Methods

COUNTRY-WIDE OLS REGRESSION

In the OLS multiple regression model, the dependent variable y (the chosen measure of poverty, here household expenditure) is statistically related to a set of N independent variables x as follows,

$$y_i = \beta_o + \sum_{j=1}^{N} x_j \beta_j + \varepsilon_i \quad \text{where } i = 1 \text{ to } M$$

where i is an index of the number of points (M) for which data are available, β_0 is the intercept, β_j are the beta-coefficients for each dependent variable, and ε is a randomly distributed error term. The reader is referred to standard texts on regression such as Draper and Smith (1988) for a fuller explanation of OLR and, for specific application to geographical problems, to Griffith and Amrhein (1997)

In addition to the beta-coefficients, the following statistics were calculated in order to evaluate model accuracy:

- The residuals at each location, and the residual sum of squares.
- The standard errors associated with each β term.
- A t statistic for each independent variable.
- A coefficient of determination statistic (R^2).

A single regression model was fitted to the aggregated household data and the environmental variables at those same locations across Uganda. The β coefficients were then applied to the predictor variables to extrapolate the relationships to estimate the average rural monthly per adult equivalent household expenditure in all pixels.

A series of diagnostic tests was also carried out to ensure that the model met the following assumptions, required for linear regression.

- Homoscedasticity – the variance of the error term must be constant for each value of y. To check this the residuals were plotted against y. Ideally, there should be no obvious pattern.
- No multicollinearity – no strong correlation should be observed among the independent variables. Bivariate collinearity was checked for with scatter plots and correlations between each pair of independent variables, and was assessed with a variance inflation factor test.
- Linearity – there should be a linear relationship between each independent variable and the dependent variable. This can be assessed with a scatterplot matrix for all variables. Non-linearity does not invalidate the OLS model but it does mean that the beta coefficients cannot fully capture the relationship. The dependent variable was transformed to ensure linearity, but the independent variables were not.
- Independence of error terms – successive residuals should not be correlated. A Durban-Watson statistic was used to check for such autocorrelation.

REGIONAL OLS

A single regression model is unlikely to capture the relationships between expenditure and the environment across an entire country, because the influence of each environmental factor, and the complex interactions among them, are likely to vary from location to location. Instead, a series of local or regional regression models might be more appropriate. Because of the potential importance of livestock, various aggregations of the livestock production system map were used to partition Uganda into zones for which to derive separate regression models. Only one of the six zones present in Uganda contained enough sample households to be treated independently, so the following three sets of aggregated categories were chosen:

- Three 'climate zones': (i) arid and semi-arid, (ii) humid and sub-humid and (iii) temperate or tropical highlands.
- Two 'farming systems': (i) livestock only and (ii) mixed crop and livestock.
- One 'dominant system': the mixed, humid and sub-humid system – the largest in Uganda.

The OLS coefficients from the first two can be combined to create country wide maps of predicted average rural monthly adult equivalent expenditure in all pixels. The 'dominant system' regression coefficients can only be extrapolated to pixels within that system. It is important to stress that within each of the bulleted headings above the categories are mutually exclusive (e.g. an area can be either 'livestock only' or 'mixed crop and livestock' only; it cannot be both, or a mixture of both). However there is considerable overlap between the bulleted headings since they form alternative ways of zoning what are frequently the same areas. Thus the regional OLS analyses using the three categorisations given above overlap considerably in the data points used (see later).

GEOGRAPHICALLY WEIGHTED REGRESSION (GWR)

An OLS regression model can be converted into a Geographically Weighted model by substituting each beta coefficient (the intercept and the dependent variable coefficients) with its local counterpart, such that the beta-coefficients can vary across space:

$$y_i = \beta_{o(u_i,v_i)} + \sum_{j=1}^{N} x_j \, \beta_{j(u_i,v_i)} + \varepsilon_i$$

where i = 1 to M, *and* (u_i,v_i) is the location in geographic space of the *i*th observation. A set of beta-coefficients (and hence a regression model) is estimated at each location based only on neighbouring, geographically weighted data points. Normally variables from data points farther away from the point in question (where $\mathrm{y_i}$ was measured) have lower weights than points closer to it and so contribute less to the regression results. If, however, $\beta(u_i,v_i)$ is constant for all (u_i,v_i) then the OLS model holds, i.e. OLS is a special case of GWR. The geographically weighted local counterparts of the residuals, standard errors, t and R^2 values (and any other associated statistic) can also be generated at each location.

Fotheringham *et al.* (2002) have developed GWR into a comprehensive statistical method. A key feature is the ability to calibrate the spatial weighting function to identify the bandwidth, i.e. the number of, or proximity of neighbouring points included, that results in a 'best-fit' model.

