Pro-Poor
Livestock
Policy
Initiative

Poverty Mapping in Uganda: An Analysis Using Remotely Sensed and Other Environmental Data

David Rogers, Thomas Emwanu, Timothy Robinson

PPLPI Working Paper No. 36

CONTENTS

Figures

Tables

PREFACE

This is the 36th of a series of Working Papers prepared for the Pro-Poor Livestock Policy Initiative (PPLPI). The purpose of these papers is to explore issues related to livestock development in the context of poverty alleviation.

In order to reduce poverty we must first describe, explain and predict its spatial distribution over large areas with as high a level of local accuracy as possible. Poverty maps are traditionally produced by exploiting links between census (wide area) and survey (smaller area coverage) data. The detailed relationships found within the survey data are extended to the census data that must share some predictor variables in common with the survey data. Both census and survey data tend to be socio-economic in nature; the mapping thus exploits the internal correlations within potentially strongly correlated data sets – one 'measure' of poverty is often correlated with another.

Rather than look at the correlates of poverty, we should like to identify its causes. We suggest that poverty is multi-dimensional and that many of its dimensions are environmentally related; people are poor because they are unhealthy, or under-fed, or without access to fuel and water *etc*. Each of these is environmental in some way or other, and a correct approach to reducing poverty might be first to identify its (environmental) causes. We have attempted to do this with survey data from Uganda and environmental data derived from multi-temporal satellite imagery that measures land-surface conditions and processes (temperature, rainfall, vegetation growth *etc*.). The same satellite data have already been used to understand the distribution of farming systems throughout Africa and to predict the distribution and intensity of insect and tick carriers of a variety of diseases, and the incidence and prevalence of the diseases they transmit.

In this analysis therefore we examined to what extent satellite data (as a proxy for environmental conditions) are correlated with household survey data. Whilst correlation obviously does not automatically imply causation, we suggest an environmental approach is more likely to reveal causes than will the traditional approach of small area mapping using census and survey data.

However, it is first necessary to establish the relative predictive accuracies of the traditional and environmental approaches. The initial results from the environmental approach, described here, are promising, though we have not yet compared them directly to small area methods.

We hope this paper will provide useful information to its readers and any feedback would be welcomed by the authors, PPLPI and the Livestock Information, Sector Analysis and Policy Branch (AGAL) of the Food and Agriculture Organization (FAO).

Disclaimer

Authors

David Rogers is Professor of Ecology at the University of Oxford, and heads the TALA Research Group. Two analytical components of the work reported here are the use of temporal Fourier-processing of satellite data, which reveals the all-important seasonal characteristics of the environment, and the discriminant analysis approaches that have been developed to model human and livestock disease and vector distributions. Both of these have been developed by the TALA Research Group.

Thomas Emwanu is Senior Systems Analyst at the Uganda Bureau of Statistics (UBOS). He is closely involved in the design, implementation and analysis of household surveys and censuses in Uganda and has worked closely with the World Bank and researchers at the Economic Policy Research Centre (EPRC), Makerere University and the International Livestock Research Institute (ILRI) in developing small area poverty mapping techniques for Uganda.

Timothy Robinson works for the PPLPI, based FAO's Livestock Information, Sector Analysis and Policy Branch (AGAL). He is responsible for the development of information systems in PPLPI and also for operations in the Horn of Africa, including the IGAD Livestock Policy Initiative, under which the work reported here will be carried forward.

Acknowledgements

The authors would first like to acknowledge the support and collaboration offered by John Male-Mukasa, Executive Director of the Uganda Bureau of Statistics (UBOS). John has been fully supportive in providing data and in encouraging Thomas to spend time on this work. We would also like to acknowledge Claudia Pittiglio for preparing much of the spatial data used in the analysis, and Federica Chiozza and Shannon Villicaña for producing the graphical outputs in this working paper.

Keywords

Poverty; welfare; livelihoods; policy; mapping; geographic information systems; remote sensing; multi-temporal satellite data; temporal Fourier processing; market accessibility; livestock; discriminant analysis; small area mapping.

Date of publication: 21 July 2006

EXECUTIVE SUMMARY

In order to target poverty reduction we need to understand, describe and explain its spatial distribution and be able to predict the degree and distribution of poverty in other regions and/or at other times. Poverty maps should therefore incorporate potential driving factors that in some way or other are associated with, and possibly even responsible for the different levels of poverty being mapped. In this analysis we start from two assumptions: a) poverty is a function of several interlinked factors, including agricultural activities, human and animal diseases, natural resources and other environmentally-determined factors; and b) many important characteristic of the environment can be described by earth-observation satellite imagery, through their ability to capture seasonal variations of a range of environmental factors.

In this analysis we explore a novel approach in which we combine household survey data with a suite of environmental variables that are either direct measures of key climatic variables (such as temperature), descriptor variables of key ingredients of poverty-generating processes (such as agricultural production systems) or proxies for constraints on the health and well-being of the human populations (such as disease-causing pathogens).

Predictions were made using a Discriminant Analysis model, in which a poverty index was estimated by the likelihood of each pixel falling within a specified "poverty" class, based on the combination of values of the predictor variables. The poverty data were derived from breakdowns of food expenditure from the 2002-2003 Ugandan National Household Survey, which covered 9,711 households in 973 communities. The predictor variables included available raster datasets: elevation, cultivated land, length of growing period, population distribution, livestock density, market accessibility (calculated as time to travel to a population centre of a certain size), and tsetse distribution; and a set of Fourier-transformed time series satellite data, derived from the Advanced Very High Resolution Radiometer, including the mean, minimum, maximum, variance, phases and amplitudes of parameters like Normalised Difference Vegetation Index, Land Surface Temperature, Air Temperature, Vapour Pressure Deficit and Cold Cloud Duration.

The analysis was performed at different spatial resolutions, ranging from 0.01 to 1 degree (approximately 1.1 km and 110 km at the equator). The overall model accuracy tended to increase with decreasing spatial resolution. Satellite-derived variables tended to dominate the list of selected variables that determine the predictions, but different predictor variables tended to be selected by the model at different spatial resolutions.

This method is appealing because it can produce estimates of the same poverty measures as those produced by the more traditional small area mapping methods, as well as an indication of the degree of statistical precision of the estimates. Work in progress will make direct comparisons between these two approaches. These preliminary results show that external, independent data appear to have at least as much descriptive power for poverty mapping as the internally correlated socio-economic data sets exploited by the small area estimates, though the precise interpretation of the correlations obtained here will require more research effort.

1. INTRODUCTION

Maps of poverty are of great use to government and development agencies alike for, no matter how such agencies wish to reduce poverty, they all need to have a 'base-line' picture of what it is they are supposed to be tackling. To realise their full potential, poverty maps should provide the following:

- a description of the spatial distribution of poverty indicators (targetting);
- an explanation for the observed spatial distribution of the poverty indicators (explanation); leading to ....
- an estimation of the degree and distribution of poverty in other regions and/or at other times and under changing conditions (prediction).

Most traditional ways of mapping poverty, and many new ones, often halt at the level of description. Various correlation and regression methods are applied to suites of socio-economic data, and the best correlations are exploited to make a spatially detailed map of whatever chosen index of poverty is preferred. By their very nature, socio-economic data are almost bound to be inter-correlated so, at best, all that these methods do is to exploit the internal correlations of a naturally correlated data set. What we end up with is an accurate description of the situation, but with no clear insight into the likely causes of the patterns we see, for the simple reason that the poverty-generating processes are neither implicitly nor explicitly included within the models themselves.

If the object of the exercise is not just to describe poverty, but also to understand it in order eventually to reduce it, then the above three steps must be undertaken in full. They should therefore incorporate potential driving factors that in some way or other are associated with, and possibly even are responsible for, the different levels of poverty being mapped. In other words, we have to look outside the socio-economic milieu in order to understand what precisely is going on within it.

Recently, excellent progress has been made using a variety of small area estimation techniques to increase the spatial resolution of descriptive poverty maps. In this approach, extensive census data (few variables, no measure of poverty) are combined with intensive socio-economic survey data (many variables, including chosen indices of poverty) in nested regression analyses that assume that the local degree of poverty is due to a combination of broad-scale regional phenomena (setting average poverty levels) and finer-scale local, or even household (idiosyncratic), level phenomena, coupled finally with an error term that often exceeds the summed total of the previous two terms. Brutally put, we end up with a relatively poor description of poverty, no explanation, and no clear idea of how to intervene to make a difference.

There seems, therefore, to be a need to move away from a static poverty mapping approach (a description of the poverty landscape) to a much more dynamic approach that attempts to reveal the underlying processes that produce the landscapes that we see.

In this analysis, we explore a novel approach in which we combine household survey data with a suite of environmental variables that are either direct measures of key climatic variables (such as temperature), descriptor variables of key ingredients of poverty-generating processes (such as agricultural production systems) or proxies for constraints on the health and well-being of the human populations (such as disease-causing pathogens). This potentially allows us to describe, explain and then predict the distribution of poverty at the highest spatial resolution of the key predictor variables. Through doing so we may be able to draw more realistic conclusions as to the likely causes of poverty.

Uganda has invested considerable efforts into measuring and mapping poverty. In 2003, poverty estimates were calculated for the year 1991, combining data on household consumption obtained from a 1992/93 Integrated Household Survey (HIS) and the 1991 housing and population census, using the small area estimation technique (Emwanu *et al*. 2003). More recently, the Uganda Bureau of Statistics (UBOS) conducted the Uganda National Household survey 2 (UNHS-2), from May 2002 to April 2003, which collected detailed data on expenditure for 9,711 households. It is these data that we have combined with remotely sensed and other spatial environmental variables in our exploratory analysis reported here.

2. MEASURING POVERTY

Human well-being has many dimensions and is perceived differently by different groups, so no single measure captures all aspects of it. Ravallion (1992) distinguishes materialistic measures, such as income and standard of living, from concepts such as 'opportunities' and the 'right to participate in society'. The International Fund for Agricultural Development (IFAD) has identified eight broad classes of poverty (Jazairy *et al*. 1992): i) material deprivation; ii) lack of assets; iii) isolation; iv) alienation; v) dependence; vi) lack of decision making power; vii) vulnerability to external shocks; and viii) insecurity. Poverty can be absolute, where individuals are unable to satisfy the minimum basic needs for survival, or it can be relative, where some function of a distribution of income or expenditure can be used to define a threshold level below which people are defined as poor.

Some efforts have been made to take a multidimensional approach to measuring poverty, for example the Priority Poverty Indicators (PPIs), developed by the World Bank, take into account measures such as nutritional status, life expectancy, under-five mortality and school enrolment rates, as well as income and expenditure. More usually, however, measures of human well-being focus on specific dimensions, such as material deprivation or levels of achievement in health or education. Most commonly, poverty measurement is based on material deprivation that is generally linked to the inability of incomes or expenditures to meet basic nutritional needs, as defined by a consumption-based "poverty line". In this way poverty rates can be estimated using head-count indices (the proportion of people below the poverty line), poverty gap ratios or severity of poverty indices (Malik 1998).

In simple economic terms (adapted from Ravallion 1996) we can define and measure poverty as follows. We first define a single monetary indicator of household welfare (y_i), *e.g.* total expenditure on consumption, or total income over some period. Next, we define a poverty line (z_i) as the cost to the *i*th household of escaping poverty. In general, the lower the value of y_i/z_i, the poorer the household. We can then generate some poverty index that incorporates the measured ys and zs, such as those in the widely used 'Foster-Greer-Thorbecke' (FGT) class of poverty indicators (Foster *et al*. 1984; Foster and Shorrocks 1988).

2.1 Welfare indicators (y)

Monetary estimates of income or consumption dominate assessments of poverty, and are certainly the only types of measure that are globally available. Moreover, compared to social indicators monetary estimates are relatively straightforward to standardise globally. Economists tend to use an estimate of current household consumption expenditures as a welfare indicator. This approach should include consumption of goods produced by the household, though the 'food basket' approach to assessing consumption expenditure fails to account for non-purchased food items, some of which (*e.g.* wild fruits, roots, blood) are difficult to estimate in market value terms. Per-capita consumption does not account for different household structures, though weighting schemes have been developed to estimate consumption 'per adult equivalent'.

2.2 Poverty lines (z)

Whether economic or social measures of poverty are chosen, thresholds must be determined that distinguish the poor from those that are not poor. In general, some food poverty line can be determined that defines the minimum requirement for survival (sometimes called the 'extreme poverty line' or the 'hardcore poverty line'). For example, in the food energy intake (FEI) approach, a monetary value is determined at which minimum food requirements can be met. The overall poverty line is then an estimate of food plus non-food household spending; this is referred to as the cost of basic needs (CBN) approach.

2.3 Poverty indices

Poverty indices are derived through functions that combine measured values of y and z in some manner appropriate to the application at hand. The simplest is the 'head count index': the proportion of total households classified as poor, *i.e.* for which incomes/expenditures are below the poverty line (y_i/z_i<1). Whilst the head count index is an intuitively simple indicator, and is good for making national comparisons, it cannot account for the degree of poverty among individuals. To overcome this, a number of 'poverty gap indexes' has been developed, that are some function of the summed differences between the poverty line and the incomes/expenditures of each household. Examples are the poverty gap index (Foster *et al*. 1984); the squared poverty gap index (Foster *et al*. 1984); the Sen index (Sen 1976); and the re-normalised Sen index (Shorrocks 1995).

The FGT poverty indicators can be summarised as:

$$\frac{1}{N}\sum_{i=1}^{Q}(z_i - y_i)^{\alpha} \qquad \text{.... 1}$$

where N = the total population, z_i is the poverty line for individual i, y_i is the welfare indicator for the same individual and Q is the total population below the poverty line. For the head count index $\alpha = 0$; for the poverty gap index $\alpha = 1$ and for the squared poverty gap index $\alpha = 2$.

2.4 Gini coefficients

Another widely used estimate of inequality is the Gini index (see for example World Bank (2006) for a detailed description). Essentially, the Gini index measures the extent to which the distribution of a welfare index (be it expenditure, income, consumption or whatever) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentage of the index against the cumulative proportion of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as the share of

the maximum area under the line. Thus a Gini index of zero represents perfect equality, while an index of 1 implies perfect inequality.

2.5 Recent results for Uganda

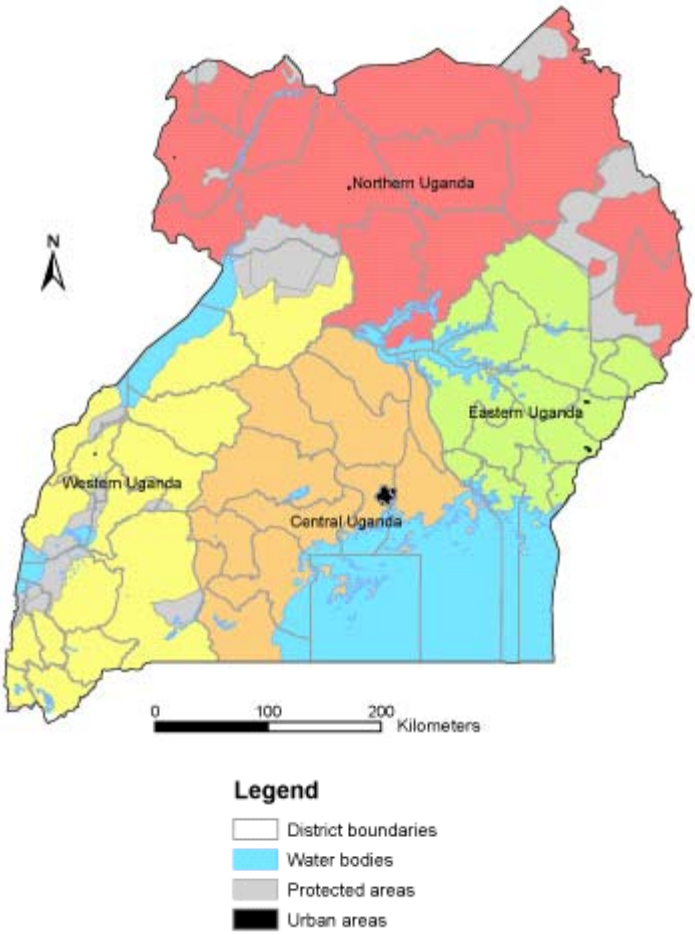
The Uganda Bureau of Statistics (UBOS) has carried out 8 rounds (see Table 1) of nationally representative surveys since 1988 in its endeavour to collect and update data on a wide range of economic, social and demographic indicators. These household surveys have had varying objectives and scope, but common to them all is a socio-economic module, which has provided useful information for monitoring welfare in Uganda.

Table 1: *Details of Uganda national household surveys, 1988-2003*

Survey Round	Dates	Households covered
Household budget survey (HBS)	Apr. 1989 – Mar. 1990	4,595
Integrated household survey (IHS)	Mar 1992 – Mar. 1993	9,925
Monitoring survey 1 (MS-1)	Aug. 1993 – Feb 1994	4,925
Monitoring survey 2 (MS-2)	Jul. 1994 – Jan 1995	4,925
Monitoring survey 3 (MS-3)	Sep. 1995 – Jun. 1996	5,515
Monitoring survey 4 (MS-4)	Mar. 1997 – Nov. 1997	6,654
Uganda National Household survey 1 (UNHS-1)	Aug. 1999 – Jul. 2000	10,696
Uganda National Household survey 2 (UNHS-2)	May 2002 – Apr. 2003	9,711

In Uganda, household surveys have been designed to be representative at the regional level, within which urban and rural households are distinguished. There are four regions: Central, Eastern, Northern and Western, shown in Figure 1.

Figure 1: *Map of the four primary regions of Uganda.*



Poverty estimates are most widely available for the surveys of 1992; 1999 and 2002, and Table 2 shows the values of the measures described above for these surveys in Uganda.

Table 2: *Welfare estimates for Uganda by region based on household surveys of 1992, 1999 and 2002. FGT0 = the head count index ($\alpha = 0$); FGT1 = the poverty gap index ($\alpha = 1$); FGT2 = the squared poverty gap index ($\alpha = 2$); and Gini = the Gini estimate of inequality. C = Central region; E = Eastern region; N = Northern region; W = Western region; U = urban; R = rural.*

		1992				1999				2002			
		FGT0	FGT1	FGT2	Gini	FGT0	FGT1	FGT2	Gini	FGT0	FGT1	FGT2	Gini
C	U	21.5	5.9	2.26	0.39	6.1	1.0	0.28	0.41	7.8	1.6	0.47	0.48
	R	52.8	19	8.95	0.33	25.2	5.8	1.95	0.33	27.6	6.9	2.49	0.37
E	U	40.6	12	5.16	0.32	17.1	4.2	1.4	0.43	17.9	4.8	2.12	0.40
	R	61.1	23.1	11.5	0.32	36.7	9.8	3.82	0.32	48.3	14.9	6.28	0.34
N	U	52.6	20.6	10.76	0.39	28.6	8.2	3.18	0.39	31.4	9.8	4.27	0.41
	R	72.2	28.7	14.66	0.33	65.4	25.4	12.75	0.32	65.0	24.2	11.95	0.32
W	U	29.7	7.3	2.6	0.35	5.7	1.0	0.27	0.39	16.9	4.5	1.73	0.44
	R	53.8	19.2	9.33	0.31	27.4	6.4	2.18	0.29	32.7	8.2	3.0	0.33

Sources: Appleton et al. (1999); Ssewanyana et al. (2004); UBOS (2003).

Broadly speaking, poverty is relatively low in both the Central and Western Regions, in 1992, 1999 and 2002, for both rural and urban areas. The Eastern Region is intermediate and the Northern Region, with over 70 percent of the rural population poor in 1992, remained the poorest region in Uganda in 1999 and 2002. The overall trend from 1992 to 2002 is a considerable reduction in all measures of poverty, in all regions but, if anything, an increase in inequality, as measured by the Gini coefficient.

3. MAPPING POVERTY

These household surveys provide quite detailed information about many aspects of welfare at the regional level, not only estimates based on material deprivation such as the FGT indictors, but also indicators based on health, education *etc*. However, the estimates cover very large and diverse areas and as such are of limited use for targeting, and can provide very little explanation of the patterns observed.

