THE USE OF GEOGRAPHICAL INFORMATION SYSTEMS IN SOCIO-ECONOMIC STUDIES

NRI Socio-economic Series 4

P Daplyn, J Cropley, S Treagust and A Gordon

NATURAL RESOURCES INSTITUTE
Overseas Development Administration

The Natural Resources Institute (NRI) is an internationally recognized centre of expertise on the natural resources of developing countries. It forms an integral part of the British Government's overseas aid programme. Its principal aim is to alleviate poverty and hardship in developing countries by increasing the productivity of their renewable natural resources. NRI's main fields of expertise are resource assessment and farming systems, integrated pest management, food science and crop utilization.

NRI carries out research and surveys; develops pilot-scale plant, machinery and processes; identifies, prepares, manages and executes projects; provides advice and training; and publishes scientific and development material.

No charge is made for single copies of this publication sent to governmental and educational establishments, research institutions and non-profit-making organizations working in countries eligible for British Government Aid. Free copies cannot be addressed to individuals by name but only under their official titles. Please quote **SES4** when ordering.

Natural Resources Institute
ISBN: 0 85954 372-2
ISSN: 0967-0548

Price: £ 5.00

This publication is printed on chlorine-free paper

Other titles in the NRI Socio-economic Series

Needs Assessment for Agricultural Development: Practical Issues in Informal Data Collection (SES1)

Commercialization for Non-timber Forest Products in Amazonia (SES2)

The Allocation of Labour to Perennial Crops: Decision-making by African Smallholders (SES3)

Gender Issues in Integrated Pest Management in African Agriculture (SES5)

Indigenous Agroforestry in Latin America: a Blueprint for Sustainable Agriculture? (SES6)

(Please contact Publishing and Publicity Services at NRI for forthcoming titles in this series)

Contents

Foreword

This series is based upon work carried out under the socio-economic research programme at NRI. Its purpose is to provide an easily accessible medium for current research findings. Whilst it is hoped that the series will be of interest to those concerned with development issues worldwide, it may be of particular relevance to people working in developing countries.

The topics covered by the series are quite diverse, but principally relate to applied and adaptive research activity and findings. Some papers are largely descriptive, others concentrate on analytical issues, or relate to research methodologies.

The aim is to present material in as straightforward a fashion as possible so that it can reach a wide audience. We are interested in the views and opinions of readers and welcome any feedback to this series.

Alan Marter
Socio-economic Research Programme

Summary

Geographical information systems (GIS) have found wide and growing applications, as digital remote-sensing data and computer technology have become more sophisticated, more easily available and less expensive. NRI recently undertook preliminary research into potential socio-economic applications of GIS.

The feasibility of utilizing spatial data, available in GIS, to model socio-economic relationships was examined. It included the following steps:

(a) identification of hypothetical relationships between socio-economic variables and location-specific variables;

(b) investigation of data sets that might permit the socio-economic variables to be modelled as a function of the spatial phenomena; and

(c) a critical assessment of the prospects for modelling socio-economic relationships utilizing GIS data.

A number of general issues concerning the availability of suitable data sets, which can constitute a serious constraint to GIS applications were highlighted in the case studies.

Recommendations are made concerning how data could be made more amenable to this type of appliction, and the criteria that should be applied in assessing the feasibility of projects involving the use of GIS in socio-economic studies.

Introduction

Resource assessment represents an integral part of the process of development in the natural resources sector. Originally based on the collection of data through laborious field-based surveys, resource assessment has broadened in recent years with the addition of two new technologies – satellite-borne remote sensors and geographical information systems (GIS).

Thus in the last decade a revolution has occurred in the ease with which complex data sets describing the physical environment can be collected, stored, modelled and interpreted. The advent of increasingly powerful, inexpensive micro-computers has made the use of such technologies more widely available to users in developing countries.

Much of the data available describe physical characteristics such as climate, topography, soil classification and so on. These data enable the assessment of the resource potential of different regions. Other data, particularly those relating to land use, provide useful indications of the allocation of resources. The linkages between the resource potential of a region and the way in which these resources are used reflect the decisions of people within that location.

This report explores the extent to which data now available through GIS might be used to identify and potentially describe the non-physical, 'human' environment in which people make decisions relating to resource allocation. As well as exploring the potential of currently available data, it seeks to highlight constraints of the explanatory power of such information systems and to recommend measures to overcome these.

The report is aimed at readers who, though they may not be conversant with the more technical aspects of GIS, are interested in utilizing the increasing volumes of information GIS contain.