The estimated beta-coefficients at each location are dependent on the bandwidth and type of kernel (or weighting scheme) that is used in the model. Here, the bi-square kernel was used, defined as:

$$W_i = \begin{cases} \left(1-\left(\frac{D_i}{h}\right)^2\right)^2 f \quad D_i < h \\ 0 \quad f \quad D_a \geq h \end{cases}$$

Where W_i is the weight assigned to point *i* (from 0 to 1), *h* is the bandwidth and D_i is the distance from the centre of the kernel to point *i*.

The most appropriate bandwidth can be chosen by means of a cross-validation (CV) procedure where the model is run for a range of bandwidths and a least squares criterion is applied to find the bandwidth that minimises the sum of the squared errors between y and the estimated value of *y* (*y'*). The equation below states that for bandwidth *h*, y'_i is computed whilst omitting the data from point *i*. This omission of the central point means that when h is very small, the model is calibrated only on its neighbouring points and not on itself. If point *i* were not omitted the CV score would tend to zero as *h* tends to zero, (and hence the weighting for all points except *i* becomes negligible) and so y'_i would tend to y_i, at each location.

$$CV = \sum_{i=1}^{N} \left(y_i - y_i^{'}(h) \right)^2$$

The bandwidth can be defined in map units or as the number of data points (nearest neighbours) to include at each location. The use of map units, whilst ideal, can only be justified if the data points are evenly distributed over the study area. The survey households are not, so the bandwidth was determined in terms of number of nearest f neighbours.

Once the model has been calibrated and the best bandwidth identified, the GWR is re-run using the best bandwidth, and a series of computationally intensive tests is run to evaluate significance in spatial variation among the GWR parameters. GWR produces localised versions of the OLS regression outputs, so in place of a table of results summarising the beta coefficients, t values, standard errors etc., localised versions of these outputs are produced for each household or household pixel, and these can be mapped. The local R^2 and t values can be interpolated to give a visual representation of the goodness of fit of the model and to map areas where the coefficients are significant, but interpretation of these local statistics is not as straightforward as it is with their OLS equivalents. Furthermore, the GWR model can be applied to the remaining rural pixels in Uganda to create an estimated rural monthly adult equivalent expenditure map. The beta coefficients can also be mapped, to show the spatially varying relationships (non-stationarity) between poverty and the environmental variables included. Finally, these beta coefficient maps can be viewed as multiband images, or clustered, in order to identify regions with common spatial relationships between poverty and environmental variables.

WORKFLOW

The dependent variable - rural monthly adult equivalent expenditure – was first aggregated to match the finest resolution of the environmental variables, 0.01 degrees (approximately 1.1 km at the equator). These values were then transformed using a Box-Cox transform (Box and Cox 1964), resulting in a normally distributed dependent variable. This process was then repeated at a series of spatial resolutions matching those of the environmental data (broadly successive doubling of pixel dimensions, quadrupling their size).

A model was built, using the 0.01 degree data, relating expenditure to environmental conditions based on previous poverty mapping work and correlation analysis. The same variables were then used at all other (coarser) spatial resolutions, and any changes in the relative importance, sign and significance of the independent variables noted.

The above approach was applied to each of the methods used, OLS, regional OLS and GWR, providing regression results and maps of predicted average rural monthly adult equivalent expenditure. A bootstrap procedure was used to compute four goodness of fit metrics and their standard errors: a) Root Mean Square Error (RMSE); b) Mean Absolute Error (MAE); c) Mean Absolute Percentage Error (MAPE); and d) the R^2 of the observed versus the expected expenditure for each regression model (OLS, regional OLS and GWR) at all resolutions. These statistics helped to determine the resolution that provided the best trade-off between predictive accuracy and spatial precision. The same four metrics were further computed for the SAE expenditure maps at district, country and sub-country levels.

Finally, the GWR coefficients were mapped at the best resolution and the spatial variation in the coefficients was investigated.

All analyses were carried out in 'R' (V2.9.2)[3] (R-Development-Core-Team 2009) running on a 32bit version of Windows XP (SP3). The following R libraries were used:

- MASS (V7.2-48) - Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S'.
- car (V1.2-15) – Companion to Applied Regression.
- relaimpo (V2.1-2) - Relative importance of regressors in linear models – Non US version[4.]
- qpcR (V1.2-1) – Used for computing Akaike's Information Criterion.
- psych (V1.0-78) – Used for basic statistical summaries.
- gvlma (V1.0) - Global Validation of Linear Models Assumptions.
- spgwr (V0.6-2) - Geographically weighted regression.
- boot (V1.2-39) – Bootstrapping regression models to generate confidence limits and standard errors.

Libraries or packages and their dependencies can be installed and updated in R, on the command line. The R code used is available from the authors.

[3] http://cran.r-project.org and http://www.r-project.org

[4] http://prof.beuth-hochschule.de/groemping/relaimpo