Crump (1997) gives two explanations for poverty: i) the individualistic theory, which assumes that people are mobile and remain in poor areas because of specific wage or price incentives (less competitive environment); and ii) the structural theory, which assumes limited mobility and a causal link between geography and levels of well-being. Spatially explicit factors such as natural resource endowment, infrastructure and access to services result in 'spatial poverty traps', and barriers to migration ensure that these differences persist. When geographically linked factors are major contributors to levels of poverty there is a clear need to map estimates of poverty at appropriate levels of spatial disaggregation. Some reviews of poverty mapping approaches and results are given in Ghosh and Rao (1994), Henninger (1998), Deichmann (1999), and Davis (2003). Expanding on our simple justification for poverty mapping, *i.e*. targeting, explaining and predicting, we can add the following details under each heading:

Targeting

1. To enable geographical targeting for interventions *e.g*. social, agricultural, emergency, environmental and anti-poverty programmes.
2. To facilitate development planning and policy formulation at the national and sub-national levels *e.g*. for planning public investments in education, health, sanitation, water, transport, and other sectors.
3. To incorporate poverty estimates into spatially explicit decision support systems. For example enabling linkage to other monitoring systems such as USAID's Famine Early Warning System (FEWS), or FAO's Food Insecurity and Vulnerability Information and Mapping System (FIVIMS).
4. To facilitate comparisons between regions/countries *etc*. through standardised estimates of poverty.
5. To facilitate information dissemination and advocacy, such as to politicians and donors.

Explanation

1. To allow visual comparison with environmental data to discern correlations (Henninger and Snel 2002).
2. To enable spatial analysis of poverty data, such as exploring clustering and other spatial patterns.
3. To determine environmental correlates of poverty (*e.g*. natural resource endowments, infrastructure) and therefore to identify appropriate development interventions.
4. To create spatial variables pertaining to welfare for use in multivariate analysis of other, related issues.

Prediction

1. To estimate poverty in data-sparse areas.
2. To predict changes in poverty, based on changes in specific variables or arising from poverty-alleviation measures.

The appropriate scale of mapping will clearly depend on the application. Information needs must be balanced against the costs of collecting data. Whereas fairly coarse estimates may be suitable for regional comparisons, these are likely to be inadequate for targeting interventions and exploring the causes of poverty. The resolution at which poverty maps can be made is very dependent on the availability of data and the required accuracy of the poverty maps. In general increasing the level of disaggregation in the data increases the errors in the estimates of poverty.

3.1 Sources of information

Henninger (1998) reviews data collection methods and sources of poverty information across the globe. There is generally a trade-off between sample size/geographic coverage and level of detail. Bottom-up approaches such as intensive anthropological/sociological studies and participatory/rapid appraisal methods typically collect detailed data but from small samples with limited geographical coverage. Such methods are useful for identifying solutions and developing interventions. Top-down approaches such as census and welfare surveys typically collect a limited range of data but from large samples, providing wide geographic coverage, and offering the possibility to map poverty.

Welfare surveys are relatively comprehensive but provide poor coverage. A census by definition gives complete coverage, but only provides limited detail, and rarely includes information on income, expenditure or consumption. The two most widespread welfare surveys are the World Bank's Living Standards Measurement Surveys (LSMS), which measure economic aspects of well-being, and USAID's Demographic and Health Surveys (DHS), which measure food supply and health care. These surveys usually only provide statistically reliable information at the provincial or regional level.

Surveys should ideally be designed to ensure good spatial coverage and statistical significance of survey data at relatively low levels of spatial aggregation. The challenge of distinguishing population differences from sampling variation would not be a problem if a survey was a simple random sample of households from the population, but this is rarely the case: typically surveys are clustered and highly stratified. Howes and Lanjouw (1997) describe some factors that can influence the analysis of welfare surveys, such as clustering, stratification and the number of sampling stages involved. These factors must be accounted for in statistical analysis of survey data so that the results are appropriately interpreted and have accurate error terms associated with them at each level of spatial disaggregation.

3.2 Small area studies – based on survey and census data

Whilst various methods have been used for poverty mapping, some reviewed by Davis (2003), the most common is the small area estimation technique, developed and exemplified in a series of World Bank studies (*e.g.* Hentschel *et al*. 2000; Elbers and Lanjouw 2000; World Bank 2000). This technique involves the application of econometric techniques to combine sample survey data with census data for

prediction of poverty indicators using all households covered by the census. The survey provides the specific poverty indicator and the parameters, based on regression models, to predict the poverty levels for the census. Usually the poverty indicator is a consumption or expenditure-based indicator of welfare, such as the proportion of households that fall below a certain expenditure level (*i.e.* the poverty line). The basic methodology is quite simple and involves the following stages:

Zero stage: the sampling strategy is understood, the comparability of data sources established and common variables between census and survey are identified.

First stage: A regression model is estimated of (log) per capita consumption or expenditure in the household survey based on variables common to the census and the survey. The model thus provides a set of empirical regression parameters. These regressions are generally nested at various spatial levels, from regional down to household levels.

Second stage: The parameter estimates are taken to the census, where they are used to predict consumption, and thus to estimate poverty and inequality for each population of interest. It is particularly important to gauge the precision of the poverty estimates by computing standard errors. Standard errors increase with the level of disaggregation and tend to "explode" at cluster sizes below a certain threshold.

In general:

$$y_i = A_i'\beta_i + \varepsilon_i \qquad \text{.... 2}$$

where y_i is the welfare indicator for household i, A_i' is a vector of independent variables (and associated parameters, β_i) common to the welfare survey and the census and ε_i is a normally distributed error term.

Ghosh and Rao (1994) review small area estimation techniques. Small area poverty estimates have been made for a number of countries, for example Ecuador (Hentschel *et al*. 2000), South Africa (Alderman *et al*. 2000; Statistics South Africa 2000), Nicaragua (Arcia *et al*. 1996); Vietnam (Minot *et al*. 2003); Epprecht and Heinimann 2004); Kenya (Ndeng'e *et al*. 2003); and Uganda (Emwanu *et al*. 2003).

3.3 Small area studies – based on survey, census and environmental data

The statistical estimation of poverty indicators can be extended to include explanatory factors that are not included in the census or survey, such as eco-climatic conditions, access to resources and markets, sanitation *etc*. In general:

$$y_i = A_i'\beta_i + B_i'\chi + \varepsilon_i \qquad \text{.... 3}$$

where y_i, A_i' , β_i and ε_i are as above and χ is a further vector of independent variables (and associated parameters, B'_i) derived from ancillary environmental variables.

Bigman *et al*. (2000) have estimated poverty indicators at the village level in Burkina Faso in this way, combining welfare survey data, census data and environmental data from a variety of sources, including local road infrastructure, public facilities, distribution of water points and agroclimatic conditions.

3.4 Recent results for Uganda

Recently the small area poverty mapping technique has been applied to Ugandan survey and census data. Emwanu *et al.* (2003) combined information from the 1992/93 Integrated Household Survey (IHS) and the 1991 Population and Housing Census (PHC) to produce baseline 1992 poverty estimates with a spatial profile ranging from the national level down to the county-level for rural areas and the sub-county level for urban areas. These estimates were then updated, using information from the 1999/2000 UNHS (a relatively small sample of the same households that were interviewed in the 1992 IHS), to show estimated poverty levels for 1999 and the relative changes in poverty levels since 1992. Emwanu *et al.* (2003) produced a range of poverty estimates at county-level based on this technique:, the head count index (FGT0); the poverty gap index (FGT1); the squared poverty gap index (FGT2); and the Gini estimate of inequality. Figure 2 reproduces the county level poverty incidence for 1992.

Figure 2: *Small area estimates of poverty incidence in 1992 at county-level. Adapted from Emwanu et al. (2003).*

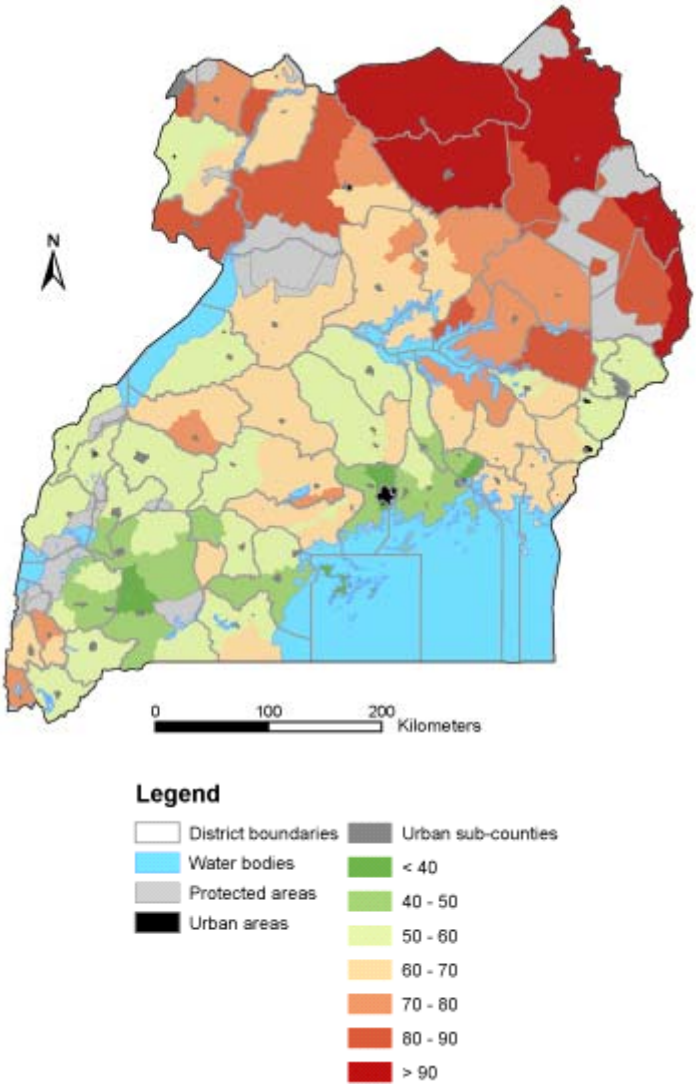
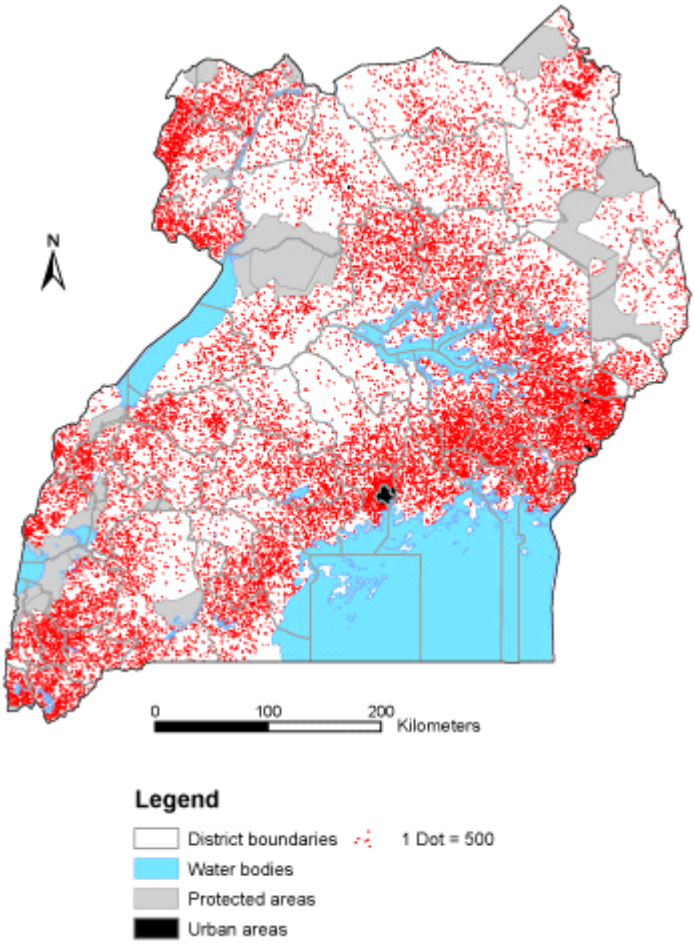


Figure 2 shows high (>50 percent) and widespread poverty across rural Uganda, and considerable geographic variation in poverty incidence at county level. Poverty was greatest in the least secure areas of the north-eastern and north-western parts of Eastern Province and several districts in Central and Western Provinces. It was lowest in the main cities, and the Eastern Province District of Jinja, the Central Province District of Mukono, and the Western Province Districts of Mbarara and Bushenyi. Poverty is more homogenously distributed in some districts than others; as an example of the latter, some counties in Mbarara District, south-western Uganda, have poverty levels of less than 30 percent whilst neighbouring counties have poverty incidences in excess of 60 percent.

Poverty data are often expressed as the proportion of poor people in an area (sometimes called the 'poverty rate') but they may also be plotted as the 'poverty density'; the number of people falling below the poverty line per unit of area. There are several ways to do this, but a particularly enlightening depiction is to create a dot-density map of poor people where each dot represents a specified number of people falling below the poverty line. An example is shown in Figure 3, produced by multiplying county level disaggregated poverty rates (from Figure 2) by the population totals, in this case we used the sub-county level rural population statistics from the 2002 housing and population census (UBOS 2002)[1].

Figure 3: *Poverty density in 1992 based on small area poverty incidence (Figure 2) and the sub-county level rural population statistics from the 2002 housing and population census (UBOS 2002).*



[1] Whilst it would be ideal to use poverty rate and population data from the same time period, these were not available.

As is often the case with such analyses (see for example similar patterns in Vietnam, Epprecht 2005), poverty density maps are almost mirror images of poverty rate maps; people are poorer where they live at low population densities. This raises the obvious question as to whether interventions should be targeted at the poorest areas, or the areas with the largest numbers of poor people. The highest rates of rural poverty in 1992 were found in the more remote northern areas of Uganda which are relatively sparsely populated, but most poor people are found in Central, Eastern and Western Provinces and closer to major urban centres. In addition poverty density 'hotspots', relatively small areas with very high numbers of poor people, are found within the districts of Mbale, Kisoro, Kasese, Masaka, Kampala and Tororo.

4. MAPPING POVERTY IN UGANDA USING REMOTELY SENSED DATA

Having reviewed the more traditional approaches to poverty mapping, particularly in the context of recent studies in Uganda, we now develop a novel approach to mapping poverty in Uganda, using remotely sensed satellite and other data as predictor variables. In developing this approach we have the following objectives, dealt with in the separate sections that follow:

1. To give a brief outline of the use of remotely sensed data in biological studies and their possible application to poverty mapping.
2. To describe ways of processing multi-temporal data that are most appropriate for describing seasonal cycles – important determinants of vector and disease distributions and descriptors of agricultural production systems; hence of likely importance in describing poverty.
3. To describe briefly the modelling approach in general.
4. To explain the concept of probabilistic mapping implied by the term 'risk maps': what they can, and cannot, do for poverty mapping.
5. To describe briefly the origin of the database used in this study, and how the training sets for the poverty maps were derived from it.
6. To describe in some detail the development of the modelling approach used for the Ugandan data.
7. To present and discuss the resulting poverty maps for Uganda.
8. To make comparisons between the poverty maps at a variety of spatial (*i.e.* aggregation) scales.

4.1 The potential importance of satellite environmental variables

Satellite sensor designs are rarely ideal for many biological studies, including poverty mapping, because of trade-offs between spectral, spatial and temporal resolution, determined by constraints of the Earth's atmosphere, or because the commissioning agencies did not have such studies in mind when the satellites were being designed. Passive satellite sensor data (*i.e.* reflections or emissions arising ultimately from the sun) have been used most commonly for the epidemiological studies during which the modelling approach applied here was developed, but there is increasing interest in radar satellites with active sensors that can produce images even under cloudy conditions. Hay *et al.* (2000) provide a very comprehensive review of the application of passive satellite sensor data, and Annex A of this paper gives a brief outline of the spectral, spatial and temporal resolution of a variety of readily available imagery, and stresses the utility of multi-temporal imagery that are a unique guide to habitat seasonality globally.

4.2 Satellite data processing

The TALA Research Group in Oxford, UK, has devised a unique way of image processing multitemporal satellite data that captures the seasonality of natural habitats and is thus ideal for describing seasonal processes. This technique is based on methods of data analysis that split the satellite signal for any channel into annual, bi-annual or tri-annual sinusoidal components, each with a characteristic amplitude

and phase. Such temporal Fourier decomposition of satellite data (named after the French mathematician Joseph Fourier) thus provides a link between the satellite signal and the biological processes that may in some way or other be linked to poverty. The former may therefore be used to describe the latter. As explained in Annex B, temporal Fourier analysis has all the statistical advantages of any good ordination technique applied to satellite data and the additional biological advantage of easy interpretation: in many senses it is the best of both (statistical and biological) worlds. We regard the Fourier variables as giving us a unique 'finger-print' of habitat type. Different habitats will differ in their Fourier components in temperature, humidity and vegetation 'space', and the trick is to discover how best to match these fingerprints to the 'scene of the crime', *i.e.* poverty or disease hot-spots.

Further details of temporal Fourier image processing are given in Annex B, together with examples of Fourier time series from West Africa, and of Fourier-processed images of the United States of America.

4.3 The modelling approach

Key to our approach to poverty mapping are the mathematical techniques used to bring the socio-economic data (from the database) together with the satellite and other data to make the poverty map. There are several 'standard' algorithms in the literature, the commonest for mapping purposes being logistic regression analysis, available in a wide variety of commercial packages. Again the TALA research group has developed its own algorithms based on the much more powerful non-linear (maximum-likelihood) discriminant analytical approach, which allows the prediction not only of binary (presence/absence) data, a restriction of most logistic regression techniques, but also continuous (*e.g.* socio-economic) and multiple category data.

The precise way in which the poverty models were built benefits from our previous experience with vector-borne diseases. Briefly, discriminant analytical maximum-likelihood methods are employed to link the poverty and satellite data in a statistically robust way. The poverty data, divided into a series of 'bins' or categories of household expenditure levels, are the dependent variables and the temporal Fourier data layers are the independent or predictor variables. For each set of dependent variable data (*e.g.* the set of poverty categories) the algorithms examine the predictor variables one at a time to discover which one maximises the discriminant criterion selected by the user (see Annex C for details). This variable is then the first selected variable of the eventual poverty map. The algorithm then goes through all the remaining variables, again one at a time, to select which one, in association with the variable already selected, maximises the same discrimination criterion. This becomes the second selected variable: and so on. The algorithm continues until some predefined stopping criterion is met, or until 10 variables have been selected. The set of selected variables is then used to make an image, or map, of predicted poverty categories, using the weighting factors determined by the discriminant analytical method. Thus these maps essentially 'fill in the gaps' between data points to make predictions of poverty at the spatial resolution of the satellite imagery and other spatial data. Further details of the modelling approach are given in Annex C.

The great attraction of discriminant analysis for the present application is that the technique makes no assumptions about the overall relationship (linear, or non-linear in a particular way) between the predictor and predicted variables. Thus a great variety of responses can be described by essentially the same technique. This therefore overcomes the restriction of the usual multiple regression approach to poverty mapping where various obvious transforms (log, square root *etc.*) may be applied, more in hope than expectation that the transform normalises the data and

ensures a linear (or simple non-linear) relationship between predictor and predicted variables.

4.4 Poverty risk maps, what they can and cannot do

A poverty risk map is a 'best-guess' representation of how an area (a region, state or continent) would look if sampled at all points for the socio-economic indicators under study. We generate this prediction by establishing the relationship between the satellite and socio-economic data for areas or points where we have records of whatever index of poverty we are trying to map. Such records provide what is called the 'training set' for the analysis. As outlined above, it is obviously important that the training set should sample as wide a range of socio-economic conditions as possible, in as wide a range of environments as possible. Only in this way will the precise statistical relationships between the indicators and satellite data be established that can then be used at full satellite resolution, to fill in the gaps between the training set sample points to make the poverty risk map. In general we have to compromise with less than complete training set data, and our output prediction maps include areas of 'no prediction' where environmental conditions are so different from any of those in any training set site that we choose not to make any predictions for them at all. What is a 'No prediction' area in one version of a poverty map may become a poverty high-risk area in later iterations of the map, as new data become available. Thus 'No prediction' areas on poverty risk maps should be treated with caution.