Objectives and approach of study

The increasing availability of location-specific data, and of the facilities for managing and manipulating them in GIS, suggests an ever-increasing capacity and facility for utilizing such data in modelling aspects of the socio-economic environment. This prospect excites considerable enthusiasm, given the practical and methodological difficulties of collecting data in the field. It also merits a full and structured examination of whether, and to what extent, the striking advantages of this approach to modelling are realizable in practice. This study was conceived to carry out just such an examination.

The first step was to consider the characteristics, capacities and potential of a GIS, and how these relate to the task of modelling socio-economic relationships.

The next step was to identify some of the likely relationships between important socio-economic characteristics in farming systems (such as landholding practices or market orientation) and other, location-specific, variables (such as soil type or topography).

These hypothetical relationships were organized within a matrix, and the most promising of these identified for modelling.

These having been selected, the next stage was to investigate the availability of 'data sets which would allow the socio-economic variables to be modelled as a function of corresponding location-specific data.

Finally, conclusions were drawn regarding the prospects for modelling socio-economic relationships using GIS data, and, if appropriate, intermediate steps identified which could be taken to enhance these prospects.

The basic methodology can be summarized as follows:

- the hypothesis of likely relationships between socio-economic and geographical variables present in GIS format;

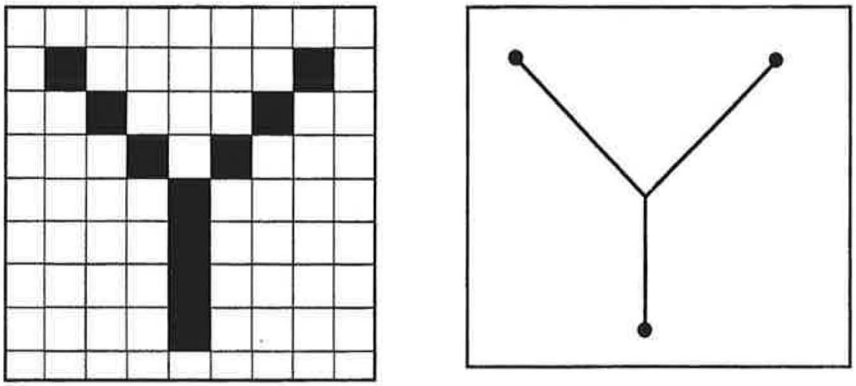
- use of a matrix based on these relationships to select human factors for which GIS data might be available;
- finding of appropriate data;
- identification of prospects for socio-economic modelling using GIS; and production of recommendations to enhance the prospects for modelling.

Geographical information systems

A geographical information system (GIS) is a system of computer hardware, software, and procedures, designed to support the capture, management, analysis and display of data referring to specific locations on earth, and which can be used for solving planning and management problems.

The data can be readily stored, retrieved, manipulated and presented. Different types of information, such as physical characteristics, population, and temperature range, can be held in the same GIS when referring to the same location; the GIS has the capability of analysing and exploring any relationships between them.

There are two distinct types of GIS format: vector



Raster-based system image

Vector-based system image

Figure 1 Data formats in geographical information systems and raster (*see* Figure 1). In a raster-based system, an area is regarded as being made up of many small, equal-sized areas, known as pixels; for each pixel, data are held on its location and attributes. In a vector-based system, the unit is the geographical feature, and information is held on the co-ordinates defining it as a point, line or polygon, together with the nature or attributes of the feature. Thus in a vector-based system, a river, for example, is recorded as such, and its course held as a series of co-ordinates. In a raster-based system, each point along the river is held as a separate pixel, with its location and its categorization as 'river'.

A vector-based system closely reflects a traditional map, and the type of information available on conventional topographic maps can be digitized and incorporated into such a system. The raster format is more suited to information such as that obtained from satellite remote sensing, which is captured as a picture comprising thousands of pixels, each of which contains attribute data.

What kind of data are held, or may potentially be held in a GIS were considered in the study.

RELATIONSHIPS BETWEEN VARIABLES

One of the strengths of a GIS is that it is possible to analyse and explore relationships between variables, providing, however, that the data are available in one database in a single format. This is frequently not the case.

Useful variables include human and physical data of the type recorded on conventional cartographic maps: the location of towns, settlements, coastlines, rivers, water bodies, vegetation boundaries, political or administrative demarcations and altitudes. Remote sensing data may be available for a variety of mainly physical features such as the density of vegetation, surface temperatures, water bodies, and main human features such as towns and transport routes. There may be information originating from surveys or a census, which may potentially be incorporated in a GIS if location data are available.

If data from diverse sources are to be related to each other, several aspects of consistency and compatibility need to be considered. Firstly, information on different topics can change at different rates over time. Some physical characteristics change very slowly, such as topography or soil types. Other characteristics, both physical and socio-economic, change more rapidly, such as the weather, or household consumption patterns. Care should be exercised when combining data originally recorded at different times.

