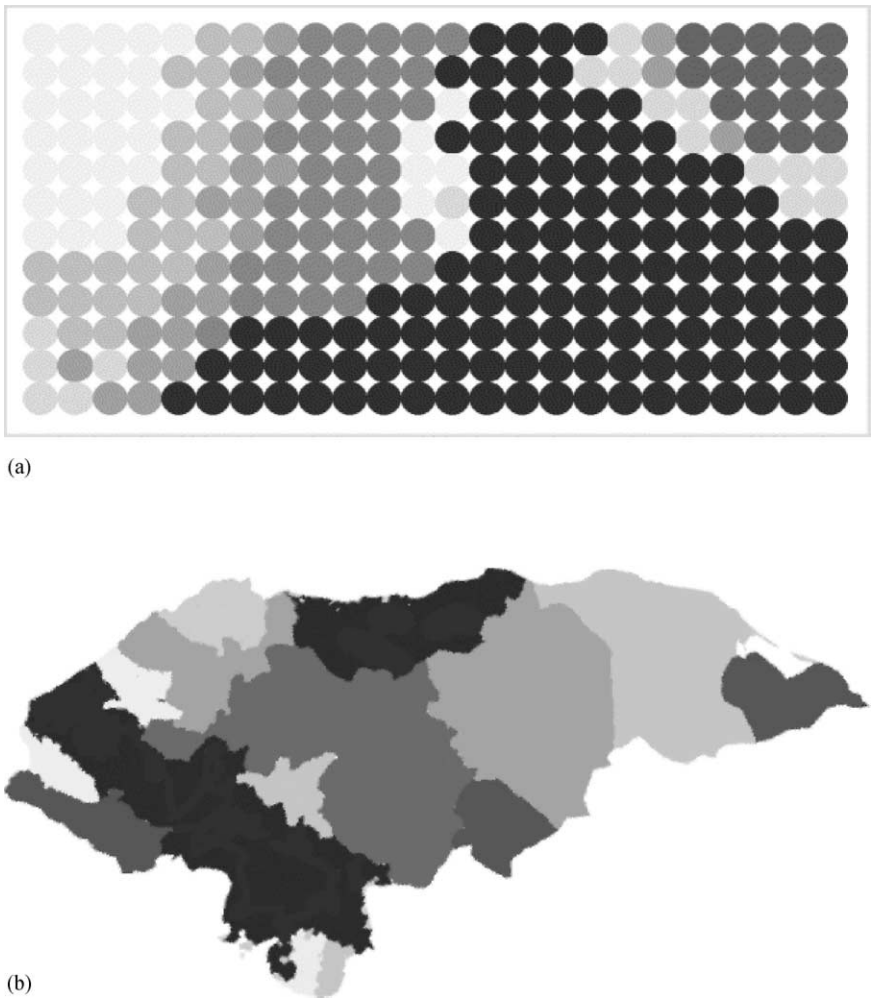
This paper has presented a series of techniques and tools that allow spatial data sets to be constructed and de-constructed in a generalised, yet context sensitive manner. We state that the outputs from such techniques can be explored and described through various user-defined levels, thus revealing spatial patterns and processes that are arguably more useful than raw data or standard representations. Further, 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.

5.1. Putting geography back into the equation

These techniques can be applied to most common GIS data types, and operate very rapidly across several scales. The cross-scale 'signatures' that are generated are a method of attaching a degree of confidence to spatial data by assessing their scale dependence and ability to faithfully represent spatial patterns. It can also be used for sensitivity analyses, which is especially important if the data is being used for spatial policy or decision-making. The method does however produce a large amount of data (one data coverage per scale analysed) and further work is required to permit equally rapid visualisation and assessment of the results. The summary images shown in Figs. 4c and 15 are ways of achieving this.

It is also possible to use local statistics to extract potential new levels of analysis from the data. There are often compelling reasons to analyse data at inappropriate 'fixed' levels. Local statistics reveal how these fixed levels relate to the underlying data, and to determine if there are areas where the fixed levels do fit the data and areas where they do not.

The regionalisation methods are part of a recent trend in quantitative geography where the user is becoming increasingly able to define their study region precisely. The zone optimisation algorithms are computationally intensive although this becomes less



(a)

(b)

Fig. 15. The SOM output map (top) with the shades of grey mapped back onto Honduras (bottom).