4.4.1 How do we know a poverty risk map is 'correct'?

Once we have produced a poverty risk map, the next step is to test its accuracy. This can sometimes be done at the model-building stage, by various boot-strap (*i.e.* sub-sampling) or jack-knife methods (*i.e.* build a model using all but one of the data points, and then examine its predictions for the omitted point; repeat this for all points in the data set and calculate the accuracy of predicting 'unknown' points), but may also be done by sampling areas on the ground predicted, by the risk map, to have a certain level of poverty.

The problem with the 'miss-one-out' technique is that, in each case, the model is often built using many hundreds or even thousands of data points. The omission of any single point is unlikely to make much of a difference in models based on the assumption of multi-variate normality. In our previous experience, mostly with vector-borne diseases, we have found that missing out even half of a large training set of data, and testing subsequent models on the omitted data, gives levels of overall accuracy that are only a few percent points less than those obtained using the entire training set. Nevertheless we are acutely aware of the need to test model accuracy; only by examining reasons why our initial models fail to predict well can we hope to improve the modelling process.

In the present case we have only relatively sparse data sets for household poverty throughout Uganda. Although the 'miss-one-out' technique might tell us about the accuracy of our current poverty risk maps (using the current data set) we cannot be certain that the data set is an unbiased sample of the poverty situation on the ground (although the sampling was designed to try to ensure this, at least at the level of the region). It is rather pointless to improve the accuracy of a model based on inaccurate or incomplete data. Instead we need to develop ways of 'best-guessing' the poverty situation from the information we have. This requires a combination of statistical manipulations coupled with insight, knowledge, or experience-based guess-work.

Clearly we will never have enough test data to prove whether or not any predictive poverty risk map is 100% accurate. Even well-resourced prediction systems (*e.g.* weather forecasting) are never tested in this way. Instead, over the course of time, sufficient observations are accumulated to give us confidence in the capabilities of our poverty mapping procedures. To begin with, all we ask is that our poverty risk maps are more right than wrong or, at the very least, provide improvement on other approaches). The role of further research is to improve the prediction exercise, to increase the odds in favour of a correct prediction.

4.5 Data sources

4.5.1 Poverty database

The modelling exercise was based upon the Uganda National Household Survey 2 (Table 1). First the database was examined for errors of location. Some of these were obvious (*e.g.* households cannot exist in Lake Victoria) others less so (*e.g.* when a unique administrative unit name was associated with latitude/longitude co-ordinates that fall outside its boundaries). Households with locational errors that were not obviously rectifiable were rejected from the analysis. Secondly a poverty index was selected for use in modelling. Of those available, we decided to model household expenditure (therefore not taking into account the regional poverty lines that were developed in the survey). Thirdly a decision was taken as to the number of categories into which to divide the continuous poverty data; a total of ten categories seemed to give a sufficiently variable measure of poverty, without expecting too much of the model in terms of discriminatory power. Preliminary analyses suggested that a finer-grained division of the poverty data was unwarranted. At each spatial resolution (see below) the category boundaries for the poverty data were chosen so as to have similar numbers of observations in each category.

4.5.2 Satellite database

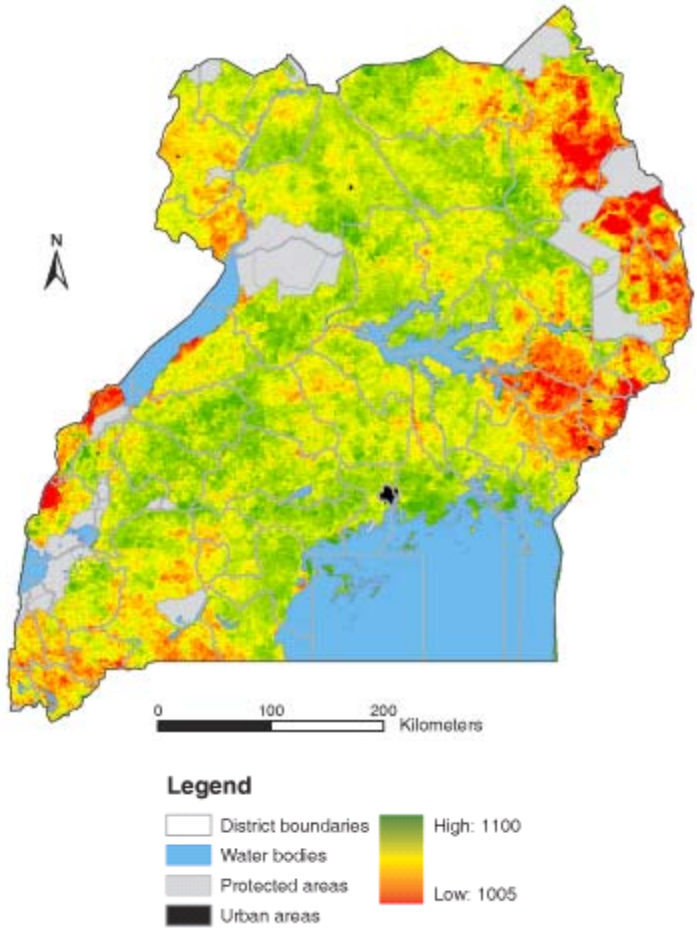
The majority of the data used in the explanatory model were satellite-derived, and most came from the 1km global AVHRR dataset made available by the NASA Pathfinder program. These data were processed by the Pathfinder program only for a limited number of months between 1992 and 1996. These data were aggregated into synoptic monthly (maximum value) composites to give a record of monthly changes in an average year. One synoptic series was produced for each of the following: the middle infra-red (MIR, AVHRR channel 3), Land Surface Temperature (LST, produced by combining information from AVHRR channels 4 and 5), the Normalised Difference Vegetation Index (NDVI, produced by combining information from AVHRR channels 1 and 2), air temperature (Tair, produced by combining LST with NDVI), and Vapour Pressure Deficit (VPD, a combination of satellite and ground-based meteorological data). Thus effectively all five bands of the AVHRR sensor were being used in the analysis. In addition to these AVHRR data, information from the European geostationary Meteosat satellite in the form of a rainfall surrogate, the Cold Cloud Duration (CCD), was obtained from the FAO ARTEMIS program[2].

The original AVHRR imagery is in the Goode's Interrupted Homolosine projection, and the CCD imagery in the Hammer-Aitoff projection (a variant of the Lambert projection). Each data series was temporally Fourier-processed to produce 10 separate data layers; the mean (1 layer), the phases and amplitudes of the annual, bi-

[2] METEOSAT data were provided by Fred Snijders, FAO.

annual and tri-annual cycles of change (6 layers in all), the maximum, minimum (2 layers) and the variance (*i.e.* original channel variance, not that of the Fourier series) (1 layer). After temporal Fourier processing, the data were re-projected to the longitude/latitude system by bi-linear interpolation to a nominal pixel resolution of 0.01 degrees (about 1.2 km at the equator). For those data layers at an original spatial resolution coarser than 1km (hence also of 0.01 degree), the data were interpolated to the same spatial resolution: this applies to the VPD and CCD imagery. As an example, Figure 4 shows the maximum value of the long-term monthly average of the NDVI. This type of image is sensitive to discriminating vegetation growth in dry areas, such as the Karamoja region of north-east Uganda in Figure 4.

Figure 4: *Maximum value of the long term monthly average Normalised Difference Vegetation Index (NDVI).*



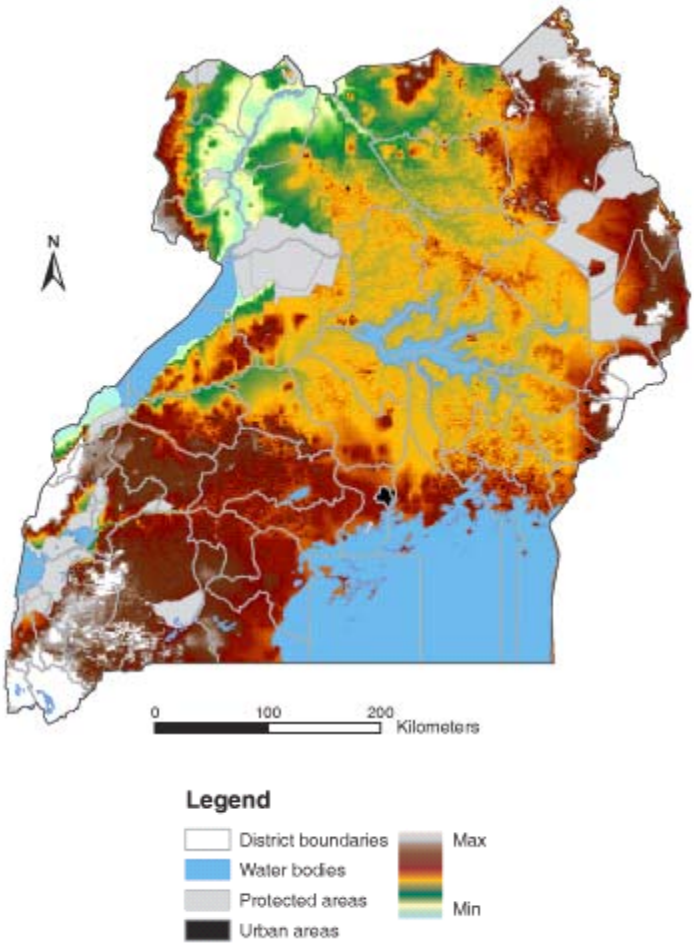
Finally, each of the 0.01 degree Fourier layers was in turn aggregated (by averaging) into a series of images with nominal resolutions of 0.02, 0.03, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75 and 1 degree. Models were developed at all of these spatial scales to investigate the relationship between predictive accuracy and spatial resolution.

4.5.3 Ancillary ground and other data

In complementary studies to this analysis of poverty in Uganda, a number of environmental variables have been mapped in digital format, many of which are likely to be closely associated with, and possibly contribute to, the causes of poverty. For the present analysis, in addition to the series of remotely sensed environmental

variables described above, the following layers were made available as potential predictor variables for the discriminant analytical modelling: digital elevation (Figure 5), human population density (Figure 6), access to markets (Figure 7), cattle, sheep, goat and pig densities (Figure 8) and probability of presence of major tsetse species (Figure 9)[3]. All of these additional data layers were clipped and/or resampled to the same spatial resolutions as, and made coincident with, the satellite data layers for Uganda. They were also aggregated by averaging, in the same way as the Fourier-processed satellite data, for the models at different spatial resolutions.

Figure 5: *Elevation derived from the Global Land One-kilometre Base Elevation (GLOBE) data set, Version 1 (Hastings and Dunbar 1998; 1999).*



[3] The emphasis here on livestock-oriented variables is due to the context of this analysis, which is to develop poverty mapping approaches in support of pro-poor livestock policy analysis and formulation – for further information see the PPLPI web site: http://www.fao.org/ag/pplpi.html.

Figure 6: *Rural population density estimated at sub-county level from the 2002 Uganda national housing and population and census (UBOS 2002).*

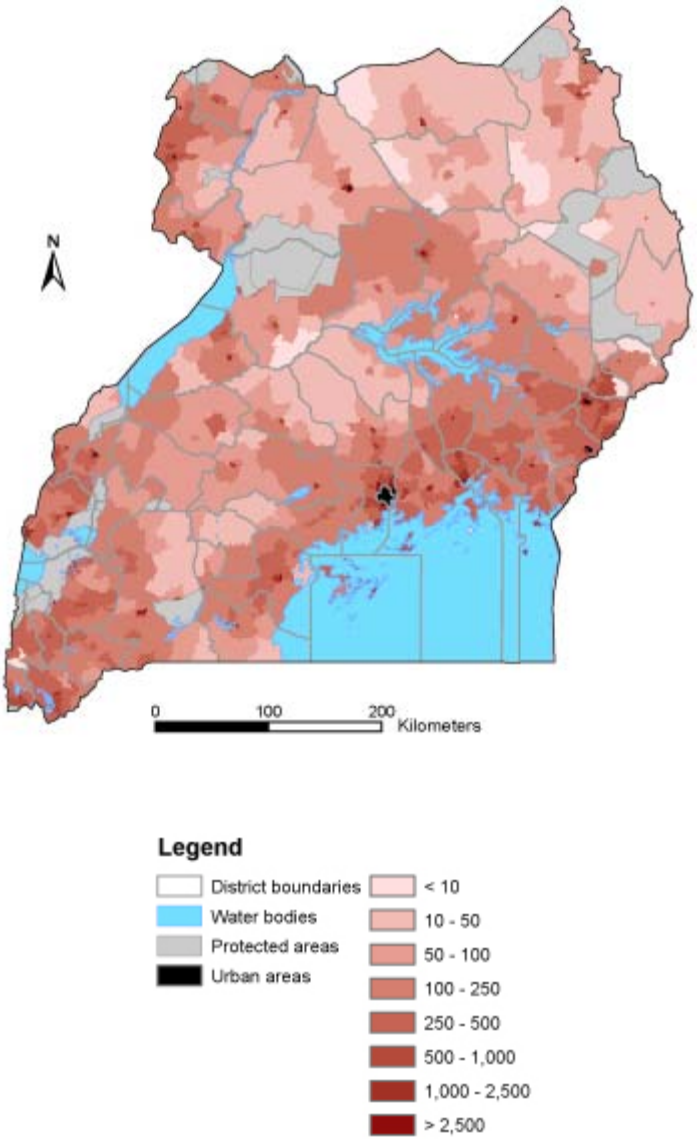


Figure 7: *Time taken to travel to populated places with more than 50,000 people. Produced using data provided by IFPRI – described in You and Chamberlin (1994).*

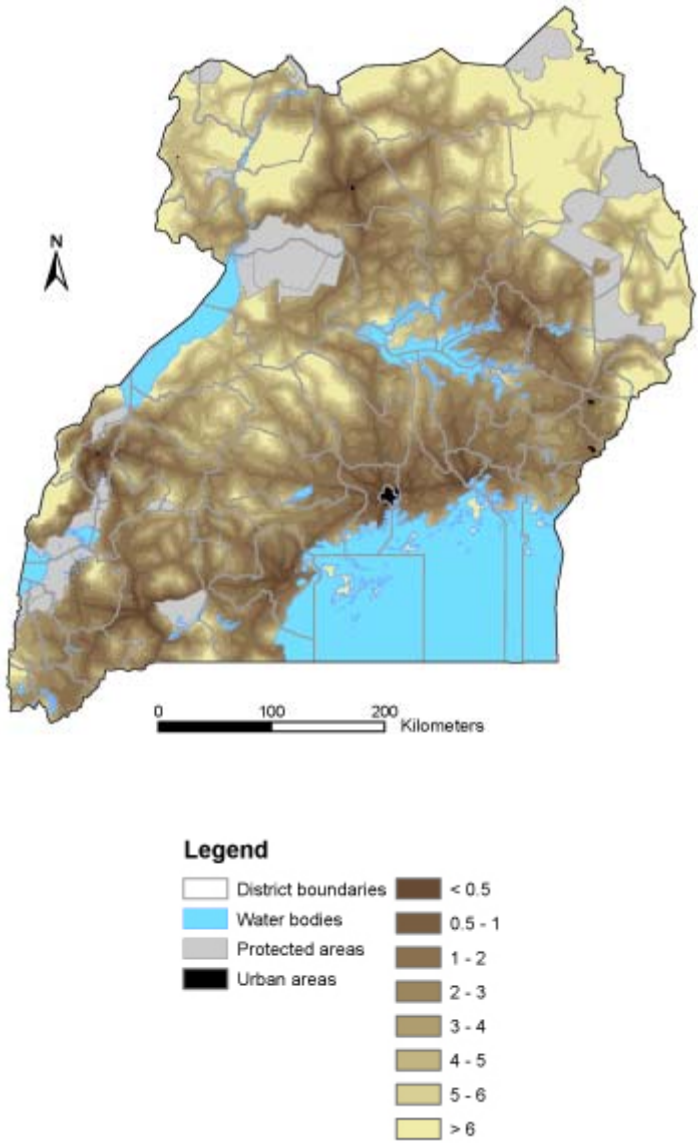


Figure 8: *Densities of major livestock species in Uganda: a) cattle, b) sheep, c) goat and d) pigs, summarised by (rural) sub-county from the 2002 Uganda national housing and population and census (UBOS 2004).*

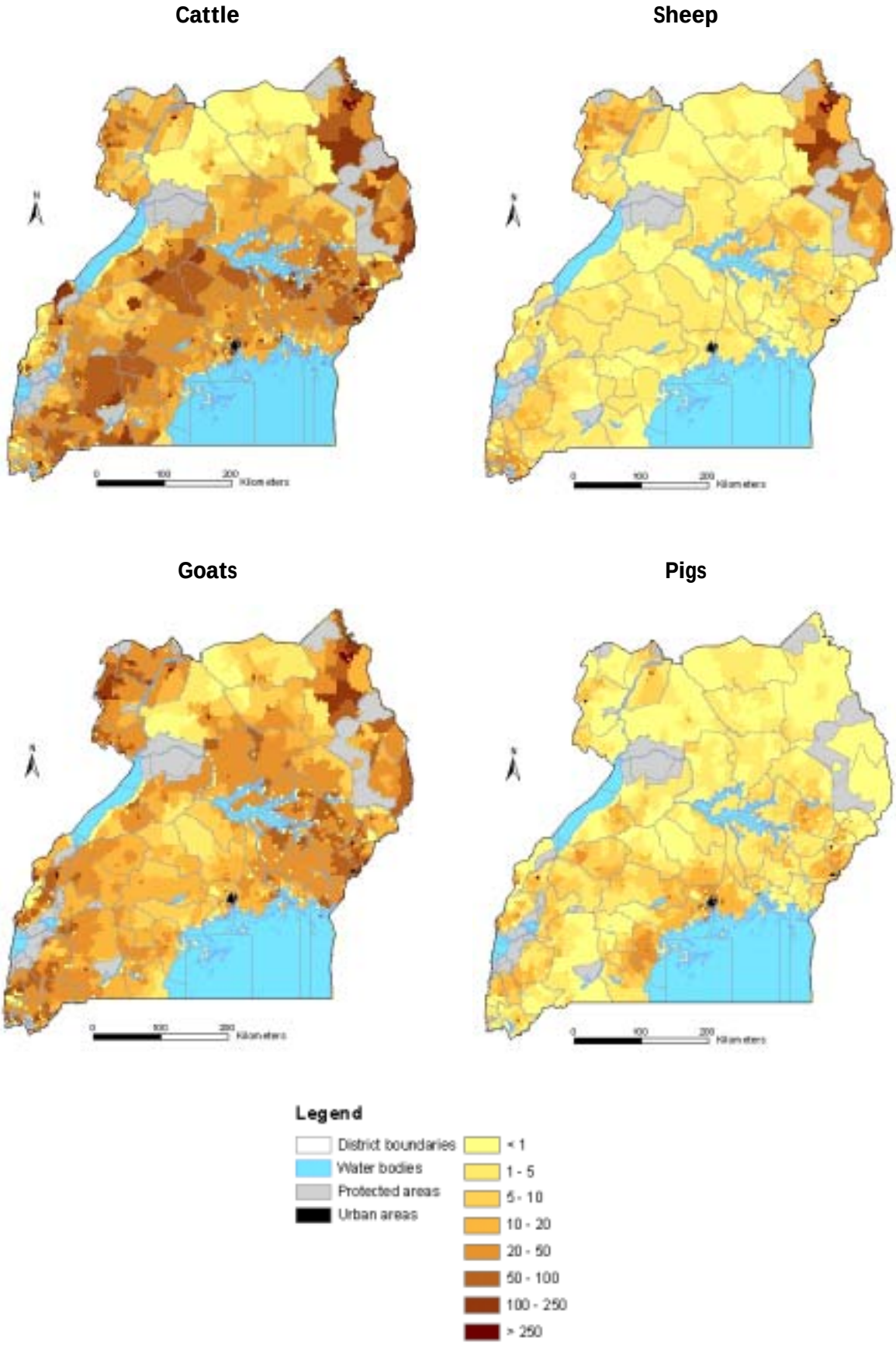
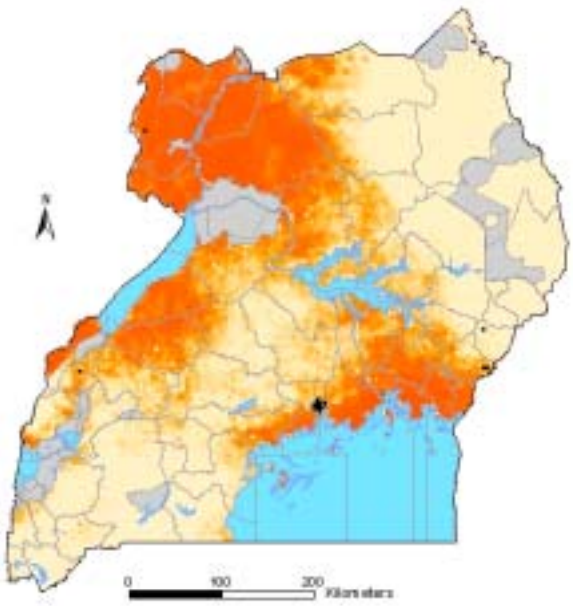