A second major area of potential inconsistency is in the level of precision or resolution of data from different sources, that is, in the fineness of the scale at which they were collected. The general rule when combining data of different levels of resolution is that the resulting resolution is that of the coarsest of the contributory sources. For data held in a GIS, however, the precision of the original data may not be apparent, and it is possible to produce information at a misleading level of apparent precision.

REMOTE SENSING

Remote sensing is the science of deriving information about the earth's land and water area from images acquired at distance, typically from aircraft or satellites. It usually relies on electromagnetic energy being reflected or emitted from these features.

The level of resolution of data is often related to the costs of acquiring it. Some remote-sensing data are freely available to developing countries on a continuous basis from weather satellites such as NOAA. Other, usually high resolution data, may have to be purchased from an organization such as SPOT or LANDSAT.

The greater the precision of the information the higher the cost per image size, and the greater the number of image pictures required to cover an area. For example, in 1987 each LANDSAT image covering an area of 185 km by 185 km was priced at £2 680. For a country like

Sudan, 73 images would be needed to give full cover, at a cost of £365 850. This price does not include the costs for processing the data, or validating the image against conditions in the field, known as ground truthing.

An NOAA image can cover about 3 000 km by 1 500 km with a resolution of one km or more. Only one sheet would therefore be needed to cover Sudan as a whole. The cost of a hardware system to receive the free data transmitted by NOAA was less than £20 000 in 1987 (Walker, 1989).

COUNTRY INFORMATION

Socio-economic data are likely to exist at many different levels of resolution, for example, households or villages, depending on the nature of the survey. Some studies may provide detailed information for a few locations, but in many cases data relating to the livelihoods of rural people are aggregated to the district or regional level.

The availability of data on socio-economic and physical characteristics varies greatly from one developing country to another. Generally, there is reasonable global information available through the United Nations relating to physical features such as land mass borders, altitudes, climate, soils and vegetation types. Data on human features are usually less available, but data sets do exist for political boundaries, major transport routes and urban locations. Data from the UN are often available at no cost, but the accuracy of the information may be variable and sometimes quite dated.

There are often institutional barriers to accessing information. In many countries availability of the best maps and aerial photographs is restricted because the maps are considered to be of military importance (Fox, 1989).

Survey data from developing countries may have been collected but never entered into a computer database. Questionnaire returns are often difficult to interpret by anyone but the original users. Although summary results of a past survey may be available, the original information rarely is.

Where socio-economic and geographical data do exist, they are potentially available for incorporation into a GIS, although the proportion of such data sets already existing in a GIS is small. Information on data available in a GIS format is obtainable from a number of specialized organizations, such as NASA (earth sciences, oceanography, atmospheric and geophysical sciences); the Canadian Centre for Remote Sensing (Landsat, SPOT and MOS); the US Geological Survey (earth sciences, water); and the World Meteorological Organisation (climate) (Walker *et al.*, 1992).

The United Nations Environment Programme (UNEP), based in Nairobi, Kenya, has a more broadly-based remit. Its Global Resource Information Database

SPOT
(Systeme Probatorire d'observation de la terre)
Resolution possible to 10 m. Images of the same location can be taken every 2-4 days.

NOAA
(National Oceanic and Atmospheric Administration)
Resolution to 1000 m. Images of the same location taken approximately twice a day.

LANDSAT
(Land Satellite)
Resolution possible to 20 m. Images of the same location can be taken every 16 days.

METEOSTAT
(Meteorological satellite)
Resolution possible up to 5000 m. Images of the same location taken approximately every 30 minutes.

(GRID), designed to develop global environmental monitoring programmes, is an increasingly important source of information on locationally-referenced global and regional data sets.

DATA COMPATIBILITY

If data on various topics of interest are not available on a GIS, it is often time-consuming and expensive first to obtain the data and then to capture them in an appropriate format. There are also likely to be problems of compatibility between different data sets with, for example, format, resolution and coverage. In the tinted box below, a project in which a GIS was used to model relationships is illustrated.

DEFORESTATION AND MALNUTRITION IN ZAIRE (Fowler and Barnes 1992).

Research is being conducted by Ohio State University on deforestation and malnutrition in Zaire and builds on a project originally funded by the US Agency for International Development (USAID). The steps of the analysis are described briefly below.

