



# Spatial structure and multivariate analysis in hillside agro-ecosystems

Andy Nelson, 2000

## Summary

### **Objective**

*Test the significance of spatial structure in hillside agro-ecosystem characterisation.*

An outstanding methodological issue relates to incorporating space or location as an explanatory variable, equivalent to more traditional variables. Recent efforts have used an approach based on multivariate statistics and factor analysis suggesting dynamic spatial modelling across geographic scale can be incorporated into more traditional dynamic temporal analysis. Location however is not a factor that has been routinely nor accurately described. Given that structural heterogeneity is widely accepted among ecosystem ecologists, acceptance of agricultural systems applications 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.

This report focuses on a relatively new methodology called Geographically Weighted Regression (GWR). GWR permits a calibrated regression model to vary spatially such that spatial drift from the global relationships can be measured directly. Furthermore the parameter variations can be mapped across space to improve understanding of the processes being modelled and to reveal system structure and system boundaries.

GWR and OLS models are compared with an example using agro-economic variables at a national level in Honduras. The regression parameters from a calibrated GWR model are analysed by a Self Organising Map (SOM), to discover the trends and patterns in the parameter variations across Honduras. The patterns are interpreted as a set of distinct agricultural systems.

The final output of the analysis suggests that this, and similar methodologies have huge potential for revealing multivariate structure that would otherwise prove difficult to visualise or understand.

## Introduction

The increasing availability of large and complex spatial data sets has led to a greater awareness that the global and whole map statistical methods are of limited application, and that there is a need to understand local variations in more complex relationships. In response to this recognition, 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.

Regression is the most commonly applied method for multivariate analysis, and is the focus for this report. However, if regression is to be considered as an analytical framework for discovering the spatial structure within complex agro-ecosystems then it is useful to reiterate the 7 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 (no perfect multicollinearity)
7. The error term is normally distributed

This places fairly restrictive limitations on what can and cannot be reasonably be expected of an OLS regression. With that in mind, we can contrast the seven assumptions with a list (table 1) of some of the characteristics of spatial data and the consequences that they have for regression modelling.

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. For instance, spatial autocorrelation immediately invalidates the application of methods based on the general linear model (GLM) such as principle components, factor analyses and regression models. Johnston (1997) observed it is somewhat paradoxical that spatial autocorrelation reflects the order that geographers are seeking to establish with laws and theories and yet its presence prevents its identification in the accepted technical models.

Since this realisation (in the 1950's) two distinct branches of research have appeared.

1. The continued application of these models, either in ignorance of the autocorrelation case or on the 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.



2. 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.

There are of course many other forms of regression (LOESS, median, and non-linear to name but few), but the purpose of this report is not to validate the latest methods and tricks included in the most recent version of SPSS or SAS. Rather it is to assess the possibilities for applying standard multivariate techniques to spatial data, and to present some novel approaches that can augment such analyses.

This report presents:

1. A method for accommodating spatial variations within a regression analysis framework.
2. A method for detecting (ecoregional) system boundaries within a multivariate analysis of agricultural variables.

## Regression models for spatial data

Of the problems listed in the previous section, there are three that require direct intervention in the design of a multivariate analysis for spatial data. In order of importance, or arguably intractability, they are:

1. The variations in relationships and hence processes over space, which is termed spatial non-stationarity or spatial drift.
2. The spatial dependencies across space and across variables, which is termed spatial autocorrelation.
3. That geographical data rarely has a linear distribution.

Here we briefly review the options for dealing with non-linearity and autocorrelation, and then we describe a model that accounts for the spatial 'drift' in linear relationships. If non-stationarity can be dealt with in a regression model then it can then be further adapted to address the two lesser difficulties of non-linearity and autocorrelation.

### **Non-linear, non-parametric and wavelet regression models**

Given that we have accepted that the world does not exhibit a great deal of linearity, non-linear regression initially seems an attractive possibility. However there are two problems here. Firstly the model specification is almost arbitrary in that there are an infinite amount of non-linear functions we can adopt, and without substantial prior knowledge, there is difficulty in justifying one function over another. This leads to the model-driven hypotheses where a model is formulated first and then validated or falsified by means of observed data. Secondly there is the difficulty of interpretation; for example, how do



we decipher statements such as *there is a high correlation between the arcsine of log population density and water supply raised to the power of 0.335*.