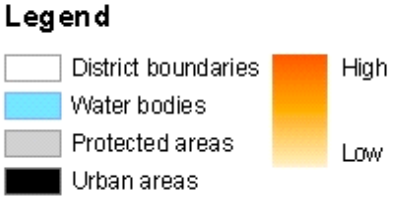
Figure 9: *Modelled distributions (probability of tsetse presence) of the three predominant tsetse species in Uganda: a) Glossina fuscipes fuscipes, b) G. pallidipes, and c) G. morsitans submorsitans (Wint 2001).*

a) Glossina fuscipes fuscipes



b) G. pallidipes

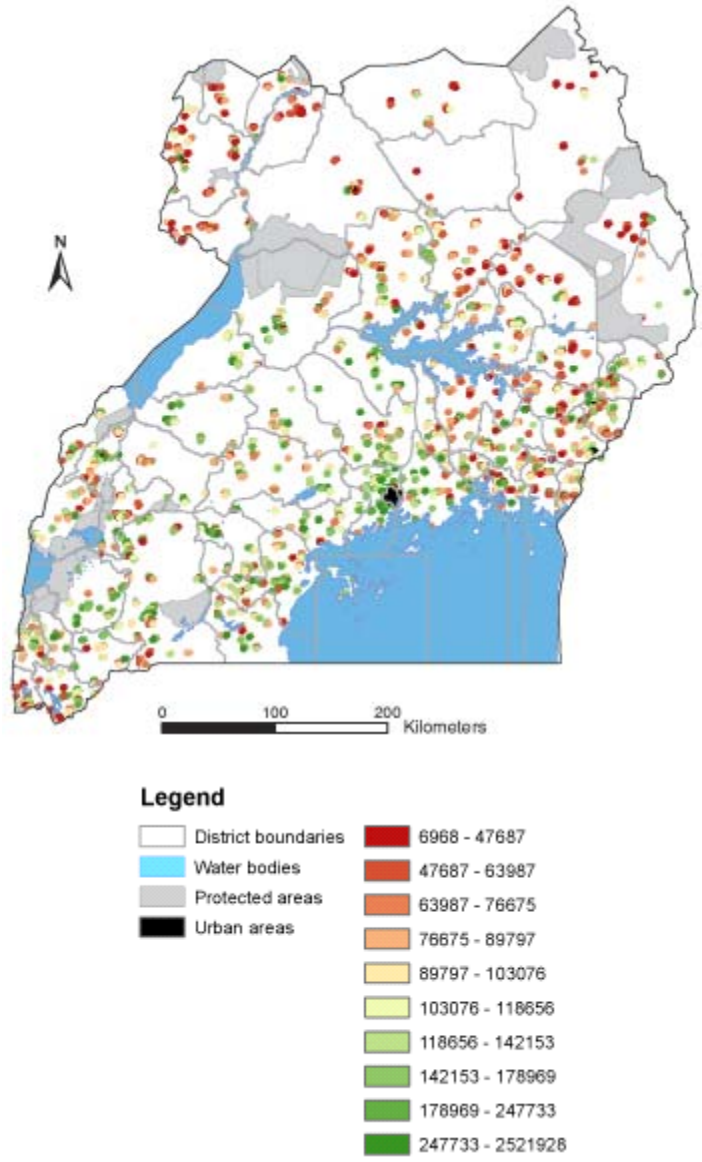
c) G. morsitans submorsitans



4.6 The modelling approach applied to the Ugandan poverty data

Two series of models were developed; each spanning the entire range of spatial resolutions of the predictor data layers, from 0.01 to 1 degree spatial resolution. Aggregating the household data to the same spatial units as the satellite data was felt to be the best way to investigate the relationships between local environmental conditions and poverty. The household data were aggregated in two stages. In the first stage, data for the households that fell within the same 0.01 degree pixel were first averaged. The resulting data are shown in Figure 10. An image was also constructed that stored the number of households contributing to each pixel's average value.

Figure 10: *Household expenditure data were averaged for all the households that fall within each 0.01 degree grid square in Uganda. These data were then assigned to one of 10 'bins' shown here (divisions were selected to give approximately equal sample sizes) that formed the basis for the satellite data analysis.*



During the second stage the household survey data image was progressively aggregated to larger pixel sizes; at each stage the weighted average household expenditure was calculated and inserted into each new image (weights were obviously the numbers of households contributing to each pixel's average value). Thus at all levels of aggregation the household expenditure image contained values that would have been obtained by simple arithmetic averaging of the data for all households falling within each pixel. At the finer spatial scales not all pixels contained a value for household expenditure, and these 'empty' pixels were not used in model construction; as spatial aggregation increased, larger and larger proportions of the (fewer and fewer) pixels contained household expenditure averages, and so could be used in the modelling. These household expenditure data layers were then used to extract the coincident satellite and other data associated with each pixel's expenditure value (at the same spatial resolution).

The analysis followed the discriminant analytical approach outlined previously. The two series of models differed only in the method used to select the predictor variables at each step of the step-wise inclusion employed. In the first series, variables were selected that maximised the kappa statistic (see Annex D). Kappa examines the categorical assignments in a confusion matrix relating observed to predicted results. In the second series of models the information-theoretic approach outlined in Annex C.2 was employed. In this case the probability of a pixel belonging to its 'correct' (*i.e.* observed) category was calculated and used to generate the corrected Akaike Information Criterion (AIC_c) at each step of variable selection. As explained in Annex C these probabilities were, strictly speaking, the Bayesian posterior probabilities calculated from the discriminant formulae. Variable selection was based on minimising the AIC_c at each step. Burnham and Anderson (2000) state that reductions in AIC_c values of greater than 10 certainly indicate better models, whilst reductions of as little as 2 indicate acceptable improvements of model fit (*i.e.* parameters should be retained in models if they achieve these levels of reductions in AIC_c values). Reductions of less than 2 are probably of no interest. As explained in Annex C, there are no formal tests of hypotheses in the information-theoretic approach involving the AIC_c.

4.7 Poverty maps for Uganda

Figures 11 to 16 show the resulting poverty risk maps at a variety of spatial resolutions spanning the entire range of model fits (0.01, 0.05, 0.10, 0.20, 0.30 and 0.35 degrees resolution). In general, and as expected, overall model accuracy increased with decreasing spatial resolution.

Figure 11: *Modelled household expenditure at full spatial resolution of 0.01 degrees (ug051150.img). Kappa = 0.146; r^2 of observed vs predicted = 0.160. The data being modelled are shown as dots (from Figure 10), with the category boundaries as indicated in the legend.*

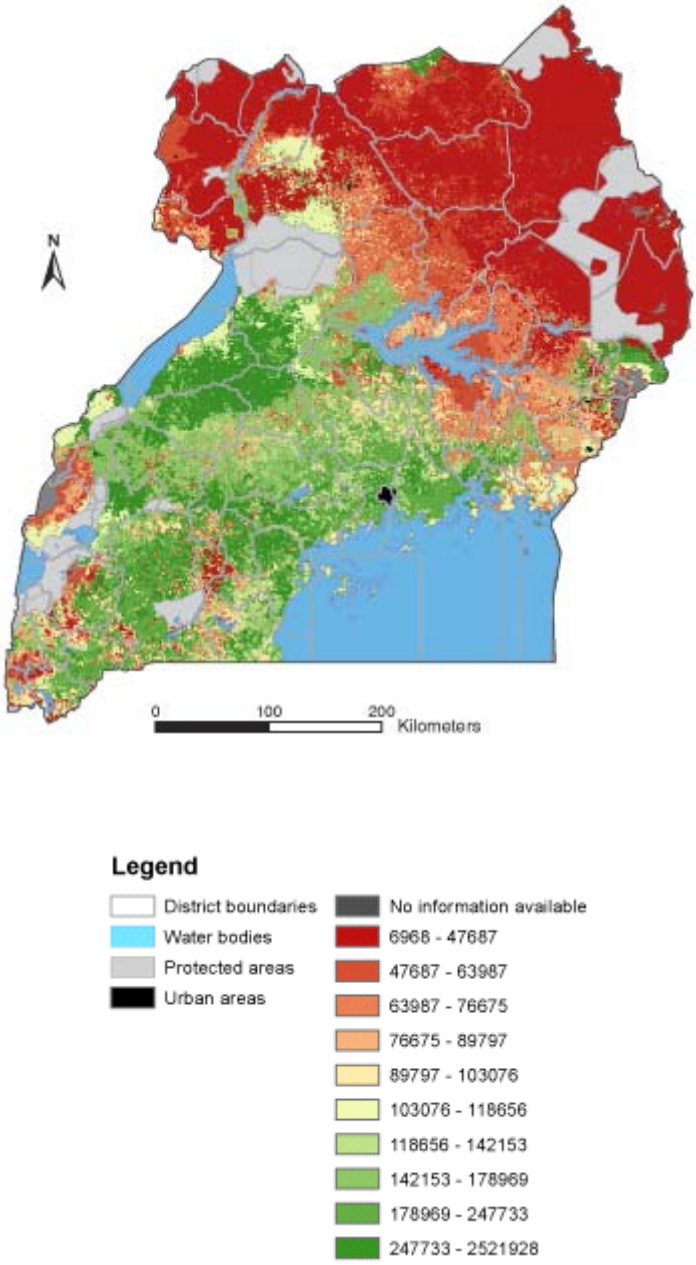


Figure 12: *Modelled household expenditure at a spatial resolution of 0.05 degrees (ug081150.img). Kappa = 0.249; r^2 of observed vs predicted = 0.219.*

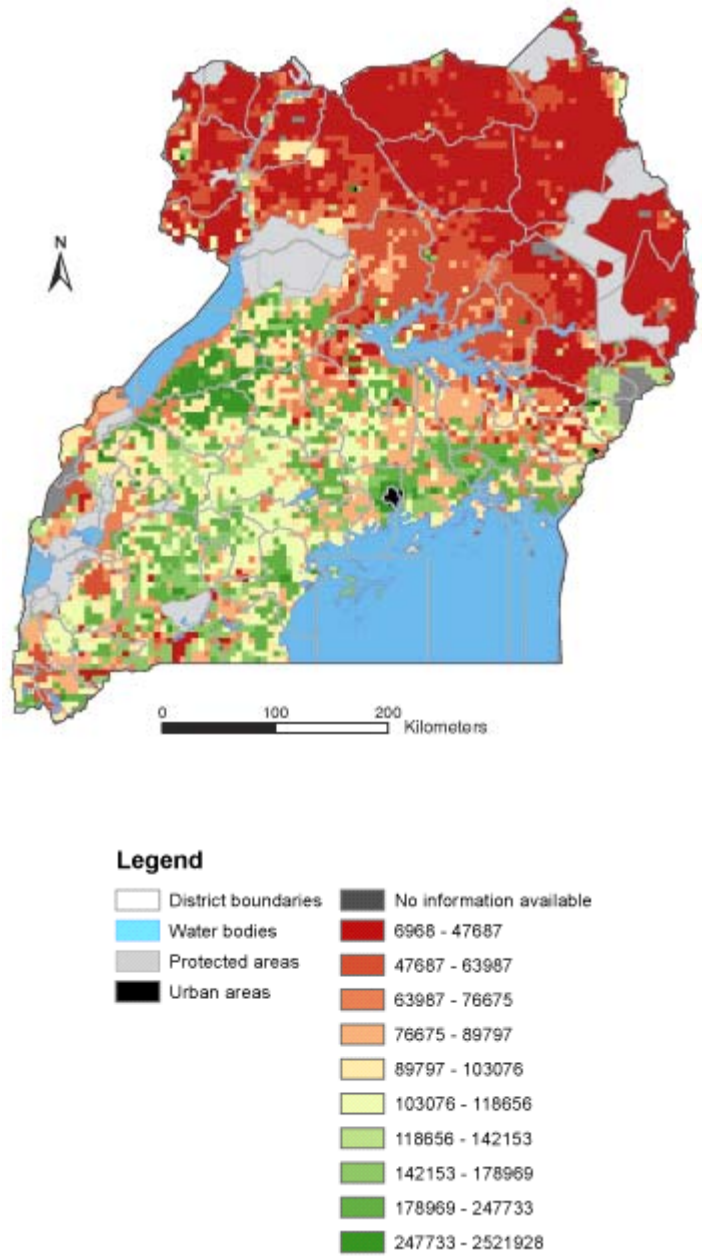


Figure 13: *Modelled household expenditure at a spatial resolution of 0.10 degrees (ug091150.img). Kappa = 0.299; r^2 of observed vs predicted = 0.269.*

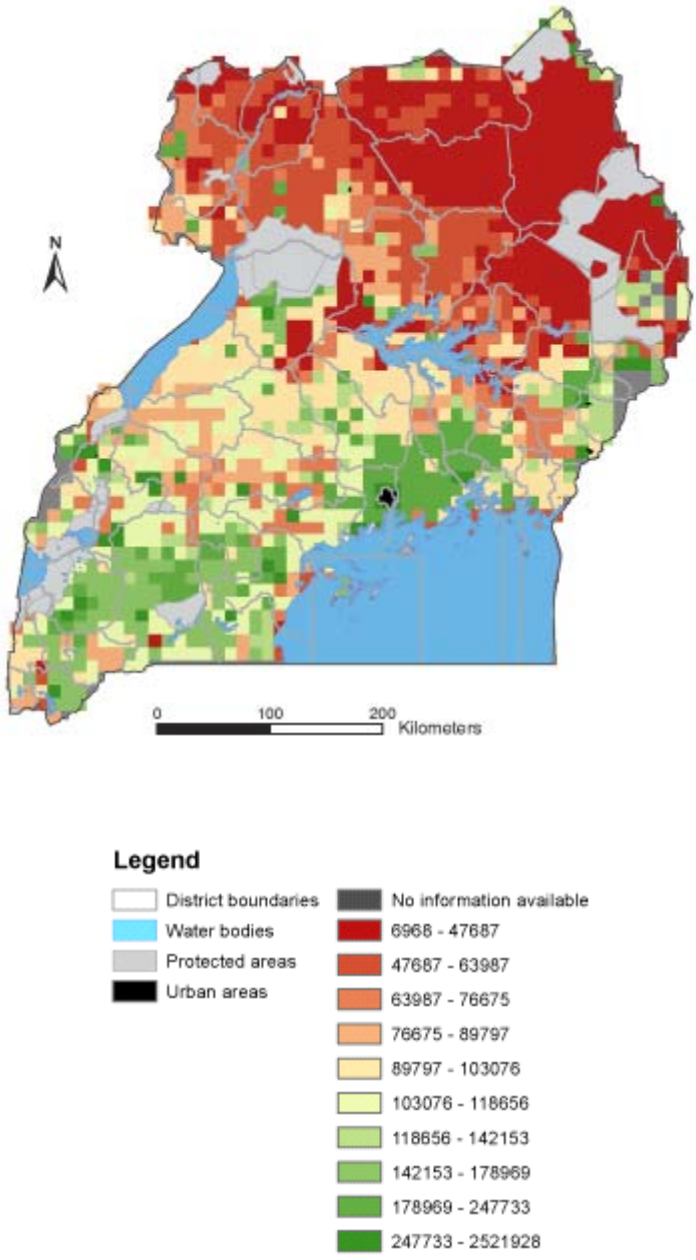


Figure 14: *Modelled household expenditure at a spatial resolution of 0.20 degrees (ug101150.img). Kappa = 0.529; r^2 of observed vs predicted = 0.372.*

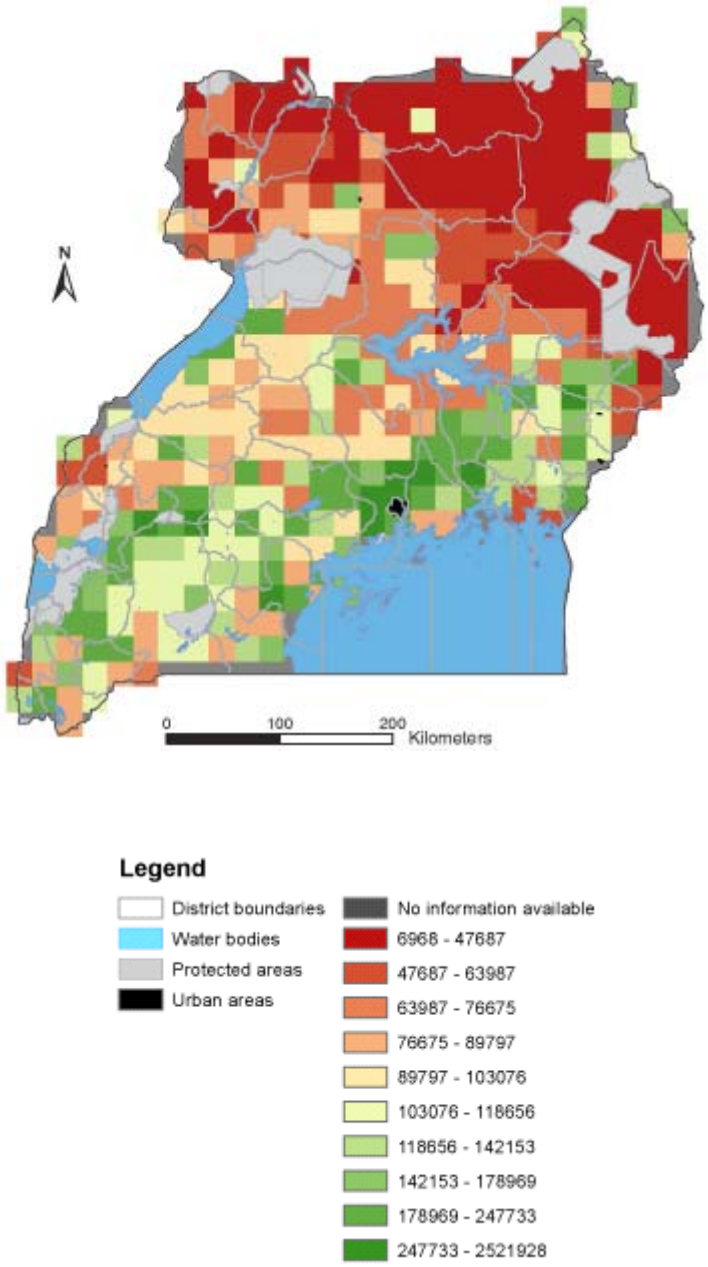


Figure 15: *Modelled household expenditure at a spatial resolution of 0.30 degrees (ug151150.img). Kappa = 0.699; r^2 of observed vs predicted = 0.617.*