Table 6
The groups defined by the self-organising map, and their general locations

Group	Region
Black	Concentrated in three very poor inaccessible areas
Very dark grey	Two plantation areas, Gracias a Dios and around San Pedro Sula
Dark grey	The ranching region of Olancho
Grey	Impoverished areas in Choluteca and the west
Light grey	The area close to Tegucigalpa
Very light grey	Central Honduras, hillside smallholdings are predominant
White	Two zones, Atlantida in the north and south-western Honduras

relevant day by day as computing power increases and hardware prices fall. The cost–distance algorithm provides another option whereby new regions can be built based on ancillary data. Freely available world-wide data sets for slope, roads, rivers, towns and land cover means that this method is not dependent on a data-rich environment. Dynamic ecoregional-sheds can be created for practically any location and any user-defined purpose, although further work on the algorithms sensitivity to data resolution is required. The concept of delineating unique regions for each case study could be difficult to introduce, since there is a deeply ingrained heritage of doing exactly the opposite. For example, watersheds have been used for practically everything from studies on poverty alleviation to transport development.

The GWR model is another methodology that uses maps as analytical tools, and if nothing else it highlights the conceptual flaws in applying non-spatial methods to spatial problems. GWR could also be run at several different bandwidths as part of an exploratory analysis. Neither the GWR nor SOM have been developed by the project, but their combination represents a first attempt to convert multivariate parameter maps into system boundaries in such a manner. This method is particularly elegant since it is an automated process, GWR locates an optimal bandwidth and generates the parameter maps, and SOM takes the parameters and generates an optimal representation of the data, with no user input or tweaking of model parameters.

SOM has also been used in this project as a data reduction method to convert census variables into indices (e.g. well-being). This kind of data reduction or ranking becomes very useful when combined with a method of detecting spatial clusters — such as Openshaw's Geographical Analysis/Explanation Machines (Openshaw and Openshaw, 1997) — to locate and explain blackspots or areas of high risk/vulnerability.

5.2. GIS and its role in interdisciplinary research

The functionality of GIS can be classified into three levels:

1. The use of GIS to do simple things that we have always done.
2. The use of GIS to do complex things that we seldom or never do.
3. The use of GIS to do new things that revolutionise thinking and create new hypotheses (Arnold and Appelbaum, 1996).

The techniques herein ('level 2' functionality) 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, there is 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. The techniques are neither perfect nor all encompassing (e.g. the GWR application did not address spatial autocorrelation, although it could be included) but they do address some of the important issues and problems that plague many spatial analysis applications.

As Gardener (1998) states, the identification of appropriate scales for analysis and prediction is an interesting and challenging problem. Although the factors producing scale-dependent patterns may not be clearly understood, accurate and reliable descriptions of scale-dependent patterns and processes are required, to design data sampling procedures and test the accuracy and reliability of methods of prediction. There is clearly some way to go before scale effects can be fully understood and accommodated, but this research has aimed to be a 'next step' in that process.

6. Conclusions

Scale effects make any single or multivariate analysis of aggregated spatial data highly suspect, and 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 presented here.

The existing methodology of optimised regionalisation has been complemented with a new concept based on a spatially explicit cost–distance function to

generate regions that can adapt dynamically. It is able to represent a wide range of factors and still remain easy to use and intuitive.

Explicit identification of the role of spatial structure of socio-economic as well as traditional biophysical factors was made possible by the novel combination of a GWR model and an unsupervised neural-network. There was considerable variation in the parameters across the country that could be interpreted at regional and national levels. The parameters from the GWR model exhibited a strong spatial structure that was subsequently revealed by the SOM algorithm and associated mapping. The SOM algorithm generated an intuitive map of seven agricultural regions that were defined by their predominant farm types. The groupings in the map were generated with no a priori decision on the number of groups to be expected or desired. The output groupings were consistent with prior knowledge of the regions. The outputs of the combined GWR model and SOM algorithm have the potential to:

- Quantify and visualise the spatial drift within a data set.
- Produce better representations and hence understanding of local phenomena where spatial variation is found to be significant.
- Help to generate new hypotheses and experiments based on this understanding.
- Automatically define regions and system boundaries for further analysis.
- Guide the refinement of model specifications for multivariate data analysis.

7. Online references (software, documentation, technical reports and colour figures)

Colour figures for this paper, project documents and technical reports

http://www.ciat.cgiar.org

Geographical Analysis Machine (GAM)

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

Geographically Weighted Regression (GWR)

http://www.ncl.ac.uk/∼ngeog/GWR

Zone DEsign System (ZDES)

http://www.geog.leeds.ac.uk/pgrads/s.alvanides/zdes3.html

Self-organising Map Package (SOM_PAK)

http://www.cis.hut.fi/research/som-research

Acknowledgements

The project was co-ordinated by CIAT, but would not have been possible without the collaboration of the Royal Agricultural College, University of Florida, University of Georgia and Leeds University.

The Ecoregional Fund generously provided the funding for this research to support Methodological Initiatives, a Dutch Government financed program, administered by the International Service for National Agricultural Research (ISNAR).

Stan Openshaw, Phil Rees, Ian Turton and the CCG research group, University of Leeds, Stewart Fotheringham, Chris Brunsdon and Martin Charlton, University of Newcastle, Teuvo Kohonen and the CIS research group, Helsinki University of Technology, Joseph Wood, University of Leicester are thanked for making many of the data analysis and spatial analysis tools used in this research was freely available via the Internet. The author thanks the four reviewers for their constructive criticism and helpful suggestions on draft versions of this paper.