If linearity is rare and non-linear models are difficult to justify and interpret, the next recourse is to non-parametric regression, where any parametric model that reasonably fits the data may be useful. There are a several possibilities with kernel estimation and smoothing splines being two of the most popular. Again each method is either mathematically complex or relies on a set of pre-defined parameters which are difficult to define without prior knowledge. Kernel estimation however does have some advantages when used in exploratory spatial data framework (Wand and Jones 1996), although the possibilities for statistical testing and further inference are minimal.

Wavelets are a relatively new mathematical technique for pattern identification and have been used to great success in image analysis. There is evidently great potential for this method, although again, the interpretation and predictive ability of the outputs are questionable. It has mainly been applied to image data, and further investigation is required for point and zonal data applications.

### **Mixed regressive spatial autoregressive models**

A regression model can contain a spatial autoregressive term, that is, a weighted sum of the values of the dependent variable at other locations. Again such a model has two difficulties. Firstly there is the assumption of a global autocorrelation function for the error terms. Activity 2.1 of this research highlights the variation of autocorrelation across even small regions for key eco-regional variables. Secondly there is the need to determine a suitable weights matrix for the autocorrelation function. In other words, the analyst is expected to know *a priori*, the degree of all interactions between all locations across the region, when really it is this spatial structure that we wish to discover, not pre-impose! The proper choice of a spatial weights matrix is an important problem, and one that has not yet obtained a satisfactory solution, especially when areal units are being studied.

### **Modelling non-stationarity in regression models**

The problems of spatial dependency and spatial non-stationarity are obviously related. If spatial varying relationships are modelled within a global framework such as a standard regression then the error terms in the global model will exhibit spatial autocorrelation, as the model is unable to deal with the spatial drift of the relationship being measured.

Other efforts to produce a spatially acceptable regression models have been reviewed in Brunson et al, (1998), and their discussion is included here in full.



The most widely known is the expansion method (Casetti, 1972; Jones and Casetti, 1992) which attempts to measure parameter 'drift'. In this framework, parameters of a global model are expanded in terms of other attributes. If the parameters of the regression model are made functions of geographic space, trends in parameter estimates over space can then be measured (Eldridge and Jones, 1991; Fotheringham and Pitts, 1995). While this is a useful and easily applicable framework in which improved models can be developed, it is essentially a trend-fitting exercise in which complex patterns of parameter estimates will be missed. The output from spatial variants of the expansion method is thus a second-order set of relationships when what is required is information on the first-order relationships.

Four other statistical methods have been proposed to handle varying parameter estimates although only one of these, geographically weighted regression (GWR), has been developed specifically for spatial data (Brunsdon et al., 1996; Fotheringham et al., 1997a; 1997b). The other three are:

1. Spatial adaptive filtering (Foster and Gorr, 1986; Gorr and Olligschlaeger, 1994).
2. Random coefficients model (Aitken, 1996).
3. Multilevel modelling (Goldstein, 1987).

The first incorporates spatial relationships in a rather ad hoc manner and produces parameter estimations that cannot be tested statistically, resulting in it finding very limited applicability. In the latter two techniques, the parameter estimates in a regression model are assumed to be random variables. In the random coefficient model, the parameter estimates are modelled as finite mixture distributions, while in multilevel modelling the distribution of the estimates is assumed to be Gaussian. In both cases, by using Bayes' theorem it is possible to obtain local parameter estimates although no assumption is made about any spatial dependency in relationships that seems unrealistic in the light of what we know about spatial processes. Although geographical variations of multilevel modelling have been attempted (Jones, 1991), they rely heavily on a predefined hierarchy of spatial units in which the continuous nature of space is essentially ignored.

GWR, on the other hand, is a relatively simple yet powerful technique, which extends the traditional regression framework by allowing local rather than global parameters to be estimated. To calibrate the model, a modified weighted least-squares approach is taken so that the data are weighted according to their proximity. There are parallels between GWR, kernel regression and kernel density estimation (Parzen, 1962; Cleveland, 1979; Silverman, 1986; Cleveland and Devlin, 1988; Brunsdon, 1991; 1995; Wand and Jones, 1995). In kernel regression,  $y$  is modelled as a non-linear function of  $x$  by weighting data in attribute space rather than geographic space. That is, data points closer to  $x_i$  are weighted more heavily than data points further away and the output is a set of localised parameter



estimates in  $x$  space. It should be noted that as well as producing localised parameter estimates, the GWR technique described above will produce localised versions of all standard regression diagnostics including goodness-of-fit measures such as  $r$ -squared. The latter can be particularly informative in understanding the application of the model being calibrated and in exploring the possibility of adding additional explanatory variables to the model.