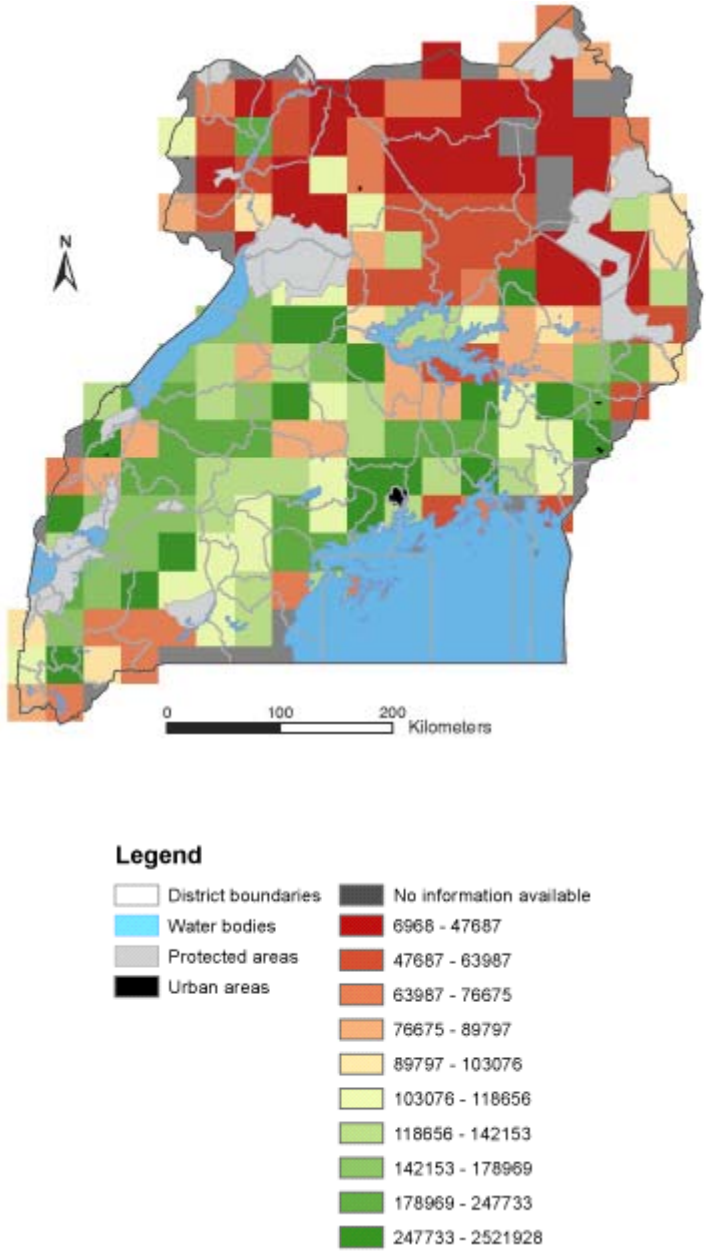


Figure 16: *Modelled household expenditure at a spatial resolution of 0.35 degrees (ug161150.img). Kappa = 0.943; r^2 of observed vs predicted = 0.973.*

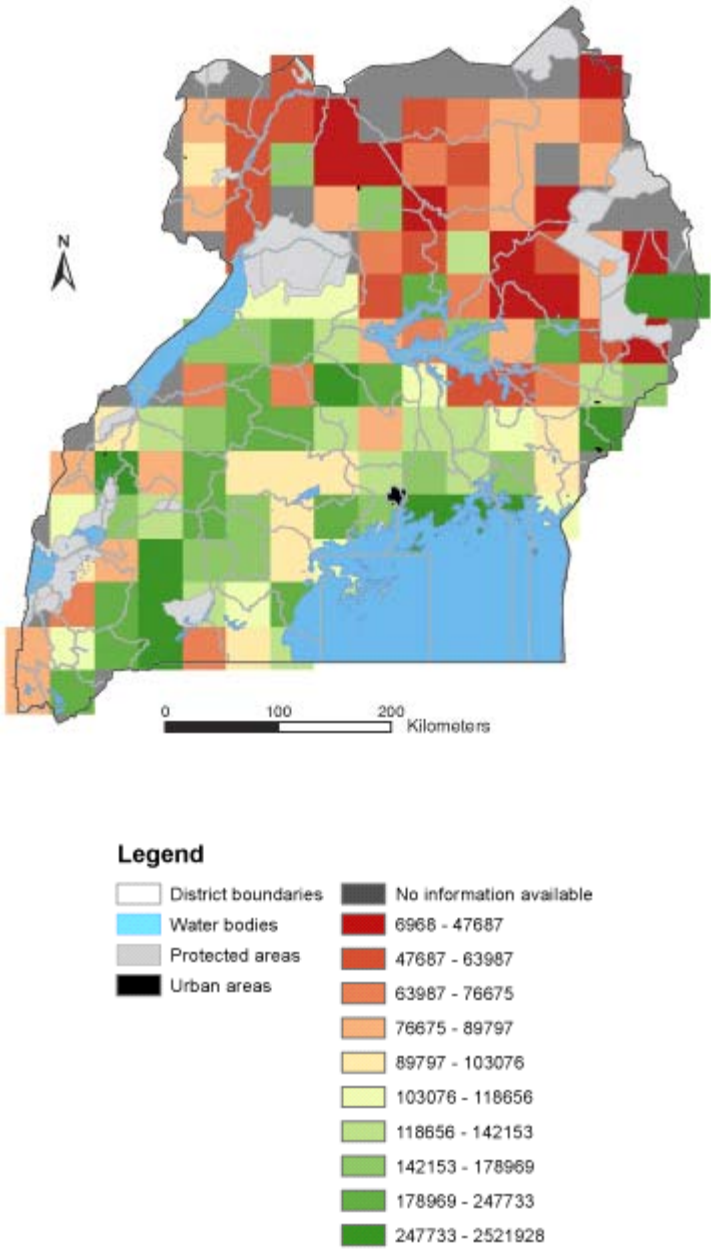


Table 3: *a) Mean values of the top ten selected variables for the 0.01 resolution model (ug051150.img, Figure 11, Kappa = 0.146; r^2 of observed vs predicted = 0.160). b) Accuracy of model description of poverty levels in Uganda: % correct, % correct plus or minus one category, or plus or minus two categories, the producer's and consumer's accuracies (see Annex D for description).*

Table 3a

Category	ug0121a0ll	ug0103a0ll	ug0125vrll	Got02DS	ug0103p1ll	ug0103mnll	Gpall	ug0120a0ll	ug0125a1ll	ug0107mxll	n (Sample)
1	38.74	36.36	17.31	0.36	3.56	30.95	0.09	29.08	33.83	44.18	274
2	37.96	35.84	17.84	0.37	3.53	30.6	0.14	27.97	32.63	43.78	273
3	37.69	35.85	18.55	0.33	3.76	30.69	0.13	27.93	35.31	43.81	275
4	37.13	35.53	18.57	0.34	3.81	30.49	0.14	25.9	34.16	43.29	275
5	36.86	35.44	18.48	0.3	4.28	30.3	0.13	25.99	35.41	43.41	273
6	36.61	35.29	18.82	0.3	4.25	30.4	0.19	25.15	35.87	43.38	273
7	36.54	35.26	18.82	0.3	4.41	30.47	0.14	24.78	37.12	43.29	275
8	36.38	35.25	18.65	0.26	4.21	30.27	0.13	24.55	34.92	43.27	274
9	36	35.35	18.81	0.25	4.52	30.53	0.13	23.8	39.04	43.31	273
10	36.14	35.43	18.99	0.22	4.73	30.47	0.16	23.57	38.8	43.57	274
All	37.01	35.56	18.48	0.3	4.11	30.52	0.14	25.87	35.71	43.53	2739

Key to variable names: ug0121a0ll - Tair mean; ug0103a0ll - Channel3 mean; ug0125vrll - CCD variance; Got02DS - Goat density 2002; ug0103p1ll - Channel3 phase1; ug0103mnll - Channel3 minimum; Gpall - *G. pallidipes* risk; ug0120a0ll - VPD mean; ug0125a1ll - CCD amp1; ug0107mxll - Price LST maximum

Table 3b

Category	Household expenditure	% correct	% correct (+/-1cat.)	% correct (+/-2cat.)	% Producer's Accuracy	% Consumer's Accuracy
1	0.0 to 47687.3	42.7	51.8	64.2	42.7	27
2	47717.1 to 63987.1	19	56.8	61.9	19	26.4
3	64013.3 to 76675.7	23.3	34.9	61.8	23.3	21.8
4	76678.2 to 89797.4	12	30.2	45.5	12	20.2
5	89875.1 to 103076.5	11	23.4	39.6	11	19.5
6	103094.7 to 118656.0	13.9	26.4	45.8	13.9	23.9
7	118692.4 to 142153.6	16	29.5	55.6	16	22.6
8	142246.8 to 178969.0	21.2	43.4	69	21.2	21.8
9	179027.0 to 247733.8	36.3	65.9	72.2	36.3	21.8
10	248340.2 to 8165266.0	36.5	57.7	65.7	36.5	23.6

Key to variable names: ug0121a0ll - Tair mean; ug0103a0ll - Channel3 mean; ug0125vrll - CCD variance; Got02DS - Goat density 2002; ug0103p1ll - Channel3 phase1; ug0103mnll - Channel3 minimum; Gpall - *G. pallidipes* risk; ug0120a0ll - VPD mean; ug0125a1ll - CCD amp1; ug0107mxll - Price LST maximum

Table 4: *a) Mean values of the top ten selected variables for the 0.35 resolution model (ug161150.img, Figure 16, Kappa = 0.943; r^2 of observed vs predicted = 0.973); b) Accuracy of model description of poverty levels in Uganda: % correct, % correct plus or minus one category, or plus or minus two categories, the producer's and consumer's accuracies (see Annex D for description); c) Variable descriptions.*

Table 4a

Category	ug3514vrll	ug3514p1ll	ug3525mxll	ug3521a1ll	ug3503a3ll	ug3507p1ll	ug3525vrll	ug3525a0ll	Gpall	mkt2fin	n (Sample)
1	0.02	6.92	174.38	7.2	1.55	3.45	14.48	101.77	0.03	4.7	13
2	0.03	6.66	198.14	5.66	1.61	3.99	17.54	121.57	0.07	4.49	14
3	0.02	6.3	191	4.82	1.48	3.62	16.73	115	0.28	4.06	13
4	0.01	6.18	185.86	4.31	1.89	3.93	16.74	118.43	0.14	4.03	14
5	0.01	5.16	193.86	2.82	1.54	4.59	18.33	121.57	0.25	3.02	14
6	0.01	5.53	194.08	3.72	1.56	4.46	17.89	119.31	0.17	3.87	13
7	0.02	5.7	204.43	2.56	1.43	4.03	18.25	124.71	0.18	3.42	14
8	0.01	5.9	201.77	3.29	1.46	4.02	18.42	123.31	0.19	2.91	13
9	0.01	5.47	185.57	2.86	1.51	4.39	17.24	116.71	0.19	2.88	14
10	0.02	5.17	188.71	3.14	1.37	4.57	17.98	113	0.15	4.44	14
All	0.02	5.89	191.82	4.02	1.54	4.11	17.37	117.62	0.17	3.78	136

Keys to variable names: ug3514vrll - NDVI variance; ug3514p1ll - NDVI phase1; ug3525mxll - CCD maximum; ug3521a1ll - Tair amp1; ug3503a3ll - Channel3 amp3; ug3507p1ll - Price LST phase1; ug3525vrll - CCD variance; ug3525a0ll - CCD mean; Gpall - *G. pallidipes* risk; mkt2fin - Distance to kmarket

Table 4b

Category	Household expenditure	% correct	% correct (+/-1cat.)	% correct (+/-2cat.)	% Producer's Accuracy	% Consumer's Accuracy
1	0.0 to 64943.5	100	100	100	100	100
2	65647.3 to 85890.6	100	100	100	100	93.3
3	88079.9 to 101556.4	100	100	100	100	92.9
4	102627.6 to 112228.8	92.9	100	100	92.9	92.9
5	112882.8 to 123129.1	92.9	92.9	92.9	92.9	92.9
6	123161.2 to 131255.5	69.2	69.2	84.6	69.2	100
7	131845.0 to 143454.6	92.9	100	100	92.9	100
8	144350.9 to 171718.8	100	100	100	100	86.7
9	172939.2 to 189457.0	100	100	100	100	93.3
10	189776.0 to 609703.6	100	100	100	100	100

Keys to variable names: ug3514vrll - NDVI variance; ug3514p1ll - NDVI phase1; ug3525mxll - CCD maximum; ug3521a1ll - Tair amp1; ug3503a3ll - Channel3 amp3; ug3507p1ll - Price LST phase1; ug3525vrll - CCD variance; ug3525a0ll - CCD mean; Gpall - *G. pallidipes* risk; mkt2fin - Distance to kmarket

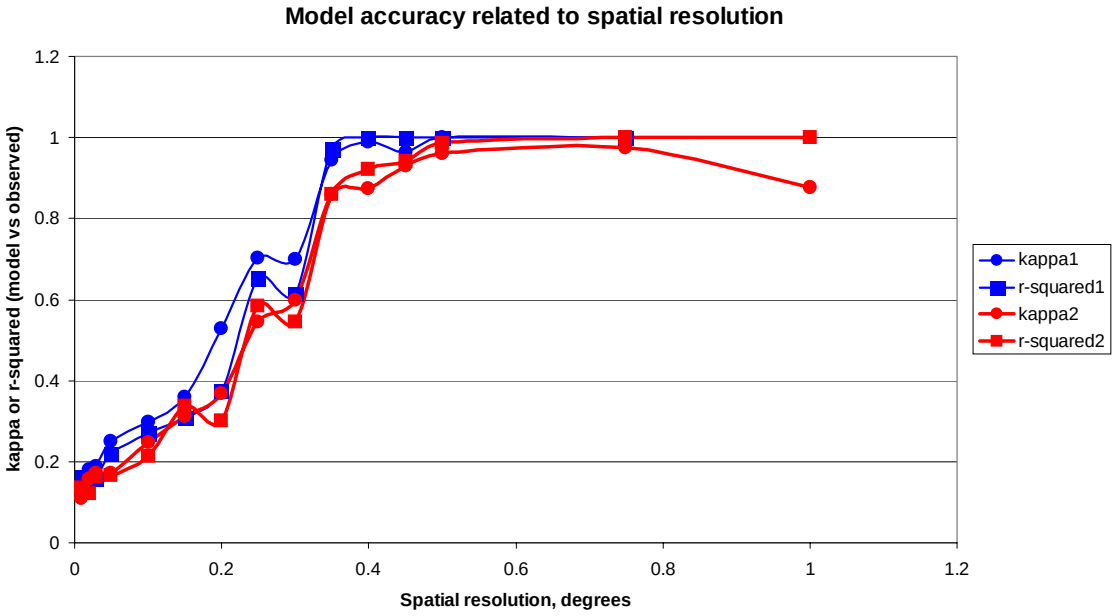
Table 3 shows details from the first model in this series (0.01 degree spatial resolution) and Table 4 shows details from the model at 0.35 degrees resolution. Beyond the latter (i.e. at larger pixel sizes, up to 1 degree resolution), model accuracy was frequently 100%, and a certain degree of over-fitting tended to occur, as indicated by the AIC_c value which began to increase as more variables were added to the list after the first few. Tables 3 and 4 show a number of features of interest, in terms of satellite and other variables correlated with the different levels of poverty. First, mean values of key variables tended to change monotonically across the range of poverty levels (see, for example, the first variable in Table 3, the Tair mean). Second, increasing household expenditure (decreasing poverty) is associated with increases in some variables and decreases in others (for example the first variable in Table 3 – the Tair mean – decreases, but the third variable, CCD variance, increases). Third, we might not expect, and we do not find, that the same variables are predictors of poverty at all spatial scales; for example, there are no NDVI variables at the 0.01 degree spatial scale, but they occupy the top two positions for the 0.35 degree resolution model. Fourth, satellite variables tend to dominate the list of selected variables. Goat density and *G. pallidipes* risk were selected at the highest spatial resolution, *G. pallidipes* risk and distance to markets at the lower spatial resolution. In general, poorer people tend to have more goats (Table 3) and richer people tend to live closer to markets (Table 4), with the exception of the richest of all categories in Table 4. To a great extent, the list of selected variables is a function of the correlation structure of the data. A variable, once selected, will tend to exclude all those other variables with which it is strongly correlated, and will therefore tend to favour the inclusion of other variables with which it is less strongly correlated, if these additional variables have some descriptive power.

Tables 3 and 4 also each show an accuracy table for each poverty category in the respective models. Accuracy tends to be lowest for the intermediate categories of poverty, and highest at the extreme ends. Model accuracy is greatly increased by allowing a plus or minus one category 'hit' to count as an accurate description; and even more so with a plus or minus two category allowance. In Table 3 (0.01 degree resolution) the consumer's accuracy is more even than is the producer's accuracy. This indicates to users of poverty risk maps that each category of poverty on the map is as likely to be correct as are all the other categories. In the particular case of Table 3 (the highest spatial resolution model and therefore probably the worst in terms of predictive accuracy) the accuracy of the poverty map is approximately twice that of a random 'guess' (consumer's accuracy about 20%; random guesses would be about 10% accurate in a 10-category map).

4.8 Map accuracy related to map scale

Figure 17 shows the accuracy of the models made at all spatial scales, from 0.01 to 1 degree. The figure shows the kappa values for models where variables were selected either to maximise kappa itself (blue dots) or to minimise the corrected Akaike Information Criterion (red dots). Each models' prediction was correlated with the observed category of each household's poverty level and the r-squared of this relationship (also shown in Figure 17 with square symbols) therefore indicates how much of the variance of household expenditure is explained by the respective models. R-squared increases rapidly from its low value at 0.01 degree spatial resolution (r^2 about 0.1 – 0.15) to more respectable values at about 0.30 degree spatial resolution (r^2 of about 0.6 to 0.7). It increases further beyond this point, asymptotically approaching 1.0 at spatial resolutions of 0.4 – 0.45 degrees, or beyond (*i.e.* larger pixels). Arguably a useful poverty map can be constructed from satellite and other data at spatial scales of about 20km grid size upwards.

Figure 17: *Poverty risk map model accuracy (y-axis) related to spatial resolution (x-axis). Blue circles – kappa values for models where the variable selection was based on maximising kappa: blue squares – proportion of the variance in the original data explained by these model (r^2). Red circles – kappa values for models where variable selection was based on minimising the corrected Akaike Information Criterion (AIC_c) (Annex C): red squares – proportion of the variance in the original data explained by these models (r^2).*



5. SUMMARY AND CONCLUSION

Here we have presented an approach to mapping poverty that enables us to move beyond description, where the more traditional small area estimates reach their limitation, and towards explaining the distribution of poverty and even to predicting changes in poverty that may result from changing conditions that are associated with the different recorded levels of poverty. A definite advantage of the standard small area technique is that policymakers in many countries are familiar with poverty and inequality indicators (*e.g.* the Foster-Greer-Thorbecke measures, the Gini coefficient, *etc*.) that are regularly reported in country poverty profiles using household surveys. The method developed here is appealing because it produces estimates of these same measures, as well as an indication of the degree of statistical precision of these estimates, for smaller administrative units. Poverty maps constructed in this fashion are more likely to be put to practical use because the statistical underpinnings of the methodology makes them more credible and more readily endorsed than the more commonly found maps based on *ad hoc* methods (*e.g.* maps based on Basic Needs Indicators).

There are two broad approaches that can be taken in the future. The first is to construct a series of attribute layers containing information judged to be of relevance in determining poverty. Agricultural production systems, and human and animal disease maps can be independently constructed from satellite and other data and brought together to 'explain' poverty. Alternatively, direct correlations can be sought between poverty and satellite data, as carried out here, thus short-circuiting the 'explanatory phase' implied by the first approach.

Clearly the first approach is the more desirable because it investigates the proximate causes of poverty. Once identified, these proximate causes can be addressed directly (for example, with an extension program in agriculture or in health). This approach only works, however, if we know in advance the full range of poverty generating processes and can somehow capture them in our data layers to be used in the model.

The merit of the second approach, therefore, is that it brings into the analysis no preconceptions as to what are the ultimate causes of poverty. By investigating the satellite and other variables associated with poverty we have a much more objective view of the potential driving mechanisms through the model's incorporation of their (likely) proxies. For example, rainfall may be a proxy for agricultural production or for malaria, or for some as yet undiscovered but equally important other factor. A poverty mapping exercise based on the first approach might include only agricultural production and malaria, but not the unknown factor, and would therefore miss even the possibility of discovering it. Arguably this third factor might be so closely correlated with either agricultural production or malaria that we have no need of measuring it at all, but this entirely misses the point of poverty mapping in the first place: to understand the processes by which people become poor, or are kept in poverty, in order, sensibly, to reverse them. Poverty mapping is an exercise in development, not in statistics. We are using what statistics we have, and over which we have some control (= knowledge), to investigate something we presently understand only very poorly.