Data set identification and capture

a) Selection of the study location on the basis of information available, in particular, relatively cloud-free satellite images.

b) Construction of a hypothetical explanatory model of malnutrition and estimation of data ideally required.

c) Inventory of all available data (many desired data sets not available or incomplete).

d) Collection of additional data on the ground to fill the most important gaps.

e) Integration of existing data to make different administrative units comparable.

f) Geographical location of 162 health clinics plotted on 1:200 000 scale map and digitized for conversion into ARC/INFO format.

g) Conversion of health centre data from being points to polygons, (health centre service areas were estimated to be of 10 km radius), using Thiessen polygons.

h) Handling of data on agricultural extension services and NGOs as 'attribute data' and assigned locations on the basis of the closest health centre.

i) Data from 1984 census assigned spatial areas at 1:50 000 level.

j) Digitalization of graphic layers for various attribute data:- 1:200 000 map from 1970s used for roads, rivers and settlements; 1:50 000 map used for creation of 1955 layer of forest cover.

k) Searching archives of satellite data. Forest cover LANDSAT images obtained for 1974 and 1986.

l) Ground-truthing of satellite data.

m) Image rectification of satellite data. Image classification and raster to vector conversion. Data finally resampled from 80 m squares to a larger grid size of 240 m squares.

GIS model formulation and operation

n) Development of a simple non-spatial model to describe malnutrition as a function of three spatially referenced phenomena: forest resources, agricultural extension and market forces.

o) The model provided the weights for spatial overlays and therefore a framework for manipulating the GIS elements.

p) From a total of 162 health centre polygons, data from only 34 polygons sufficiently complete to fit the requirements of the model.

q) Operation of the GIS model to explain health status. Model was iteratively developed to weigh the explanatory power of variables for different parts of the sub-region.

r) The model was made more sophisticated by inclusion of the interactions of variables over time by using a simulation programme to make the model more sophisticated.

This particular project demonstrates both the potential and the complexities of handling different types of data which relate to a developing country. In this case a great deal of the challenge was in data acquisition as much as in analysis. Even in countries where data are readily available, such as Canada, research scientists using GIS and remote sensing to study wetlands were forced to conclude that their work 'demonstrates the complexity and inter-disciplinary nature of an initially seemingly simple task' (Fabbri, 1992).

Identification of relationships

DEVELOPMENT OF A CONCEPTUAL MATRIX

The major catalyst behind this study has been the apparent relationship between location-specific variables, amenable to use in GIS, and the corresponding use of natural resources. Such socio-economic relationships have frequently been based on empirical, observed interactions; others are wholly theoretical, but nonetheless thought to exist. However, progress has been made towards studying these relationships in a more formal manner in only a few instances.

The first step in bringing together this diverse range of apparent relationships was to canvass suggestions from individuals actively involved in natural resource development. Views were sought from specialists in, for example, arable crops, horticulture, livestock, fisheries and forestry, with a range of geographical experience. The aim was not to compile a comprehensive set of such relationships, but rather to collate a sufficiently wide range to select commonly observed/theorized relationships which could be investigated further.

As would be expected from relationships based largely on observation, those identified by the specialists were relatively specific and included:

- **Nigeria**
  High population and density ———> greater use of manure and compound waste
- **Papua New Guinea**
  Increasing proximity to town ———> higher incidence of horticultural production for sale

In some cases suggested relationships involved intermediary stages:

- **Uganda**
  Poor roads —> lower commodity prices —> lower ex-farm sales

In all, some 50 such relationships were suggested. Although covering a wide range of geographical locations and natural resource sectors, when these were examined and considered in closer detail it was possible to group the relationships into broad categories. This became simpler when intermediary stages were ignored and the relatively specific examples expressed in broader terms.

For example, the relationships observed in Papua New Guinea and Uganda are essentially the same. When summarized as above, the left-hand side of each can be described as *communication links*, and the right-hand side as *market orientation*. Similarly, activities such as use of manure, terracing land, and the construction of windbreaks can collectively described as *conservation-related activities*.

The original relationships could be categorized into broader groupings and therefore moved from those observed or theorized in a comparatively narrow context to those relevant on a wider basis. Although any subsequent work on explaining these in a more formal manner (i.e. modelling) will need to be carried out over closely defined boundaries, such methodologies are likely to have broader applications. For example, data from south-western Uganda might indicate a close correlation between high rainfall levels at certain times of the year with the undertaking of soil conservation activities by farmers, such as the construction of ridges to limit run-off – the higher the rainfall, the greater the activity. In northern Uganda, *low* rainfall levels at certain times of the year appear to result in a greater incidence of soil conservation activities such as the building of windbreaks. Such examples highlight limits on the application of individual models, though not of the modelling methodology itself.

The definition of these broader categories for both location-specific and socio-economic variables also made construction of a matrix of relationships feasible. In all, 9 location specific and eleven socio-economic topics, or groups of variables, were defined as shown below.