Brunsdon et al. (1996) and Fotheringham et al. (1997b) consider the choice of the spatial weighting function in the GWR framework and also demonstrate how the function can be calibrated. The calibration generates a weighting function that provides information on the scale of a particular process and it can be adapted spatially so that regions that have less information correspond to a less steep weighting function.

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 parameters we can impose boundaries where a parameter 'flips' between a positive and negative contribution and overall we can delineate the regions where the model does not provide a good fit and then consider another set of dependent variables based on the sign changes in the parameters for these regions. The next section explains the GWR model and relates it to a standard OLS regression.

## Methodology

Starting with the standard regression model, a vector of dependent variables  $\mathbf{Y}$  are related to a matrix of independent variables  $\mathbf{X}$ , as follows (Eq.1)

$$y_i = \sum_j x_{ij} \mathbf{b}_j + \mathbf{e}_i \quad \text{Eq. 1}$$

Where  $\hat{\mathbf{a}}$  is a vector of regression coefficients and  $\hat{\mathbf{a}}$  is a random vector whose distribution is

$$N(0, \mathbf{s}^2, 1) \quad \text{Eq. 2}$$

From this is it simple to extend to the GWR model (Eq. 3)

$$y_i = \sum_j x_{ij} \mathbf{b}_j(u_i, v_i) + \mathbf{e}_i \quad \text{Eq. 3}$$



Where  $(u_i, v_i)$  is the location in geographic space of the  $i$ th observation. If  $\hat{\alpha}_j(u_i, v_i)$  is constant for all  $(u_i, v_i)$  then the OLS model of Eq. 1 holds.

The method used to calibrate Eq. 3 is a moving window approach where at each location a weighted regression is performed and each observation carries a weight  $w_{ki}$ , where  $w_{ki}$  is a monotonically decreasing function of the distance  $d_{ik}$  between the points  $(u_i, v_i)$  and  $(u_k, v_k)$ . The weighting function is typically of a Gaussian distance decay type where

$$a_{ik} = \{1 - (d_{ik} / h)^2\}^2 \quad \text{Eq. 4}$$

And  $h$  is the 'bandwidth' or radius of the circle of influence around each observation. The choice of bandwidth is crucial if the best fitting model is to be found, however running the model with a range of bandwidths is a very useful technique for assessing scale dependence and discovering levels of organisation within a set of independent variables. Should an optimum bandwidth be desired then a least squares cross-validation method for choosing  $h$  automatically can be applied. Thus GWR produces a localised set of parameters representing the spatial drift in the system

## Modelling spatial variation in a hillside agro-ecosystem

There are two hypotheses that may be tested with respect to a GWR model that extend its utility beyond exploration.

1. Does the GWR model describe the data better than a global regression?
2. Does the coefficient  $\hat{\alpha}_{ij}$  vary over space for a given  $j$ ?

In the first case, the bandwidth parameter is helpful. Global models will have an optimal bandwidth close to the regions extent, and for models with local trends this will not be the case. In the second case, for each sample point  $i$ , the parameter estimate  $\hat{\alpha}_j$  is computed, giving  $N$  estimates of the coefficient under study, for which we can calculate the standard deviation, referred to as  $\hat{\sigma}_j$ . In this way we can detect the divergence of the system variables from the global linear model.

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



another report hence the justification, background and variable derivations are not repeated here. The variables to be examined are:

### **The data**

Each variable has been calculated at the aldea level (figure 1). The spatial distribution of the dependent variable (average production per worker) is shown in figure 2 as an interpolated surface. Note that this interpolation is for visualisation purposes and does not play any part in the model. There is a strong spatial trend 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.

Coefficients of correlation were calculated for each variable, and are presented in Table 2, with only two pairs of variables showing some degree of correlation. The same data is represented as a scatter plot diagram in figure 3.

The independent variables are mapped individually in figure 4 (again as interpolated surfaces), with their legends explained as follows:

Months of rainfall (most rainfall in the north and east, driest regions in the south west)

Rural population density (generally low with a few hotspots of high density)

Accessibility to the nearest market (the eastern region is notable for its lack of infrastructure)

Accessibility to the two major ports (on the Pacific and Caribbean coasts)

Education level (generally low for the entire country)

Gini coefficient (generally high for the entire country)

Technical assistance or extension (high in the east and south west)

Use of credit (again high in the east and south)

Temporary labour available in the village (high values in the western and central regions)

New seed adoption (a similar pattern to accessibility to market)



### OLS Regression results

Given that there is little collinearity in the independent variables an OLS regression is applied, with the following results.