References

Arnold, C.G., Appelbaum, R.P., 1996. The use of GIS to measure spatial patterns of ethnic firms in the Los Angeles garment industry. In: Aldenderfer, M., Maschner, H. (Eds.), Anthropology, Space and GIS. Oxford University Press, Oxford, pp. 44–54.

Fischer, M.M., Henk, S.J., Unwin, D., 1996. GIS, spatial data analysis and spatial modelling: an introduction. In: Fischer, M., Scholten, H., Unwin, D. (Eds.), Spatial Analytical Perspectives on GIS. Taylor & Francis, London, pp. 3–20.

Fotheringham, A.S., 1997. Trends in quantitative methods. I. Stressing the local. Prog. Hum. Geog. 21 (1), 88–96.

Fotheringham, A.S., 1998. Trends in quantitative methods. II. Stressing the computational. Prog. Hum. Geog. 22 (2), 283–292.

Fotheringham, A.S., Wong, D.W.S., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ. Plann. A 23, 1025–1044.

Fotheringham, A.S., Charlton, M., Brunsdon, C., 1997. Measuring spatial variations in relationships with geographically weighted

regression. In: Fischer, M., Getis, A. (Eds.), Recent Developments in Spatial Analysis. Springer, Berlin, pp. 60–82.
Gardener, R.H., 1998. Pattern, process and the analysis of spatial scales. In: Peterson, D.L., Parker, V.T. (Eds.), Ecological Scale. Colombia University Press, New York, pp. 17–34.
Gehlke, C., Biehl, K., 1934. Certain effects of grouping upon the size of the correlation coefficient in census trade material. J. Amer. Statist. Assoc. 29, 169–170.
Griffith, D., Amrhein, C., 1997. Multivariate Statistical Analysis for Geographers. Prentice Hall.
Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge.
Hobbs, R.J., 1998. Managing ecological systems processes. In: Peterson, D.L., Parker, V.T. (Eds.), Ecological Scale. Colombia University Press, New York, pp. 459–484.
Jelinski, D.E., Wu, J., 1996. The modifiable areal unit problem and implications for landscape ecology. Landscape Ecol. 11 (3), 129–140.
Johnston, R.J., 1991. Geography and Geographers: Anglo-American Human Geography Since 1945. Edward Arnold, London.
Kaski, S., Kohonen, T., 1996. Tips for processing and color-coding of self-organising maps. In: Deboeck, G., Kohonen, T. (Eds.), Neural Networks in Financial Engineering. World Scientific, Singapore, pp. 195–201.
Kohonen, T., 1997. Self-organising Maps. Springer, Berlin.
Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P.C., 1998. People and Pixels. Linking Remote Sensing and Social Science. National Academy Press, Washington, DC.
Martin, D., 1998. Optimising census geography: the separation of collection and output geographies. Int. J. Geog. Inf. Sci. 12 (7), 673–685.
Morain, S., 1999. GIS Solutions in Natural Resource Management. Balancing the Technical-Political Equation. Onward Press, Santa Fe.
O'Neill, R.V., King, A.W., 1998. Homage to St Michael; or, why are there so many books on scale? In: Peterson, D.L., Parker, V.T. (Eds.), Ecological Scale. Colombia University Press, New York, pp. 3–16
Openshaw, S., 1984. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography 38. GeoBooks, Norwich.
Openshaw, S., 1996. Developing GIS-relevant zone-based spatial analysis methods. In: Longely, P., Batty, M. (Eds.), Spatial Analysis: Modelling in a GIS Environment. GeoInformation International, London, pp. 55–73.
Openshaw, S., Clarke, G., 1996. Developing spatial analysis functions relevant to GIS environments. In: Fischer, M., Scholten, H., Unwin, D. (Eds.), Spatial Analytical Perspectives on GIS. Taylor & Francis, London, pp. 21–38.
Openshaw, S., Openshaw, C., 1997. Artificial Intelligence in Geography. Wiley, London.
Openshaw, S., Rao, L., 1995. Algorithms for reengineering 1991 census geography. Environ. Plann. A 27 (3), 425–446.
Openshaw, S., Taylor, P.J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley, P. (Ed.), Statistical Applications in the Spatial Sciences. Pion Press, pp. 127–144.
Peterson, D.L., Parker, V.T., 1998. Ecological Scale. Colombia University Press, New York.
Veldkamp, A., Fresco, L.O., 1996. CLUE: a conceptual model to study the conversion of land use and its effects. Ecol. Model. 85, 253–270.
Wood, J.D., 1996. The geomorphological characterisation of digital elevation models. Ph.D. Thesis. University of Leicester.