CONCEPTUAL MATRIX

	Cropping patterns	Intensity of cultivation	Tree establishment removal	Conservation related activities	Livestock density	Off-farm activities	Market orientation	Marketing patterns	Crop processing activities	Use of purchased inputs	Land-holding practices
Soil type	✓	✓	✓	✓							
Rainfall	✓	✓	✓	✓		✓			✓		
Topography	✓	✓	✓	✓							✓
Temperature	✓								✓		
Communication links							✓	✓	✓	✓	✓
Population density	✓	✓		✓		✓	✓			✓	✓
Surface water (fresh)	✓	✓	✓		✓						
Settlement distribution	✓					✓	✓	✓	✓	✓	✓
Pest/disease distribution	✓		✓		✓					✓	

Note: ✓ = presence of a relationship

Location-specific variables

Soil type
Rainfall
Topography
Temperature
Communication links
Population density
Surface water
Settlement distribution
Pest / disease distribution

Socio-economic variables

Cropping patterns
Intensity of cultivation
Tree establishment / removal
Conservation related activities
Livestock density
Non-farming activities
Market orientation
Marketing patterns
Crop processing activities
Use of purchased inputs
Landholding practices

These variables were subsequently arranged in a matrix. The corresponding cells in the matrix represent observed and theoretical relationships. This arrangement also led to the identification of further relationships not originally suggested and therefore worked as a conceptual tool. The matrix is set out in the box.

The matrix also served the important purpose of clarifying the extent to which socio-economic variables relating to the use of natural resources were typically influenced by more than one location-specific variable. In many cases this was already apparent. Continuing with the above example, it was clear that conservation-related activities, although influenced by rainfall, were also determined by the toposequence of the location. (Again, this relationship is not constant. In southern Uganda, the more undulating the terrain the greater the observed bunding to reduce run-off. In the north of the country the reverse is observed, the greater the undulation, the lower the activity to reduce wind erosion).

Other linkages highlighted by the matrix were less apparent, such as the impact of population density on conservation-related activities.

Thus in most cases, the relationships originally identified are represented by individual cells within the matrix. Seeking to describe such relationships in isolation is of limited use; the relationship between socio-economic and all relevant location-specific variables must be considered in order to produce a meaningful model. In terms of the matrix, these constituent cells are represented by columns beneath each of the socio-economic variables.

Examination of the required parameters in each relationship indicate that data on variables in the 'physical', left-hand side are likely to be more widely available than data for the 'human' right-hand side variables.

Meteorological records indicating rainfall and temperature levels at a typically wide range of specified locations are available in many areas of the developing world. Similarly, maps indicating topography, soil type, areas of surface water, communication links and settlement distribution are also available.

Much of this information has been compiled through traditional, ground-based surveying techniques. In more recent years the development and application of remote sensing techniques has dramatically improved the availability of such data. The use of GIS has facilitated the analysis and presentation of such data.

The same cannot be said of the socio-economic variables. Without exception, these reflect the results of human decisions, and in most cases such decisions are made, and vary, at the household level. It therefore follows that significant measures of variables such as the use of inputs, market orientation and crop processing activities are obtainable only from surveys with the household as the unit of study.

Outline of relationships

By moving from the concept of cells to columns within the matrix, the following relationships were identified.

Conservation-related activities = *f* (rainfall, topography, population density, soil type)

Intensity of cultivation = *f* (rainfall, topography, surface water soil type, population density)

Tree establishment = *f* (rainfall, topography, surface water, pest distribution soil type)

Livestock density (extensive) = *f* (non-crop vegetation intensity, surface water availability, prevalence of livestock pests)

Importance of off-farm activities within household = *f* (seasonality of rainfall, population density settlement distribution)

Market orientation = *f* (communication links, population density, settlement distribution)

Marketing patterns = *f* (communication links, settlement distribution)

Crop processing activities = *f* (rainfall, temperature communication links, settlement distribution)

Use of inputs = *f* (communication links, population density, access to irrigation, settlement distribution, pest distribution)

Cropping patterns = *f* (soil type, topography, rainfall, temperature, surface water, population density, settlement distribution, pest / disease distribution)

Landholding practices = *f* (topography, communication links, settlement distribution, population density)

The types of data required to measure each of the socio-economic and location-specific variables were identified. In many instances, the same variable can be quantified by different types of data. This is particularly the case for the socio-economic variables making up the left-hand side of the relationship. Where this is the case, the various alternatives were specified. (*see* Appendix 1). As indicated above, the variables on the right hand side of each relationship are available, if not over a comprehensive, then at least a significant, proportion of the developing world.