In conclusion, what we have been able to show here is the step beyond exploiting correlations within internally correlated socio-economic data sets (the traditional small area mapping approach) to a situation where we have been able to show that external, independent data appear to have at least as much descriptive power for poverty mapping. The precise interpretation of the correlations obtained here will require more research effort but at least we have shown that this effort is both justified and appropriate. It is time to take poverty mapping out of the realm purely of socio-economics.

REFERENCES

Alderman, H., Babita, M., Lanjouw, J., Lanjouw, P., Makhatha, N., Mohamed, A., Ozler, B. and Qaba, O. (2000) *Is census income an adequate measure of household welfare? Combining census and survey data to construct a poverty map of South Africa*. Washington D.C.: Statistics South Africa and the World Bank.

Arcia, G., Mendoza, H. and Ichan, R. (1996) *Mapa de Pobreza Municipal de Nicaragua*. Research Triangle Instituted, Research Triangle Park, North Carolina.

Appleton, S., Emwanu, T., Kagugube, J. and Muwonge, J. (1999) *Changes in poverty in Uganda, 1992-1997*. Working Paper WPS/99.22, Centre for the Study of African Economies, Oxford University.

Bigman, D., Dercon, S., Guillaume, D. and Lambotte, M. (2000) Community targeting for poverty reduction in Burkina Faso. *The World Bank Economic Review* **14,** 167-194.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Crump, J.R. (1997) Teaching the political geography of poverty. *Journal of Geography* **96,** 98-104.

Davis, B. (2003) *Choosing a method for poverty mapping*. Rome: Food and Agriculture Organization of the United Nations.

Deichmann, U. (1999) Geographic aspects of inequality and poverty. Text for World Bank's Web Site on inequality, Poverty and Socio-economic Performance. at: http://www.worldbank.org/poverty/inequal/index.htm. pp 13.

Elbers, C. and Lanjouw, P. (2000) Intersectoral transfer, growth, and inequality in rural Ecuador. *World Development* **29,** 481–496.

Emwanu T., Okwi, P.O., Hoogeween, J.G. and Kristjanson, P. (2003) *Where are the poor? Mapping patterns of well-being in Uganda*. Kampala: Uganda Bureau of Statistics.

Epprecht, M. (2005) *Geographical Dimensions of Livestock Holdings in Vietnam: Spatial relationships among Poverty, Infrastructure and the Environment*. Pro-Poor Livestock Policy Initiative (PPLPI) Working Paper 24.

Epprecht, M. and Heinimann, A. (2004) *Socioeconomic Atlas of Vietnam: A depiction of the 1999 Population and Housing Census*. NCCR North-South, Hanoi and Berne.

Foster, I. and Shorrocks, A. (1988) Poverty orderings. *Econometrica* **56,** 173-177.

Foster, I., Greer, I. and Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica* **52,** 173-177.

Ghosh, M. and Rao, I.N.K. (1994) Small area estimation: an appraisal. *Statistical Science* **9,** 55-93.

Hastings, D.A. and Dunbar, P.K. (1998) Development and assessment of the Global Land One-km Base Elevation Digital Elevation Model (GLOBE).

International Society of Photogrammetry and Remote Sensing, Archives, **32,** 218-221.

Hastings, D.A. and Dunbar, P.K. (1999) *Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Documentation, Volume 1.0*. Key to Geophysical Records Documentation (KGRD) 34. National Oceanic and Atmospheric Administration, National Geophysical Data Center, 325 Broadway, Boulder, Colorado 80303, U.S.A.

Hay, S.I., Randolph, S.E. and Rogers, D.J. (eds) (2000) *Remote Sensing and Geographical Information Systems for Epidemiology*. San Diego: Academic Press.

Henninger, N. (1998) *Mapping and geographical analysis of poverty and human welfare – Review and assessment*. Report prepared for the UNEP/CGIAR initiative on GIS. Washington D.C.: World Resources Institute, pp 97.

Henninger, N. and Snel, M. (2002) *Where are the poor? Experiences with development and use of poverty maps*. Washington, D.C.: World Resources Institute and UNEP/GRID-Arendal. pp 66. (http://population.wri.org/pubs_description.cfm?PubID=3758).

Hentschel, I., Lanjouw, I.O., Lanjouw, O. and Poggi, I. (2000) Combining census and survey data to trace the spatial dimensions of poverty: A case study for Ecuador. *The World Bank Economic Review* **14,** 147-165.

Howes, S. and Lanjouw, L.O. (1997) *Poverty comparisons and household survey design*. LSMS Working Paper No. 129. Washington, D.C.: The World Bank

Jazairy, I., Alamgir, M. and Panuccio, T. (1992) *The state of world rural poverty: an inquiry into its causes and consequences*. International Fund for Agricultural Development (IFAD). New York: New York University Press.

Malik, S. J. (1998) *Rural Poverty and Land Degradation: A Reality Check for the CGIAR*. CGIAR

Minot, N., Baulch, B. and Epprecht, M. (2003) *Poverty and Inequality in Vietnam: Spatial patterns and geographic determinants*. International Food Policy Research Institute and Institute for Development Studies, Hanoi.

Ndeng'e, G., Opiyo, C., Mistiaen, J. and Kristjanson, P. (2003) *Geographic dimensions of well-being in Kenya. Where are the poor?* Volume 1. Nairobi: The Regal Press Kenya Ltd. pp. 164.

Ravallion, M. (1992) *Poverty comparisons, a guide to concepts and methods*. LSMS Working Paper Number 88. Washington D.C.: The World Bank.

Ravallion, M. (1996) Issues in measuring and modelling poverty. *The Economic Journal* **106,** 1328-1343.

Sen, A. (1976) Poverty: an ordinal approach to measurement. *Econometrica* **46,** 437-446.

Shorrocks, A. (1995) Revisiting the Sen poverty index. *Econometrica* **63,** 1225-1230.

Ssewanyana, N.S., Okidi, A.J., Angemi, D. and Barungi, V. (2004) *Understanding the determinants of income inequality in Uganda*. CSAE WPS/2004-29, Centre for the Study of African Economies, Oxford University.

Statistics South Africa (2000) *Measuring poverty in South Africa*. Pretoria: Statistics South Africa. pp. 107.

UBOS (2004) *Report on the Agriculture Module, piggy-backed onto the Population and Housing Census 2002*. Entebbe: Uganda Bureau of Statistics (UBOS).

UBOS (2003) *Uganda National Household Survey 20002/2003: Report on the socio-economic survey*. Entebbe: Uganda Bureau of Statistics (UBOS), November 2003, pp. 48.

UBOS (2002) *2002 Uganda population and housing census: provisional results*. Entebbe: Uganda Bureau of Statistics (UBOS).

Wint, W. (2001) *Kilometre resolution tsetse fly distribution maps for the Lake Victoria Basin and West Africa*. Report by the Environmental Research Group Oxford (ERGO) to FAO/IAEA Joint division, December 2001.

World Bank (2000) *World development report 2000/2001: Attacking poverty*. New York: Oxford University Press.

World Bank (2006) *World development report 2006: Equity and development*. New York: Oxford University Press.

You, L, and Chamberlin, J. (1994) *Spatial analysis of sustainable livelihood enterprises of Uganda cotton production*. Washington, D.C.: International Food Policy Research Institute (IFPRI) Environment and Production Technology Division (EPTD) Discussion Paper No. 121.

ANNEX A: SATELLITE IMAGERY

From a poverty-mapping perspective the choice of remotely sensed data is determined by the spectral, spatial and temporal resolution capabilities of satellites, briefly defined in the following sections.

A.1 Spectral resolution

Satellite sensors detect reflected sunlight or infra-red radiation emitted by all bodies above absolute zero. Data are most readily available in three to seven wavebands or channels in the human-visible and near-to-thermal infra-red part of the electro-magnetic spectrum (0.3-14 μm wavelengths).

A.2 Spatial resolution

Earth observing satellites produce data with spatial resolutions of 0.61- 2.4m (QuickBird), 1-4m (Ikonos-2), 10-20m (Satellite pour l'Observation de la Terre, SPOT), 30-120m (Landsats 1 - 5) or 15-60m (Landsat 7). Images, made up of picture elements or 'pixels' of these sizes, have swath widths of ~11km (Ikonos), ~60km (SPOT) and 185km (Landsat). The 'vegetation instrument' on SPOT-4 has a spatial resolution of 1km and a 2,250km swath width.

Meteorological satellites have lower spatial resolutions, with pixel sizes down to 1.1km (National Oceanographic and Atmospheric Administration Advanced Very High Resolution Radiometer, NOAA AVHRR), and a correspondingly wider swath width of ~2400 km. Geo-stationary satellites maintain a constant position relative to the earth, giving spatial resolutions of 1-8km (GOES for the Americas) or 2.5-5 km (Meteosat 4-6 for Europe/Africa) and images of the entire Earth half-disk.

A.3 Temporal resolution

Higher spatial resolution satellites have a repeat frequency of 11 (Ikonos), 16 (Landsat) or 26 (SPOT) days. Orbiting meteorological satellites produce two images per day of the entire Earth's surface, whilst geostationary ones produce 2 images per hour to monitor weather systems: both are referred to as 'multi-temporal'.

A.4 New satellites and sensors

New systems promise greater spectral and spatial resolutions and greater signal stability over time. These include Quickbird (5 channels with 0.61-2.4m. resolution and a potential 3 day return time), the Moderate Resolution and Imaging Spectrometer (MODIS, 36 channels with 250m – 1km resolution, 1 to 2 day return time) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER, 14 channels with 15m–90m resolution, 16-day return time) instruments on board the TERRA spacecraft (http://terra.nasa.gov), and the Spinning Enhanced Visible and Infrared Imager (SEVIRI) on the geostationary Meteosat Second Generation

(MSG, 12 channels, half-disk images every 15 minutes) platform (http://www.esa.int).

A.5 Images for poverty mapping

Imagery is adversely affected by atmospheric contamination such as clouds and other aerosols. The low revisit frequency of the higher spatial resolution satellites prevents the recording of important seasonal determinants of pathogen transmission rates. In contrast, frequent images from NOAA-AVHRR, Terra/MODIS(/ASTER) and Meteosat sensors can be combined to produce relatively cloud-free monthly maximum value composites (MVCs), of much greater use in studying dynamical epidemiological processes.

Data from each satellite channel may be used directly to describe biological and other events, or may be processed to produce indices related to ground-based variables such as soil surface temperatures. Commonly-used products include the middle infra-red band (MIR) from AVHRR channel 3, and Land Surface Temperature (LST, derived from AVHRR Channels 4 and 5), both related to the Earth's surface temperature; the Normalised Difference Vegetation Index (NDVI), derived from AVHRR channels 1 and 2, related to plant photosynthetic activity; near-surface air temperature (Tair) derived from LST and vegetation index measurements; and Cold Cloud Duration (CCD) from Meteosat, correlated with rainfall in convective precipitation systems.

ANNEX B: TEMPORAL FOURIER PROCESSING OF SATELLITE DATA

Temporal Fourier analysis describes variations through time of satellite signals as the sum of a series of sine curves with different frequencies and amplitudes. It is applicable to regularly collected data such as maximum value composite monthly AVHRR data, {xt.}, collected over one or more years (for simplicity, temporal Fourier analysis should only be applied to entire years of data, not to partial years).

The Fourier series representation of {xt} is found from the following:

$$x_t = a_0 + \sum_{p=1}^{N/2-1}[a_p \cos(2\pi pt/N) + b_p \sin(2\pi pt/N)] + a_{N/2}\cos \pi t$$

$(t = 1,2,....N)$ B.1

with coefficients $\{a_p, b_p\}$ defined as follows:

$$a_0 = \bar{x}$$

$$a_{N/2} = \sum(-1)^t x_t / N$$

$$\left.\begin{aligned} a_p &= 2[\sum x_t \cos(2\pi pt/N)]/N \\ b_p &= 2[\sum x_t \sin(2\pi pt/N)]/N \end{aligned}\right\} p = 1,.....N/2-1$$

…. B.2

Despite the rather daunting appearance of Equation B.1 its interpretation is relatively straightforward. For any particular value of p in the summation term, Σ, the terms in the square bracket define a single sine curve with a period of N/p time units (frequency = p/N). This is because the sum of a cosine and a sine curve with the same argument ($2\pi pt/N$ in Equation B.1) is another sine curve of the same period but with a different amplitude and displaced in time by an amount dependent on the relative contributions of the cosine and sine curves to the total. These relative contributions are determined by the coefficients a_p and b_p respectively, which therefore fix the amplitude and timing of the peak(s) of the combined curve within the interval from t = 1 to N. Equation B.2 simply describes how to estimate these important coefficients from the sample data, and also suggests that there are ($N/2$) pairs of coefficients of this sort ($aN/2$ can be regarded as the final 'pair' of coefficients, since in this final term $\sin\pi t$ is always zero, and bN/2 is therefore also zero), implying that there are also ($N/2$) different sine curves (each with a different period of oscillation) in the description of xt in Equation B.1. Each of these curves is called an harmonic, so there are $N/2$ harmonics overall.

Rp = the amplitude of the pth harmonic $= \sqrt{(a^2_p + b^2_p)}$

and ϕp= the phase of the pth harmonic $= tan^{-1}(-b_p/a_p)$ B.3

Thus Rp and ϕp uniquely define the amplitude and position of the pth harmonic. Further details of time series analysis may be found in (Chatfield 2004; Diggle 1990).

The inverse Fourier transform is a way of re-constructing the original signal from its (forward) Fourier transform. Since the calculation of the forward transform (Equation B.2) uses the same expressions (sine and cosine curves) as the inverse transform (Equation B.1), a single algorithm may be used, with care, for both forward and inverse Fourier transforms of data sets.

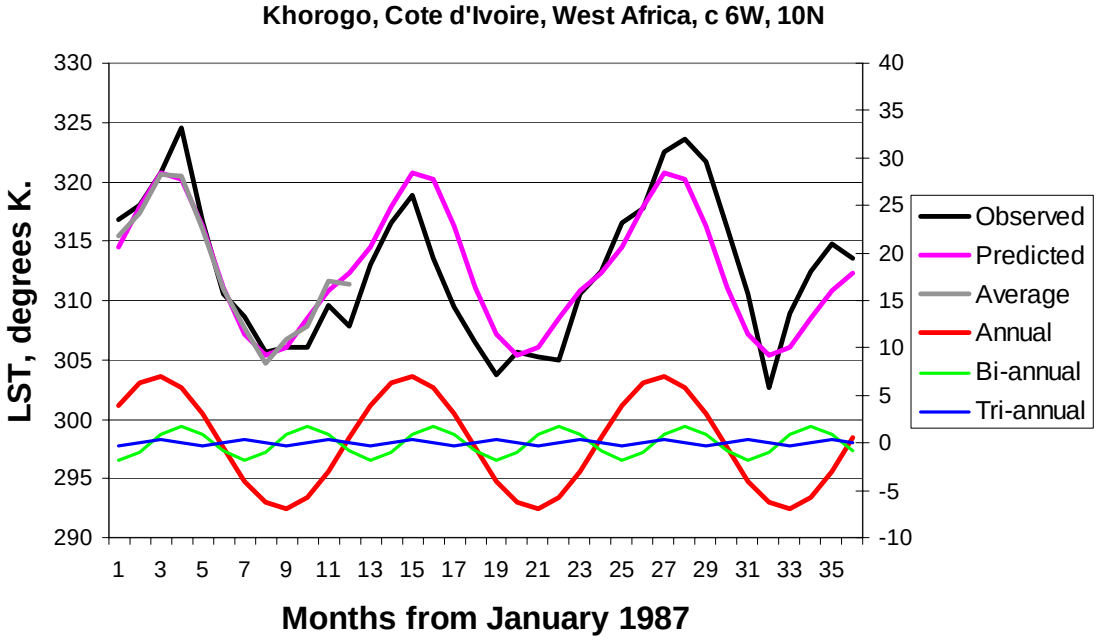
Temporal Fourier analysis has a number of useful characteristics. The sum of all of the harmonics exactly describes the original time series. This means that the harmonics are orthogonal to (*i.e.* uncorrelated with) each other. The variance of each harmonic (for a sine curve this is simply the square of its amplitude) therefore contributes additively to the total variance and each harmonic may be examined in turn to determine its contribution to overall variance. Harmonics of low amplitude may be dropped from re-constructions of any signal from its Fourier transform thereby achieving efficient data ordination (the reduction of a data set without severe loss of information). Similarly the omission of high frequency harmonics often achieves noise reduction (smoothing), a useful operation with time series of remotely sensed data. In work on a variety of insect and tick vectors the harmonics with periods of 12, 6 and 4 months were generally the most useful descriptors of habitat variability and these are here called the annual, bi-annual and tri-annual Fourier components. Figure B.1 illustrates these three important Fourier components of the Land Surface Temperature (LST) and Normalised Difference Vegetation Index (NDVI) time series of satellite data from a single point in northern Côte d'Ivoire, West Africa, and shows how their sum provides a smoothed description of the raw data. Figure B.1 shows how Fourier analysis captures the different shapes of the NDVI and LST cycles, and how it can quantify the time delay between the peak vegetation activity, and peak temperature each year.

Figure B.2 shows examples of temporal Fourier imagery of North America. The data from each pixel in the set of monthly images for the USA were processed in the way shown in Figure B.1 and the resulting details of the mean, amplitudes and phases were stored in a series of output images (one for each variable). The set of temporal Fourier images – 10 images per original satellite data channel (the mean; phases and amplitudes of the annual, bi-annual and tri-annual cycles; the maximum, minimum and variance) – are the inputs into our discriminant analysis models (Annex C).

We prefer temporal Fourier analysis to other methods of data ordination (= the reduction of a large data set to a few useful components, without great loss of information) such as Principal Components Analysis (PCA, the method traditionally favoured by remote sensors) because its output is more readily interpreted biologically. Both temporal Fourier analysis and PCA produce results which have the important statistical property of orthogonality *i.e.* the results additively contribute to explaining the variance of the original signal, so there is little to choose between these two methods on statistical grounds. As biologists, however, we appreciate the importance of annual, bi-annual, tri-annual cycles of physical events (temperature, rainfall, *etc.*) for describing biological processes and such descriptions are available only from temporal Fourier analysis, not from PCA.

Further details of temporal Fourier analysis are given in (Rogers and Williams 1994; Rogers *et al*. 1996; Rogers 1997; Rogers 2000a).