Dependent mean	=	4.22410107	
Number of observations	=	3212	
Number of predictors	=	10	
Error Sum of Squares	=	0.1773	
Error Degrees of Freedom	=	3201	
<b>Root Mean Square Error</b>	=	<b>0.371</b>	

  

<b>parameter</b>	<b>estimate</b>	<b>std. error</b>	<b>t. test</b>
-----	-----	-----	-----
Intercept	+3.843	0.0377	101.5
RAINM	+0.050	0.0020	25.3
ACCESS_P	+0.003	0.0002	14.4
GINI	-0.400	0.0353	-11.5
SCHOLAR	+0.014	0.0012	11.6
TECNI	+0.001	0.0001	3.8
CREDIT	+0.001	0.0001	4.8
LABTEMP	+0.009	0.0023	3.9
SEED	+0.003	0.0002	12.1
POPDENS	-0.014	0.0011	-12.4
ACCESS_T	-0.002	0.0004	-4.8

All variables are significant (at the 1% confidence limit) and the parameter estimates have the signs that we would expect (positive for rainfall, education, credit, technical assistance, labour and improved varieties, and negative elsewhere). However the fit (0.377) is not great, although it is not too bad. Could a GWR model improve the fit and would its application be justified?



### GWR Regression results

A GWR model was applied to the data, and a cross-validation procedure used to determine the optimum bandwidth, which in this case was 51km (bandwidth shown in figure 2). The results of the GWR model are summarised below where the minimum, quartile, median and maximum values for each parameter are presented for each parameter, as the model varies across the region. Clearly the total variation for each local parameter lie well outside the range of their global values and standard errors. Every parameter (except the intercept) changes sign across the region, and these deviation from the global parameter are visualised in figure 5, where blue regions are below the global value and red values are above. Darker colours indicate greater variation from the global value (in white).

Geographic Weighted Regression Bandwidth 51km					
Label	Minimum	L.Quartile	Median	U.Quartile	Maximum
Intercept	+0.788	+3.962	+4.363	+4.542	+6.423
RAINM	-0.229	-0.001	+0.011	+0.033	+0.290
ACCESS_P	-0.007	-0.003	-0.000	+0.003	+0.016
GINI	-0.675	-0.532	-0.419	-0.314	+1.149
SCHOLAR	-0.059	+0.008	+0.010	+0.012	+0.049
TECNI	-0.003	+0.000	+0.001	+0.001	+0.003
CREDIT	-0.003	+0.000	+0.001	+0.001	+0.004
LABTEMP	-0.048	+0.004	+0.010	+0.016	+0.081
SEED	-0.026	+0.002	+0.002	+0.003	+0.073
POPDENS	-0.118	-0.021	-0.016	-0.013	+0.014
ACCESS_T	-0.010	-0.001	+0.001	+0.004	+0.007

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. 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, hence creating surfaces with gradual trends.

### Comparing the two models

Report 2.1 (example application 2) presents the results of a cross-scale correlation analysis between agricultural income and three independent variables (rainfall, improved varieties and population density). There are clearly local patterns in these relationships that contradict the global pattern. This is the first



evidence for suggesting that the global regression model is an inefficient estimator in many regions of Honduras, as the relationships between the dependent and independent variables change across space.

Firstly we can compare the distribution of the percentage errors in both models with a histogram, (figure 6). Both models exhibit a normally distributed error term, with a narrower distribution visible in the GWR model. The GWR model has a tendency to overestimate the agricultural income compared to the underestimation in the OLS model

We can map the residuals (figures 7 and 8) from both models to see which geographical location relates to which residual, hence we can detect any spatial autocorrelation of the errors, and visualise the differences between the two models in terms of predictive accuracy.

Both models exhibit similar patterns of errors, both with string autocorrelation, indicating that some kind of auto-regressive function would be appropriate, although as stated earlier, autocorrelation is not the most serious issue for modelling spatial data. The slight improvement of the GWR model that was visible in the histogram of errors is more clearly evident here.

A further comparison of the two models can be achieved via an ANOVA test (based on the methodology in Fotheringham et al, 1999). Table 3. This reveals that there is a reduction in the residual sum of squares when the GWR is used; this implies that the null hypothesis of the OLS model should be rejected.