Despite the far narrower coverage achieved by household level surveys, there are no apparent methodological constraints to obtaining data that indicate left-hand side variables, and for many, data sets have already been collected. The practical issue is that of matching socio-economic data sets with GIS data of appropriate and consistent levels of resolution and accuracy and, for some variables, timeliness.

SELECTION OF RELATIONSHIPS FOR FURTHER ATTENTION

Although all the above relationships can be modelled in methodological terms, the impact of factors not embraced by the models makes some relationships less realistic propositions than others.

Clearly, all the above relationships are simplifications of actual situations – this is common to all social-science models. However, empirical evidence indicates that some of the socio-economic factors described are more strongly influenced by exogenous factors than are others. Those most strongly influenced by exogenous factors which cannot be incorporated will produce models with the lowest explanatory power and therefore the lowest utility.

Each of the relationships above was therefore assessed on the basis of the degree of impact of exogenous factors. Those most strongly affected were judged to be livestock density and landholding practices.

Human decisions determining livestock density, although influenced by the location specific variables of 'non-crop vegetation intensity', 'fresh surface water availability' and 'prevalence of pests', are also strongly influenced by cultural factors. Of these, religion is the most important and in many societies can override all of the above physical determinants. This relationship was therefore regarded as being inappropriate for modelling.

The prevailing influence of exogenous factors on landholding practices led to a similar decision. Issues such as government policies on land reform, re-settlement programmes and proximity to international frontiers, together with cultural factors such as kinship structures, frequently outweigh the influences of topography, communication links and settlement distribution.

The impact of exogenous factors, particularly government policies and cultural factors, was also considered to be moderately strong on use of inputs, tree establishment, intensity of cultivation and conservation-related activities. These were also therefore deemed less appropriate for modelling.

A further socio-economic group of variables, that of cropping patterns, was considered to be influenced by such a wide variety of locational and exogenous factors, probably with complex interactions, that the prospects of success in modelling were again felt to be small.

There remained four variables: market orientation, marketing patterns, crop processing, and non-farm activities which were judged to be less affected by exogenous factors, and to be the most promising candidates for modelling relationships.

Market orientation

Market orientation refers to the extent to which output is sold as opposed to being consumed on-farm, and is linked to communications, settlement distribution and population density. The presence or otherwise, of traders and middlemen has also been mooted as an additional determining factor. Whilst this is certainly the case, the argument that this represents an exogenous factor is difficult to sustain: their presence is more likely to be related to the same location-specific variables.

Marketing patterns

Marketing patterns refer to the degree to which farming households are involved in downstream marketing of produce, as opposed to selling at the farm gate. Such patterns are strongly influenced by communication links and settlement distribution. Exogenous factors exist, perhaps the most important being the influence of government policies which may result in compulsory selling of commodities at the farm gate at fixed prices. These in turn remove options for marketing and thus much of the variation in patterns. However, geographic regions over which such factors apply can usually be defined accurately, and exclude in the development and application of models.

Crop processing

With regard to crop processing activities, the influence of cultural and ethnic factors is likely to play a significant role in determining the nature of processing undertaken. However, the general adoption of processing as a strategy is more closely linked to the location-specific variables described (*see* page 7). It was clear that attention would need to be restricted to individual commodities, otherwise the influence of exogenous variables would again be great. Two major commodities in particularly were commonly processed, under a wide variety of methods and circumstances, and were therefore identified for consideration, namely cassava and fish.

These four relationships were therefore selected as being the most promising for further investigation. The importance of off-farm activities within rural households appears to be linked with the population density of the location (for example in densely populated areas where access to land is limited income from non-farming activities may be necesssary to supplement domestic food production). The distancce to an urban centre may also play a determining role in that this may reflect employment opportunities. In some areas, the seasonality of rainfall results in temporary migration from rural households to undertake non-farming activities. Clearly the areas over which such relationships exist may be narrowly defined and the influence of compklicating factors such as proximity to non-urban employment centres (for example sugar mills etc) cannot be discounted.

Nonetheless, the scope for modelling this, and the other four relationships described above were selected for further investigation.

Case studies

Having selected the relationships better suited for further study, an investigation into the existence and availability of data on their constituent variables was made.

An exhaustive search of all potential sources of data was beyond the scope of this study, and was not judged to be appropriate. The aim was to determine the extent to which compatible data sets existed, describing each variable within relationships, rather than to compile a comprehensive list. The objective was therefore to identify a relatively small number of complete data sets and to gauge the extent to which this objective was met.

Organizations and/or individuals with experience in the study of the socio-economic characteristics within each relationship were selected to undertake these case studies, thereby exploiting their comparative advantage.