Figure B.1: *Example of temporal Fourier processed Land Surface Temperature (upper) and NDVI (lower) time series from a single point in northern Côte d'Ivoire. In each case three years of monthly AVHRR data are shown as the black lines (the additional grey line in year 1 is the 3-year average). The annual, bi-annual and tri-annual Fourier cycles are shown in red, green and blue respectively (notice the second, zero-centered scale for these on the upper graph, right hand axis) and their sum is shown as the violet line super-imposed on the raw data. Notice how the Fourier decomposition manages to capture subtle details of the seasonal cycle in both variables.*



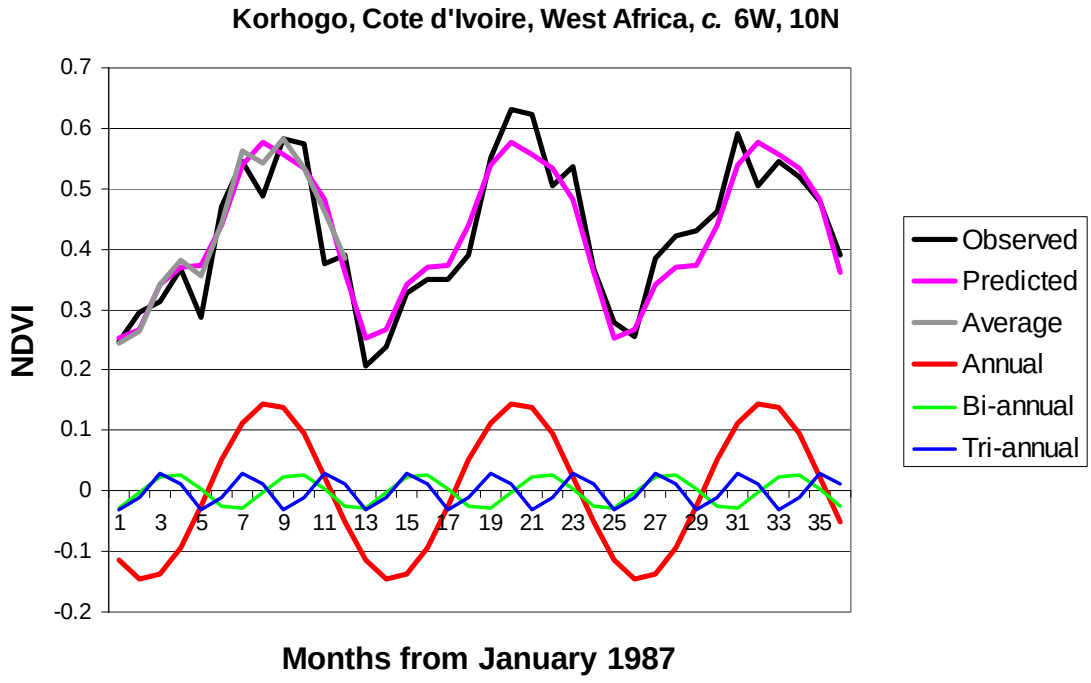
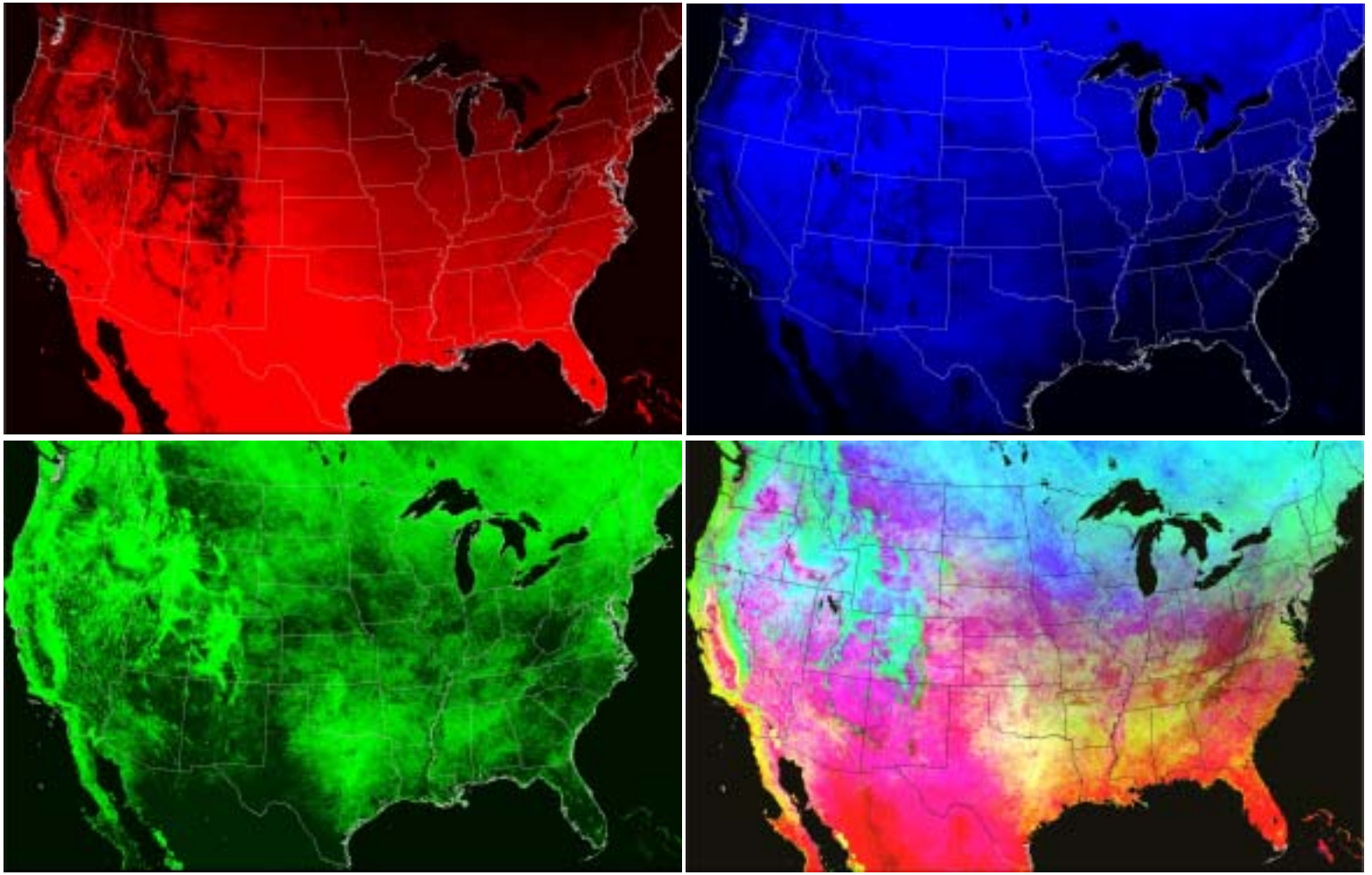


Figure B.2:. *Example of temporal Fourier processed Land Surface Temperature images of the USA. The mean is shown in red (upper left), the annual amplitude in blue (upper right), the annual phase in green (lower left) and the combination of all three images in the multi-colored image, lower right. The red image shows that average temperature is higher in the South and decreases North-wards. The blue image, however, shows that the seasonal variation in temperature is greatest in the North. The green image (where later phase is brighter colored) shows that the seasonal peak of temperature tends to be later in the North. Notice how topographic features - such as the Rocky Mountains - affect the regional patterns.*



References for Annex B

Chatfield, C. (2004) *The Analysis of Time Series: An Introduction*. 6th ed. Boca Raton: Chapman & Hall.

Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction*. Oxford: Clarendon Press.

Rogers, D.J. and Williams, B.G. (1994) *Tsetse distribution in Africa: seeing the wood and the trees*. In "Large-scale ecology and conservation biology. 35th symposium of the British Ecological Society with the Society for Conservation Biology, (1993)." (P. J. Edwards, R. M. May and N. R. Webb. eds.) Chapter 11. University of Southampton: Blackwell Scientific Publications, pp. 249-273.

Rogers, D.J., Hay, S.I. and Packer, M.J. (1996) Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology* **90,** 225-241.

Rogers, D.J. (1997) *Satellite Imagery and the Prediction of tsetse distributions in East Africa*. In "Diagnosis and Control of Livestock Diseases using Nuclear and related techniques". Vienna: International Atomic Energy Agency, pp. 397-420.

Rogers, D.J. (2000) Satellites, space, time and the African trypanosomiases. *Advances in Parasitology* **47,** 129-171.

C.1 Discriminant analytical methods

In its simplest form, discriminant analysis assumes both multi-variate normality and a common within-group co-variance of the variables for all points defining vector, disease or poverty presence and absence. Co-variances are estimated from the training set. Means of multi-variate distributions are referred to as centroids and are defined by mathematical vectors $\{ \bar{\mathbf{x}}_v \}$ where v is the number of dimensions (= variables) (here we follow the usual convention that heavy type indicates a row or column vector, or a matrix, and the bar above a variable indicates the mean). The Mahalanobis distance, D^2, is the distance between two multi-variate distribution centroids, or between a sample point and a centroid, and is defined as follows:

$$D^2_{12} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{C}_w^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$
$$= \mathbf{d}' \mathbf{C}_w^{-1} \mathbf{d} \qquad \text{.... C.1}$$

where the subscripts now refer to groups 1 (*e.g.* for vector absence) and 2 (*e.g.* for vector presence), $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ and $\mathbf{C}^{-1}{}_w$ is the inverse of the within-groups covariance (dispersion) matrix (Green 1978) (the single quotation, ', indicates the transpose of a row or column vector). Thus D^2 is the Mahalanobis Distance between the sample centroids adjusted for their common co-variance. Equation C.1 may be used in a number of ways. Firstly it may be used to assign new data points to one or other category (of presence or absence) by examining the value of D^2 between each point and each of the training-set defined centroids. The point is then assigned to the group for which D^2 is a minimum. Secondly, the equation may be used to calculate the probability with which each data point belongs to each of the training set groups. This involves defining the position of the point within the multi-variate distribution around each centroid (most easily achieved by calculating D^2 which is distributed as χ^2 with *(v-1)* d.f., where *v,* as before, is the number of variables defining each centroid). In general these measures are normalised by dividing each by the sum of all measures (*i.e.* the sum of the probabilities across all classes in the training set) to give posterior probabilities, defined as follows:

$$P(1|\mathrm{x}) = \frac{p_1 e^{-D^2{}_1/2}}{\sum_{g=1}^{2} p_g e^{-D^2{}_g/2}}$$

and C.2

$$P(2|\mathrm{x}) = \frac{p_2 e^{-D^2{}_2/2}}{\sum_{g=1}^{2} p_g e^{-D^2{}_g/2}}$$

where P(1 | x) is the posterior probability that observation **x** belongs to group 1 and P(2 | x) the posterior probability that it belongs to group 2 (Green 1978) (the exponential terms in Equation C.2 are those of the multi-variate normal distributions defining groups 1 and 2; all other terms of the multi-variate distributions are the same in numerator and denominator and therefore cancel out (Tatsuoka 1971). In Equation C.2, p_1 and p_2 are the prior probabilities of belonging to the same two groups respectively, defined as the probabilities with which any observation might belong to either group given prior knowledge or experience of the situation. In the absence of any prior experience it is usual to assume equal prior probability of belonging to any of the groups; in the simple case of two-group discrimination, therefore, $p_1 = p_2 = 0.5$. Great care should be taken with the normalisation step of Equation C.2 since it assumes that observation **x** must come from one or other of the classes defined in the training-set data. This emphasises the importance of carefully selecting the training set to be representative of all possible presence and absence sites, not just some of them. In general it is advisable to produce along with the output image of predicted probabilities a second image of the Mahalanobis Distance to the nearest cluster in the training set *i.e.* the cluster to which each pixel is assigned. This image can then be examined to find areas where the Mahalanobis Distances are very large and therefore where predictions are likely to be inaccurate.

As indicated earlier, Equations C.1 and C.2 should be modified when the assumption of common covariances is obviously invalid. Areas of different degrees of poverty may well differ in their environmental characteristics, requiring separate multi-variate descriptions of their climatic conditions. Since discriminant analysis is based on categorical assignments, it is first necessary to split up the continuous poverty variable into a series of bins. Each level of poverty (*i.e.* each 'bin') is then treated in the model as a separate multi-variate normal distribution, with its own co-variance characteristics, and the posterior probabilities are calculated for each bin separately. In the simplest case of two bins or groups only (for example one for poor and one for non-poor), Equation C.2 is then modified as follows:

$$P(1|\mathbf{x}) = \frac{p_1 \left|\mathbf{C}_1\right|^{-1/2} e^{-D^2{}_1/2}}{\sum_{g=1}^{2} p_g \left|\mathbf{C}_g\right|^{-1/2} e^{-D^2{}_g/2}}$$

and C.3

$$P(2|\mathrm{x}) = \frac{p_2 \left|\mathbf{C}_2\right|^{-1/2} e^{-D^2{}_2/2}}{\sum_{g=1}^{2} p_g \left|\mathbf{C}_g\right|^{-1/2} e^{-D^2 g/2}}$$

where $|\mathbf{C}_1|$ and $|\mathbf{C}_2|$ are the determinants of the co-variance matrices for groups g = 1 and 2 respectively. The Mahalanobis distances in Equation C.3, calculated from Equation C.1, are evaluated using the separate within-group co-variance matrices $\mathbf{C_1}$ and $\mathbf{C_2}$ (Tatsuoka 1971). When there is more than a single class of presence or absence data (*e.g.* multiple categories) the summation in the denominators of Equation C.3 covers the entire set of g>2 groups and there are as many posterior probability equations as there are groups. With unequal co-variance matrices the discriminant axis (strictly speaking a plane) that separates the two groups in multi-variate space is no longer linear, and Equation C.3 then effectively defines the maximum likelihood solution to the problem (Swain 1978).

There is no obvious rule about the use of expected or observed prior probabilities in Equations C.2 or C.3. Use of observed (generally training-set) prior probabilities shifts the

equi-probability contours towards the smaller groups, resulting in a larger proportion of assignments to the classes with larger group sizes. This shift frequently increases predictive accuracy. In the present case, however, the poverty data were split (binned) to give approximately equal sample sizes, so the prior probabilities were about equal.

Further details of multi-variate analysis may be found in several useful texts (Tatsuoka 1971; Green 1978; Krzanowski and Marriott 1995; Legendre and Legendre 1998). Further details of the application of these techniques to vector and disease mapping are given in (Rogers and Randolph 1993; Rogers and Williams 1994; Rogers *et al*. 1996; Robinson *et al*. 1997; Rogers 1997; Rogers 2000; Rogers 2006).

C.2 Application to poverty mapping

As indicated above, there is a variety of criteria by which variables might be selected during analysis. These are outlined in Annex D. At each step, each variable is examined in turn for its ability to maximize the test statistic. This procedure should maximise the ability of the technique to distinguish different levels of poverty. In each run of the model the number of bins into which to divide the poverty data can be varied – in the present example it was eventually felt that a set of 10 bins gave the best results, and only these are reported here.

C.3 A brief introduction to the information-theoretic approach

A significant contribution to the modelling and variable selection approach described above is provided by the work of Burnham, Anderson and others who promote what is called an information-theoretic approach that appears to tackle a number of problems that arise with the more traditional approach (Burnham and Anderson 2002). Whilst much of what Burnham and Anderson write about concerns the philosophy of model building and the rôle of Null Hypotheses in this activity, their work also has implications for the ways in which models of different structure are compared and (importantly for present purposes) how variables should be selected during the modelling phase.

In this approach it is assumed that there exists an n-dimensional and unknowable truth (the real distribution of diseases in the present case) that models can only attempt to approximate rather than describe completely. There exists, therefore, a certain distance ($I(f,g)$) between model (g) and reality (f) that is captured by the Kullback-Leibler information or distance measure which is defined as:

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx \qquad \text{.... C.4}$$

for continuous functions and

$$I(f,g) = \sum_{i=1}^{k} p_i . \log\left(\frac{p_i}{\pi_i}\right) \qquad \text{.... C.5}$$

for discrete distributions such as the Poisson, binomial *etc*. In Equation C.4, full reality f is considered fixed whilst g varies over a range of models indexed by θ. In C.5 there are k

possible outcomes of the underlying random variable. The true probability of outcome i is p_i, whilst the modelled outcome is π_i, with $\sum p_i = \sum \pi_i = 1$.

These rather fearsome looking equations are really quite simple. It is obvious from both, for example, that in the unlikely event that the models perfectly describe reality, $g(x|\theta) = f(x)$ in Equation C.4 and $\pi_i = p_i$ in Equation C.5. The logarithmic terms will therefore be zero (because log(1)=0) and the Kullback-Leibler distance, $I(f,g)$, will thus also be zero in each case. The greater is the discrepancy between model and reality, the larger will $I(f,g)$ become. Thus the K-L distance is a guide to model accuracy and may be used to select the best from a set of candidate models for any particular situation.

There is one obvious problem, however, and that is that we do not know in each case what the truth ($f(x)$ or p_i) actually is. Taking the continuous case as an example, Equation C.4 can be re-arranged as follows:

$$I(f,g) = \int f(x)\log(f(x))dx - \int f(x)\log(g(x\mid\theta))dx \qquad \text{.... C.6}$$

with the following statistical expectations

$$I(f,g) = E_f[\log(f(x))] - E_f[\log(g(x\mid\theta))] \qquad \text{.... C.7}$$

each with respect to the distribution f. The first expectation on the right of Equation C.7 will be unknown (because it is the expectation of reality) but constant (reality does not change!). The second expectation on the right of Equation C.7 will vary, depending both upon the model and its current parameters. This means that although $I(f,g)$ cannot be evaluated exactly, it can be estimated up to a constant C (viz. $E_f[\log(f(x))]$)

$$I(f,g) = C - E_f[\log(g(x\mid\theta))]$$

or $$\text{.... C.8}$$

$$I(f,g) - C = -E_f[\log(g(x\mid\theta))]$$

The left hand side is a relative directed distance between f and g and thus the value of the right hand side can be used to select between different candidate models. A model with a lower value of this quantity is better than one with a higher value. Because we do not know C we can never know just how good our 'best' model really is, but the difference between models is a guide to how much better is our best model than any others in the candidate set.

In the discussion so far it is assumed that the parameters of the candidate models are already known. In reality they must be estimated from a set of data. Akaike showed that in

practice the K-L distance could be estimated from the empirical log-likelihood function evaluated at its maximum point. The practical equivalent of Equation C.8 is what has since become known as the "Akaike Information Criterion" or AIC, defined as follows:

$$AIC = -2\log(\ell(\hat{\theta} \mid y)) + 2K \qquad \text{.... C.9}$$

where $\log(\ell(\theta|y))$ is the value of the log-likelihood at its maximum point (*i.e.* the maximum likelihood estimate) and K is the number of estimated parameters in the model. It is clear from Equation C.9 that the first term on the right-hand side will tend to decrease as the number of parameters in the model increases (because a model with more parameters is almost bound to fit a dataset better than one with fewer parameters) whilst the second term ($2K$) will obviously increase. This achieves a neat balance between over-fitting a model (too many parameters, AIC penalised with a large value of $2K$) and under-fitting a model (too few parameters, AIC large because the first term is large).

A modification of the AIC was suggested by Hurvich and Tsai (1989) for the situation where the sample size is small in relation to the number of fitted parameters. This modification, the corrected AIC or AIC_c, is calculated as follows:

$$AIC_c = -2\log(\ell(\hat{\theta} \mid y)) + 2K\left(\frac{n}{n-K-1}\right) \qquad \text{.... C.10}$$

where n is the sample size and all other terms are as in Equation C.9. In general, unless the sample size is large in relation to the number of estimated parameters, Equation C.10 is to be preferred over Equation C.9.

The modelling approach recommended by Burnham and Anderson involves proposing a set of candidate models for the biological situation involved, then fitting these models to the data and calculating the AIC or AIC_c values. As mentioned before, the absolute values of these quantities are usually of little interest, but differences between them are very informative. The AIC difference (Δ_i) is defined as follows:

$$\Delta_i = AIC_i - AIC_{\min} \qquad \text{.... C.11}$$

where AIC_{min} is the minimum AIC for any candidate model in the set of models, and the model with this minimum value is the current best one. Despite the very wide possible range of absolute values of AIC, AIC differences of approximately greater than 10 indicate models that have very little support and therefore can be omitted from further consideration, whilst AIC differences of less than 2 are indicative of strong support. Given any particular

set of models, the likelihood of one of the models within the set (g_i), given the data, is proportionately related to the AIC difference by the following:

$$\ell(g_i \mid x) \alpha \exp\left(-\frac{1}{2}\Delta_i\right) \quad \text{.... C.12}$$

These likelihoods are usually normalised across the entire set, R, of candidate models to determine a set of Akaike weights, w_i that sum to 1.0:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)} \quad \text{.... C.13}$$

These weights are thus an effective way to scale and interpret the AIC difference values.

Equations C.11 to C.13 involve comparisons *between* models, and Equation C.13 refers to a particular set of models. Thus one can only conclude that a particular model has an Δ_i, likelihood or Akaike weight relative to some one (Δ_i) or all other models (w_i) in a particular set of models. Choice of a candidate set of models therefore becomes crucial. If a candidate model is dropped from the set, or a new model is added, the various quantities should be recalculated. However, a quantity called the evidence ratio w_i/w_j, where *i* and *j* are just two of the candidate models, is *not* affected by any other model in the candidate set, but just by the two models being compared. Evidence ratios may be used to judge how much better one model is compared with another, regardless of any other models in the candidate set.