The OLS model had an overall  $R^2$  value of 0.37, which we can compare with a map of the  $R^2$  from the GWR model (figure 9). The shading indicates where the GWR was able to improve on the global fit (light green for slightly better, darker green for substantial improvements). Regions in red or orange are where the GWR model could not improve upon the OLS fit, and regions in yellow are where both models provided similar ( $\pm 1\%$ ) fits. The blue line highlights those regions where the GWR model provided an improved fit to the data.

This implies that the parameters in the regression model provide a reasonable (0.4 – 0.6) fit for the regions of Olancho and EL Paraiso and a good fit (0.7+) for the Mosquita region and Colon. The northern coast and western borders with Guatemala and EL Salvador however are not well represented by the chosen parameters, and perhaps a new model should be proposed for these area. In this way, a study region can be divided into two or more sub-regions, each with its own set of regression parameters.



From these comparisons, there is a strong argument for using a GWR model for explaining the relationship between agricultural income and several agro-economic variables. The OLS  $R^2$  value and GWR improvement of the model fit should be viewed in light of:

1. The model contained many variables, making a high  $R^2$  value unlikely.
2. The high resolution of the data, implying considerable noise and error in the data.
3. Several proxy variables were used including the dependent variable.

### Summary

The two hypotheses that were presented in this application have been verified. The GWR model does describe the data better than a global regression, and the coefficients  $\hat{\alpha}_j$  do vary over space for a given  $j$ ?

## Defining system boundaries and spatial structure

### Interpretation of GWR results

The most interesting outputs from the GWR model are the parameter surfaces, which essentially characterise the inherent spatial structure in the various agro-ecosystems present in Honduras. In this case, with variables such as income per capita, population density, access and technical assistance, the surfaces could also indicate regional 'pathways of development' or common patterns of resource management.

Interpretation of regression results is often difficult, biased, subjective and intractable, especially when huge databases covering large heterogeneous areas are being used! However given that there are considerable variations in the parameter estimates across the country, then some form of interpretation or at least discussion would be interesting for two reasons. Firstly the variation in parameters and overall fit imply that several agricultural systems are present. Secondly that the range of the variations (with every parameter exhibiting positive and negative contributions) and their strong patterns would help in the delineation of system boundaries. Here we introduce a novel method that can help interpretation and suggest system boundaries based on the regression parameters.

### Self-Organising Maps for spatial pattern detection

The parameter surfaces are interpolations based on the 3500 aldea locations. Hence each aldea has a set of 11 (intercept value and 10 independent variables) parameters. If there are distinct systems, i.e. two or more regions with contrasting parameters, then we can discover these by applying some form of classification or clustering to the parameters. Any spatial structure or system boundaries would be visible if there is some degree of clustering in the parameters and in their spatial distribution.



A vast number of different algorithms to perform clustering are available. Choosing a suitable algorithm and applying it correctly requires thorough knowledge of both the algorithms and the data set. There must exist enough clustering tendency in the data set in order that the use of clustering algorithms would be sensible at all, and as different algorithms tend to find clusters of different shapes, the suitability of the shapes to describe the data set must be verified.

The Self Organising Map (SOM) has been presented in report 2.2, as a suitable method for representing the inherent structure in multivariate data; only here we are not exploring the multivariate data themselves but their significance within a region.

### **Applying the SOM**

The 11 parameters for 3500 locations were normalised and a SOM structure of 24x12 neurons was used to represent the data structure. Training took over one hour on a PC laptop and a very low quantisation error of 0.8 was achieved. The output map is shown in figure 10 where similar colours indicate clustering and rapid changes in colour represent the different patterns within the parameters.

Clearly there are strong patterns and trends within the parameter estimates. In report 2.2 it was possible to overlay country codes onto the map giving an immediate impression of the geographic patterns in the data, but although each aldea has a unique code, it would be impossible to overlay so many codes on to the map. Instead the colour codes were applied directly to a map of aldea boundaries (figure 11). Areas in grey contain no data, representing aldeas where there were insufficient data. The aldea boundaries have been simplified to speed up the drawing of the map (due to current limitations of the in house software). Figure 12 shows the same data without the aldea boundaries.

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 blue areas on the eastern coast, but in general, the groups are compact and contiguous. Table 4 lists the seven groups, their locations, and the main agriculture associated with each region.

The 7 groups relate well to the predominant farming trends and socio-economic trends that are known to exist in Honduras, although the grouping of Atlantida and a large swathe of south-western Honduras seems counter intuitive. With these boundaries in place an attempt at interpretation can be made. Here we briefly describe the parameters associated with two of the seven groups.