For each of the selected topics, the following tasks were specified:

- to identify four geographical areas, countries, regions or districts in the developing world, for which compatible data measuring the socio-economic characteristic and the location-specific variables influencing this (in the hypothetical model), exist;
- to provide details of what information is available

for each variable, the unit of study, the date(s) to which this information relates, and the geographical area over which it is available. Details of how this information can be obtained and its format should also be indicated if possible;

- if four locations cannot be identified, then details of the nearest to complete data sets should be provided;
- it is important that the units of study are spread over a sufficiently wide geographical area to register distinct differences in the corresponding, location-specific variables. For example, all households in the same village will share the same communication links and population density, but households in different villages will not. The data sets should therefore comprise a minimum of 10 area units of similar size, far enough apart for differences in location-specific variables to be identified – as a broad indicator, at least 10 miles apart;
- it is important that the geographical area covered by the survey be stated as precisely as is possible, and that the data available are specified for each variable;
- for some variables, the specified measure may not be readily available, but the data required to derive this will be; this should be made explicit in the review.

The resulting case studies are reproduced in Appendix 2. The salient features of data availability are summarized below.

Market orientation and marketing patterns were considered together, since they are usually closely related and are likely to be represented in the same data sources. Three possible case study data sets were identified; two of them were multi-country and linked with each other. All the data sets identified had satisfactory geographical coverage, and included topics from which appropriate measures of both market orientation and marketing pattern could be derived, though for one of the single country studies the marketing structure is highly atypical.

Non-farm activities were represented by some measure of their contribution to the household economy. Data for eight possible case studies were identified, six in Kenya and one each in Guatemala and Mexico. The relevant variable in each of these case studies was the either the proportion of time spent in, or proportion of income derived from, non-farm activities. The number and distribution of locations was satisfactory in four of the studies but in only one of these, the Kenya Rural Labour Force Survey of 1988/89, were the data more recent than 1979. Considering the availability of corresponding location-related data, there exists for Kenya a series of 1:50 000 maps showing settlement patterns, amongst other characteristics. These maps date individually from between 1969 and 1987. There is also a series of 1:250 000 climatic maps, which show the locations of meteorological stations, and rainfall graphs for them. The most recent population census for Kenya was for 1979 - this is not known to be mapped on a detailed level.

For cassava processing, the only case study identified was the multi-country Collaborative Study on Cassava in Africa which initially involved six countries, each with at least 30 locations. The first phase of the study did not include any measure of the importance of cassava processing. From the third continuing phase, it should be possible to ascertain the proportion of those households growing cassava which also process it. In this study the data collected in the first phase included the map reference of the village location, the distance to the nearest city, and the state of the road leading to it. Preparation of the sampling frames included the extraction of the relevant rainfall and temperature data from the Centro Internacional de Agricultura Tropical (CIAT) climate database. Population density data were also prepared for mapping, but came from the 1963 census. The most recent census was in 1992 and detailed population distribution data are unlikely to be available in the near future.

Four case studies were initially identified on fish processing, but two did not contain usable indicators and one of the remaining studies related only to a small geographical area. The fourth study relates to a survey of Nias

Island, Indonesia, conducted under the Bay of Bengal Programme. It covered 11 locations, and was conducted in 1989/90. Employment in fish curing, and fish marketing topics, were covered, but the precise questions are not known. The study recorded the locations of the survey villages, their distances from the nearest administrative headquarters and the type of access. No climatic data are known to be available.

Conclusions

The case studies highlight a number of general issues concerning the availability of suitable data for modelling relationships between socio-economic and locational variables.

In some instances, methodological issues relating to the way data were collected have a bearing on their usefulness. For example, apparent flaws in sampling during the Integrated Rural Surveys in Kenya (1976-79) draw into question the quality, and therefore the utility, of the data on non-farming activities. The frequency and timing of both this and the subsequent Rural Labour Force Survey (1988/9), also in Kenya, means that seasonal variation may not have been 'captured', thus also affecting data quality.

In other cases, limits on the size of the data set, or the geographical area to which it relates, have a bearing on its usefulness. For example, the 1977 study of non-farming activities by Lavrijsen in Western Kenya, although containing data from over 150 households, covers only three geographically separated sites (villages).

Larger scale surveys overcome this constraint. Some were identified in the case studies:

Market orientation and patterns

- Living Standards Measurement Survey (multi-country).
- Social Dimensions Adjustment Priority Survey (multi-country).
- Evaluation Study, Government of Sind (Pakistan).

Crop processing

- Collaborative Study of Cassava in Africa (multi-country)

However, these examples highlight a further potential constraint, that of access to existing data. Data from all the above surveys have typically undergone at least some degree of aggregation prior to publication. Although a given survey may offer the potential to identify, say, market orientation at 10 locations, the data within the public domain may not.