For some biological systems where the mechanisms are fairly well understood, the set of candidate models may be easy to define. For example if we seek a model for plant growth we might generate a series of potential models that involve the quantity of available sunlight, water or soil nutrients, in various combinations. The information-theoretic approach is ideal in this situation because what we really seek is some idea of the relative importance of variables we know, or suspect, to be of importance. In the case of poverty levels, however, it is much more difficult to identify a set of 'reasonable' predictor variables (all may be important in one way or another) and so we tend to fall back upon the step-wise or data-mining methods described earlier in this Annex. Although we defend our own application of such step-wise methods, the same methods are often employed in data-mining exercises when a variable or quantity of interest (*e.g.* stock prices, car sales) is modelled using large collections of potential predictor variables. It was precisely the mindless application of data-mining methods that Burnham and Anderson's approach was designed to avoid. As they point out, given a sufficient number of potential predictor variables, data-mining methods are bound to come up with some or other descriptive model. Nevertheless it seems that we can learn from the information-theoretic approach even for poverty modelling. For example we could generate a set of candidate models which described poverty levels using different sorts of variables (socio-economic, environmental,

temperature, humidity, vegetation indices, *etc*.) and select between them. Burnham and Anderson are sympathetic to this approach, if only because it is, in their view, the lesser of two evils: *"While we do not condone the use of information theoretic approaches to blatant data dredging, we suggest it might be a more useful tool than hypothesis testing in exploratory data analysis where little a priori knowledge is available. Data dredging has enough problems and risks without using a testing-based approach that carries its own set of substantial problems and limitations."* (Burnham and Anderson 2002).

The information theoretic approach provides a completely different paradigm from the traditional statistical approach to model building. There are no formal levels of any test statistic that determine 'significance' of one result over another, and therefore no formal hypothesis testing either. As Burnham and Anderson point out, there are many areas of life and science that involve numbers that do not readily fall within the realm of traditional statistical testing. For example one does not ask for a formal test of significance if a soccer match is won by 3 goals to 1 or by 10 goals to 1. One infers that the winners in the second match were considerably better than their opponents, in comparison with the winners of the first match. How large should be the differences in goals scored for them to be judged 'significant' is irrelevant in this case. The match results simply give us evidence for the greater superiority of the winners (compared with the losers) of the second match compared with the first, and allow us to rank the teams in a tournament situation. Model selection and multimodel inference is in many ways more like a tournament. We seek the best possible candidate from a whole suite of models to do the job we have in hand. We are able to say how much better is this model compared to all the other models we have constructed, and we are able to discard at least some models because some or other information metric (the Akaike weight, or the evidence ratio) puts them so much lower than the current best model. There are, however, no threshold values for any of these metrics, signifying 'significant' in one case or 'not significant' in another, because such formal statistics are inappropriate in this situation. Burnham and Anderson go so far as to say that the use of null hypothesis testing for model selection must be considered *ad hoc* (albeit a rather refined set of *ad hoc* procedures), whereas there is a sound theoretical basis to the information-theoretic approach to model selection criteria. (There remains a role for formal hypothesis testing in more experimental situations where the experimenter can define treatment and control groups that differ only in a single or limited number of variables, although even here it is not so much the significance of the effect that is of interest, but the size of the effect.)

C.4 The information-theoretic approach to poverty mapping

Since the output of discriminant analysis can be expressed as a probability (strictly a Bayesian posterior probability of belonging to a particular category of poverty), the likelihood ℓ is simply

$$\ell = \prod_i \Pr(Y_i) \qquad \text{.... C.14}$$

where $\Pr(Y_i)$ is the probability of the observed outcome, defined as

$$\Pr(Y_i) = P_i^{Y_i} Q_i^{1-Y_i} \qquad \text{.... C.15}$$

where P_i is the predicted probability of belonging to the observed category of poverty – the one that was actually observed for this data point ($Y_i = 1$), and Q_i is the complement of P_i (*i.e.* the probability of belonging to any of the other categories of poverty).

The most convenient form of the log. likelihood function of Equations C.14 and C.15 is the following:

$$\log(\ell(\hat{\theta} \mid y)) = \sum_{i \in A_1} \log P(x_i, \theta) + \sum_{i \in A_0} \log Q(x_i, \theta) \quad \ldots. \text{C.16}$$

where A_1 and A_0 denote the sets of observations with Y=1 and Y=0 respectively (Cramer 2003). Thus it is possible to calculate the corrected Akaike Information Criterion AIC_c from Equations C16 and C10. One could therefore use this to select between models and, equally importantly, to decide how much better the best model is compared with the others. It is this approach that is used for variable selection in one set of the models reported in this working paper.

The Akaike weights are also useful in helping to determine the relative importance of the predictor variables. If the current best model contains variable x_1, say, but has only a modest Akaike weight, then it is clear that there is considerable model uncertainty and therefore only weak evidence for the importance of x_1 as a predictor variable. However the Akaike weights can be summed for all models across the set that contain x_1, or x_2, or x_3 *etc*. and these summed weights reflect the relative importance of these variables across all models (Burnham and Anderson 2002). It will generally happen that the sum of the Akaike weights for a variable will exceed the Akaike weight of the best model (in which the variable may or may not occur) and it can also happen that a variable not in the 'best' model can have a summed Akaike weight that exceeds that of any other variable, even those included in the 'best' model. These summed Akaike weights therefore highlight the relative importance of each variable regardless of which models the variable occurs in. This procedure can also be extended to pairs of variables, or to interaction effects between variables (if interaction terms are included in the candidate models). For the correct conclusions to be drawn about any particular variable, it is advisable to use a set of candidate models in which the variables being compared occur about the same number of times. Obviously this is more likely to be the case for important variables than for unimportant ones.

Needless to say the issue of the type of models we should use for distribution mapping, and the criteria we should use for variable selection, are still debated. A recent review of distribution modelling strongly favours the information theoretic approach (Rushton *et al*. 2004). A later article in the same journal redresses the balance with a plea for pluralism (Stephens *et al*. 2005). Some of the issues raised by these articles are discussed further in (Rogers 2006). For present purposes the information-theoretic approach provides us with a new method for selecting predictor variables. During each round of step-wise inclusion the variable is selected that minimizes the AIC_c. In other words the selected variable brings the model closer to the 'truth' than any other candidate variable. In practical terms the real difference between variable selection using AIC_c and, for example, some variant of the kappa statistic is that the latter method is calculated based on the categorical prediction of training set observations: a prediction is judged either correct or not, depending upon thresholding the calculated posterior probability (Equation C.3) such that a correct prediction (of a poverty category) is predicted for $p_{posterior} \geq 0.5$, and an incorrect prediction occurs otherwise. Thus the prediction is judged correct whether $p_{posterior} = 0.51$ or 0.99, and kappa will therefore be the same in each case. The AIC_c statistic, however, uses the posterior probabilities directly, in which case AIC_c will indicate a better model fit for the second outcome ($p = 0.99$) than for the first ($p = 0.51$). Importantly AIC_c also indicates just how much better is a model with *n+1* predictor variables compared with one containing only *n* variables. The penalty for including more variables (parameter *K* in Equation C.10) tends not to be severe when there are large numbers of training set data points; in such cases,

therefore, the more predictor variables, the lower is the AIC_c (and most reductions in this quantity exceed 10, indicating that each variable is making a susbstantial improvement in model fit) . With fewer data points, however, AIC_c tends to decrease initially but then increases as more variables are added. This was certainly noticeable when modelling poverty at the coarser spatial scales, when there were relatively few observations in each poverty group.

References for Annex C

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Cramer, J.S. (2003) *Logit Models from Economics and Other Fields*. Cambridge: Cambridge University Press.

Green, P.E. (1978) *Analyzing Multivariate Data*. Hinsdale, Illinois: The Dryden Press.

Hurvich, C.M. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Krzanowski, W.J. and Marriott, F.H.C. (1995) *Multivariate Analaysis Part 2: Classification, Covariance Structures and Repeated Measurements*. London: Arnold.

Legendre, P. and Legendre, L. (1998) *Numerical Ecology*. Amsterdam: Elsevier.

Robinson, T.P., Rogers, D. and Williams, B. (1997) Mapping tsetse habitat suitability in the common fly belt of Southern Africa using multivariate analysis of climate and remotely sensed vegetation data. *Medical and Veterinary Entomology* **11,** 235-245.

Rogers, D.J. and Randolph, S.E. (1993) Distribution of tsetse and ticks in Africa: past, present and future. *Parasitology Today* **9,** 266-271.

Rogers, D. J. and Williams, B. G. (1994) *Tsetse distribution in Africa: seeing the wood and the trees*. In Large-scale ecology and conservation biology. 35th symposium of the British Ecological Society with the Society for Conservation Biology, (1993). Chapter 11 (Eds, Edwards, P. J., May, R. M. and Webb., N. R.), pp. 249-273. Oxford: Blackwell Scientific Publications.

Rogers, D. J., Hay, S. I. and Packer, M. J. (1996) Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology* **90,** 225-241.

Rogers, D.J. (1997) Satellite Imagery and the Prediction of tsetse distributions in East Africa. In "Diagnosis and Control of Livestock Diseases using Nuclear and related techniques". Vienna: International Atomic Energy Agency, pp. 397-420.

Rogers, D.J. (2000) Satellites, space, time and the African trypanosomiases. *Advances in Parasitology* **47,** 129-171.

Rogers, D.J. (2006) Models for Vectors and Vector-borne diseases. *Advances in Parasitology* **62** (*in press*).

Rushton, S.P., Ormerod, S.J. and Kerby, G. (2004) New paradigms for modelling species' distributions? *Journal of Applied Ecology* **41**, 193-200.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. and Martinez del Rio, C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology* **42**, 4-12.

Tatsuoka, M.M. (1971) *Multivariate Analysis: Techniques for Educational and Psychological Research*. New York: John Wiley & Sons.

ANNEX D: MODEL ACCURACY

Here we deal first with the accuracy of binary predictions – of presence or absence of something, for example a disease; or the extreme poor versus all others. Then we discuss multiple categories (*e.g.* of household income groups). In many cases the same accuracy metrics can be applied to these different sorts of problem. Some of the above measures of accuracy included in the discussion below are reviewed by Congalton (1991).

D.1 Accuracy Metrics

Risk maps contain predictions about the training set sites that are either correct or not. However, producers of risk maps (generally the modellers, also know as 'Producers' in some accuracy metrics) and users (also known as 'Consumers' in some accuracy metrics) of the maps often have different objectives, and so favour different accuracy metrics, a brief summary of which is provided here and in Table D.1. A simple measure of accuracy is the overall percentage correct fit. This metric works for both binary (*e.g.* presence/absence) and multiple-category (*e.g.* poverty category) models. One disadvantage is that it does not distinguish whether there are more errors in the predictions of absence or of presence – 'mistakes' that might have quite different consequences from the users' point of view. Another disadvantage is that the metric is affected by the relative numbers of the presence and absence categories. A map with many more absence than presence observations can give a high overall accuracy, but may still describe presence sites poorly. In all such cases it is wise to look at the errors. In the case of a simple risk map of the presence or absence of a disease, for example, incorrect predictions are of two sorts: false positives (a false prediction of presence) and false negatives (a false prediction of absence). The significance of each is quite different. A false positive prediction indicates an area judged to be suitable for the disease or vector that is currently free of it. It is a common ecological observation that not all sites that are suitable for any organism are actually occupied by it. For geographical reasons the organism may never have arrived there; for historical reasons it may have experienced a population crash and become (temporarily) extinct there by the time the training set sample was taken. The same applies to diseases, especially those that are spreading within new continents. It takes some time for a disease to spread to occupy all suitable niches and, in the *interim*, risk maps for it are likely to show many false positives. Thus false positives do not necessarily indicate errors in modelling disease distributions, and should therefore not be used to conclude that a model is inaccurate or just plain wrong.

False negative predictions – incorrect predictions of absence – give greater cause for concern. When a disease occurs in an area predicted to be 'unsuitable' for it, there is clearly something wrong with the prediction, and probably therefore the underlying model. False negative predictions form a much firmer basis for questioning the accuracy of a predictive presence/absence model.

Clearly models must reach an acceptable compromise between false negative and false positive predictions. A model which predicted absence everywhere would have zero false positives, but a high level of false negatives; a model predicting presence everywhere would show the reverse. A correct balance between false negatives and false positives is the mark of a good model. In our risk mapping exercises, and for the reasons mentioned above, we always pay more attention to false negative than false positive predictions. With very few exceptions our presence/absence models have more false positive that false negative predictions and, in absolute terms, very few of the latter.

Sometimes modellers and users are interested in the diagnostic capabilities of the model – whether or not it can identify known positive and negative sites accurately. Here remote sensing borrows from the medical and veterinary worlds. 'Sensitivity' is the ability of a model correctly to identify known positive sites and 'specificity' is its ability to identify known negative sites. Again either measure on its own conveys some information about model performance but both measures are required to establish overall model reliability. By separating out sensitivity and specificity, however, it is sometimes easier to identify where a model is going wrong. This separation also allows us to weight our management decisions, since mis-diagnosing a positive disease case might be more serious than mis-diagnosing a negative case. The model can therefore be adjusted to increase whichever metric reduces the error rate of the more expensive 'mistake'.

Related to sensitivity and specificity are two other measures of accuracy targeted at different types of users. The Producer's Accuracy is the accuracy of a model in predicting the training set data. It answers the question "Of all known disease sites, both presence and absence, what percentage does the model classify correctly". It is thus equivalent to the percentage correct fit (see above). The Consumer's Accuracy answers the question "Of all sites predicted to be of a certain type, what percentage of them actually are of that type?" Clearly from the modeller's point of view the Producer's Accuracy may be of more interest, but from the application's point of view, the Consumer's Accuracy is more relevant. This accuracy will allow the user to know in advance the likelihood that any site predicted to have a certain level of a disease will have been correctly identified. Intervention decisions can then be taken with a known level of errors of commission and omission (*e.g.* treating, with insecticides, sites that do not need it, or not treating sites that do).

Another useful measure of model accuracy is the kappa index of agreement, which derives from both the psychometric and remote sensing literature (with rather different applications in the two fields). Kappa varies between -1 (predictions totally opposite to observations) through zero (no better than random agreement of predicted and observed) to 1.0 (perfect fit of the model to the data), and it is generally accepted by the remote sensing community that the following definitions apply to ranges of the kappa statistic: poor, $\kappa < 0.4$; good, $0.4 < \kappa < 0.75$ and excellent, $\kappa > 0.75$. These values are traditionally applied to the land-cover classification of high-resolution Landsat imagery, but we have found that they also seem reasonable for quantitative models of disease intensity. Our brief experience with poverty mapping suggests that the same values are acceptable here, too.

A metric much favoured by those who use logistic regression modeling is the 'Area Under the Curve', or AUC plot (also known as the Receiver Operating Characteristics, or ROC plot), where the Sensitivity is plotted on the y-axis and (1 – Specificity) is plotted on the x-axis (Fielding and Bell 1997). The x- and y-values are determined at several different threshold probabilities (*i.e.* cut-off points separating predicted presence and absence), and these points are then joined by a smooth curve. Clearly at an extremely low threshold probability, all sites are predicted to be disease-present: sensitivity is 1.0 and specificity is zero, hence (1-Specificity) is also 1.0. At an extremely high threshold probability all sites are predicted to be disease-absent: sensitivity is zero and specificity is 1.0, therefore (1 - Specificity) is also 0.0. These two extremes define the two points 0,0 and 1,1 in the x-y plane, and intermediate threshold probabilities define points that join these two extremes. Just as in the case of calculating kappa, models with no skill should nevertheless occasionally make correct predictions, purely on a random basis, and this is indicated by a straight line on the AUC plot, joining 0,0 with 1,1. A good model gives a curve on these axes which is concave to the origin: from a very low level of sensitivity on the y-axis, sensitivity increases rapidly whilst Specificity also stays high, and therefore (1-Specificity) remains low. The steeper the increase in this line as it departs from the origin, the greater will be the final Area Under the Curve, and the better the model fit to the training set data.. The AUC plot is favoured by logistic regression users because the threshold probability that gives the best fit of the model to the data is often not 0.5: in general it is the probability at which the slope of the AUC curve is 1.0.

Finally the information-theoretic approach described in Annex C has recently been proposed for modelling and this usually involves the use of what are called 'information criteria' to assess model performance. Burnham and Anderson's book describes the approach in some detail (Burnham and Anderson 2002) and describe Akaike's Information Criterion (AIC), or small-sample variations of this, as the yard-stick by which to assess model performance. For the present application the corrected AIC (AIC_c) is the most appropriate and this was used in developing the models described in this report. It is calculated from the predicted posterior probabilities of belonging to the observed categories of poverty (as outlined in Annex C). The value of the AIC_c is high for models that describe the data only poorly, and decreases as model descriptive power increases. In the information-theoretic approach the concept of statistical significance is reduced but nevertheless the absolute differences in the AIC_c values of a series of models allows us to identify which model describes the data best and to decide how much better this model is than other candidate models (*e.g.* with fewer or more parameters) being considered in the current 'set'.

D.2 Accuracy metrics for quantitative risk maps

The above discussion concerned mainly binary data sets (presence/absence; poor/non-poor, *etc.*). When we have training set information on some continuous variable like household expenditure, our modelling approach follows a different course of binning the data as outlined above. We divide the range of the poverty metric into a small number of categories. We then model using essentially the same maximum likelihood techniques as before. Errors in this sort of modelling are rather more difficult to quantify with a single accuracy metric. For example, an incorrect prediction of poverty category 2 in a county or area that actually recorded poverty category 1 is a less serious error than is a prediction of poverty category 10 for the same area. Although we may be able to use many of the metrics used for presence/absence risk mapping (previous section) they will not distinguish a 'near miss' from a 'far miss' of this sort. Ideally we need to look carefully at the accuracy metric applied to each category separately, and somehow adjust for near and far misses in predictive skill. We can do this for all the metrics listed in Table D.1, with the possible exception of the Area Under the Curve (it would be difficult to justify on theoretical grounds the application of the AUC to each poverty category separately, although we might do so for a single category of great importance, such as the 'extreme poor').

References for Annex D

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing and Environment* **37,** 35-46.

Fielding, A.H. and Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24,** 38-49.

Table D.1: *Accuracy metrics used in geographic information systems (GIS) and remote sensing.*

Accuracy metric	Range of values	Description	Advantages	Disadvantages
% Correct	0 to 100%	Overall percentage accuracy, all categories combined.	Simple and easy to calculate.	Presence and absence sites given equal weight. Metric usually affected by prevalence.
% False positives	0 to 100%	% of total training set sample wrongly predicted as 'presence'.	Simple and easy to calculate.	Should be considered with its complement – false negatives.
% False negatives	0 to 100%	% of total training set sample wrongly predicted as 'absence'.	Simple and easy to calculate.	Should be considered with its complement – false positives.
Sensitivity	0 to 1	Ability correctly to identify positives.	Derived from diagnostics. Useful measure of positive test accuracy.	Concentrates on positives only. Should be considered with its complement – specificity.
Specificity	0 to 1	Ability correctly to identify negatives.	Derived from diagnostics. Useful measure of negative test accuracy.	Concentrates on negatives only. Should be considered with its complement – sensitivity.
Producer's Accuracy	0 to 100%	Ability to predict correctly the training set data.	A guide to the modeller to identify where current models are wrong.	Not particularly useful to users.
Consumer's Accuracy	0 - 100%	Accuracy of model predictions.	A guide to the user to indicate the probability with which each model prediction is correct.	An important metric for operational use, but not particularly useful to the modeller in identifying model errors.
kappa	-1 to +1	Index of Agreement for positive and negative samples combined.	Adjusts for chance model agreement with training set data (for which kappa = 0). Applicable to multiple categories of presence/absence or abundance.	Sensitive to overall prevalence at high and low prevalence levels.
AUC	0 to +1	AUC is the Area under the Curve of a plot of Sensitivity (y-axis) against (1-Specificity) (x-axis , sometimes called the Receiver Operating Characteristics (ROC) plot. Wilcoxon's algorithm is used to calculate the AUC	Effectively combines sensitivity and specificity to assess model accuracy. Commonly used in logistic regression analyses where probability thresholds to achieve best fit (for presence/absence) are often NOT 0.5. Less affected than kappa by high/low overall prevalence.	Rather more time consuming to calculate than other methods, and more difficult to interpret. Only works for binary (presence/absence) situations.
AIC	0 to ∞	AIC is Akaike's Information Criterion used in information-theoretic models	Estimates the difference between a model's performance and some unknown, ultimate truth. Models with lower AICs are better than those with higher AICs.	AIC is used to compare models on an arbitrary scale. Absolute and relative differences between models are more informative, and can determine which models to drop from a candidate set of 'possible' models.