### **Light blue - Hillside region**



Productivity in the central hillsides is positively affected by rainfall and little else. Population density, although not very significant has a constantly positive effect. High population density can lead to increased use of marginalised land (negative impact on income) or intensification of land use, if technical assistance and labour are available (positive impact).

This region does indeed have relatively high levels of assistance, credit, improved varieties and labour, and the income level is reasonable. Extension, intervention and access to technology do seem to have impact in the Honduran hillsides, although the effect is by no means consistent. Whereas the eastern and western regions of Honduras have fairly consistent parameter values, the hillsides region exhibits strong variation for all parameters except rainfall and population density.

### **Green and Purple - Southwestern region**

The south west is characterised for its long dry season and a high percentage of exceptionally poor land, where subsistence farming is the predominant form of agriculture. There is also huge NGO involvement in this zone, implying technical assistance through extension. The parameters suggest that the little income that is generated is related to the educational level (perhaps their willingness to adopt new measures?) of farmers and access to very local markets rather than intervention or assistance. Accessibility and transport availability in this zone is also very poor, and the results suggest that improvements in infrastructure development could lead to a greater involvement of the rural population in the local economy. It is this region that contains the largest errors in both models, implying that the most impoverished type of smallholder is not well represented by the specified model. Furthermore the SOM patterns exist over a small and contained region.

### **General**

There are also country wide patterns in the parameter variations. As far as general trends are concerned there are a few interesting and intuitive patterns.

1. There is an inverse relationship between education and temporal labour, suggesting that labour costs are greater where literacy is higher.
2. There is an almost perfect polarity between access to ports and access to local markets. Regions are dependent on one or the other, not both.



## Conclusions

### Outputs compared with a restatement of the objective

*Test the significance of spatial structure in hillside agro-ecosystem characterisation.*

Explicit identification of the role of spatial structure of socio-economic as well as traditional biophysical factors was made possible by the combination of a geographically weighted regression model and a non-parametric data classifier.

The example presented here, incorporating a range of key agro-ecosystem variables, has indicated that there is a significant degree of spatial variation both within Honduras and within the hillside region. It was also shown that a regression framework that incorporated spatial drift could offer significant improvements over OLS regression models as the residual errors were reduced and local patterns that contradicted the global model were revealed.

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 GWR model and SOM algorithm have the potential to:

- Quantify and visualise the spatial drift within a dataset.
- Produce better representations and hence understanding of local phenomena where spatial variation is found to be significant.
- Help 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.

## Future work

### GWR

The GWR framework is a relatively simple adaptation of the classical OLS model, which lends itself to further improvements such as:



1. Developing a spatially auto-regressive GWR, where the choice of bandwidth could play a role in the design of the troublesome spatial matrix.
2. Including more robust measures of parameter significance.
3. Generalising the procedure to a locally adaptive measure of variance and other statistical measures.

Also, GWR could be run in a purely exploratory fashion by repeating the process at several scales where the rate of change of parameters across scale would indicate regions of scale dependence.

## Online References for Geographically Weighted Regression

### Homepage

<http://www.ncl.ac.uk/~ngeog/GWR/>

### Software.

<ftp://ftp.ncl.ac.uk/pub/users/nmec>

### Useful References to GWR

Fotheringham, A.S., Brunson, C., and Charlton, M.E., 2000, Quantitative Geography, London: Sage

Brunson, C., Fotheringham, A.S., and Charlton, M.E., 1999, Notes on parametric significance tests for geographically weighted regression, Journal of Regional Science, 39(3), 497-524

Fotheringham, A.S., Brunson, C., and Charlton, M.E., 1998, Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis, Environment and Planning A, 30(11), 1905-1927

Brunson, C., Fotheringham, A.S., and Charlton, M.E., 1998, Geographically weighted regression - modelling spatial non-stationarity, Journal of the Royal Statistical Society, Series D-The Statistician, 47(3), 431-443

Brunson, C., Fotheringham, A.S., and Charlton, M.E., 1998, Spatial nonstationarity and autoregressive models, Environment and Planning A, 30(6), 957-993

A.S. Stewart Fotheringham, 1997, "Trends in Quantitative Methods I: Stressing the Local" Progress in Human Geography, 21: 88-96

C. Brunson, A. Stewart Fotheringham and M.E. Charlton, 1996, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity", Geographical Analysis, 28(4), 281-298

A.S. Stewart Fotheringham, M.E. Charlton and C. Brunson, 1996,



"The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity",  
International Journal of Geographic Information Systems, 10: 605-627

## Online References for Self Organising Maps

*Please refer to report 2.2.*