Access to 'raw' data may be required. In some cases, particularly in older surveys conducted before the advent of magnetic storage, only aggregated results may have been kept and raw data discarded. In other cases, data which have been kept may not be stored centrally, making retrieval much harder.

Where raw data do exist in a manageable form, the issue of ownership frequently affects accessibility. Data are often controlled by national statistical or planning authorities, or the relevant international co-ordinating agency, and permission from such bodies must be sought to gain access to the data. This may be time-consuming and difficult.

Thus even where data sets which measure socio-economic parameters exist, the extent to which these are suitable for use in modelling hypothetical relationships may be limited by:

- methodology of collection;
- size and geographical spread of sample; and
- access to disaggregated data.

Turning to location-specific variables (i.e. those represented on the right-hand side of theoretical relationships), work conducted in the selected case studies was only able to provide indicators of availability in general terms. Specific details, particularly those relating to the precision and resolution of information, and the time period to which it relates, would require further investigation.

However, these indicators highlighted issues relating to the compatibility of data on location-specific information with socio-economic parameters. Maps exist for most areas of the developing world, but in many cases survey work for these may have been carried out some time ago. Indicators of, for example, communication links that could in theory be derived from these would not be compatible with more recently collected socio-economic data. Access to maps is also sometimes limited or difficult.

Even where apparently compatible data sets exist, a lack of information on the location to which socio-economic data relates may limit its potential for modelling. In many instances, only a village name is used to reference collected data. This is not surprising, since geographic location has typically been unnecessary for the achievement of the objectives for which the data were collected. Even where this remains the case, the cross-referencing of data collected in the future with locational co-ordinates may greatly enhance its future utility. Technology such as global positioning systems are now accessible at low cost and are easy to use. Simple, hand-held units cost approximately £400.

The case studies also highlight the need to ascertain more about the characteristics of apparently comprehensive data sets. For example, information on rainfall or temperature may appear to be available and complete in some areas. However, such a data set may have been derived using models from a much smaller set of recorded information, drawing its accuracy into question.

This case study work identified a range of data sets which may be suitable for use in theoretical modelling. Upon investigation, most of these proved to have problems of availability, accessibility or definition. The most promising appeared to be the data provided by the Living Standards Measurement Study on market orientation and marketing patterns.

The identification of case-study data sets concentrated on the most promising relationships in terms of data requirements, and utilized the knowledge of specialists working in those subject areas. The limited suitability of the data identified must therefore reflect either a paucity of suitable data, or a lack of information on their existence or their suitability and compatibility.

Whichever of these is the case, further investigation (such as modelling) of data relationships will only be of value if knowledge of, and access to, existing data is improved. Initiatives to collate and document information on data sets have been made, the most notable being the Global Resource Information Database (GRID) of the United Nations Environment Programme. Such initiatives have made some progress but further work is required both to update the database with new data when collected, and to broaden it to include further existing data.

In summary, when considering the feasibility of modelling relationships between socio-economic variables and location-specific variables using GIS, the present problems are:

- lack of compatibility between socio-economic variables and locational variables in terms of resolution and/or timing;
- lack of access to data (especially in a disaggregated form);
- limited geographical coverage of socio-economic variables;
- lack of location references for socio-economic data
- data not available in computerized form;
- lack of information on concepts and definitions for socio-economic data;
- locational data derived from implicit models (for example, observed rainfall data at two points used to derive levels at intermediate points); and
- lack of information on data quality.

Although the general approach of modelling relationships in this way is a promising one, the issues of availability and suitability of data need to be addressed first.

Recommendations

The conclusions drawn above have implications, in terms of recommendations for the future, in two main areas. Firstly, it is clear that in the great majority of cases, adequate data with which to attempt to model relationships between socio-economic and locational variables do not exist in a suitable and accessible form. For the existing data to be usable much effort will be needed to access, computerize and document them, and to assess their quality and their compatibility with other data, and to make them accessible to potential users. Such efforts would, however, realize the large potential or latent value of much existing information and it seems likely that the value added would be considerably greater than the cost incurred.

The steps required to achieve this involve:

(a) obtaining and recording map references of survey locations for socio-economic surveys;
(b) producing standard information on survey size, design, concepts and definitions, and depositing this at a central archive;
(c) computerizing existing location-related data, including topography, climate, soils, settlements and infrastructure, population distribution and characteristics; and
(d) depositing such computerized location-related data in a central archive, such as GRID.

The second area of recommendations concerns the appropriate criteria for use in assessing the feasibility of projects involving the use of GIS in socio-economic studies. It is clear from the above that there are many serious potential constraints, and that it is vital to check any proposed project to assess whether any of these apply, and if so whether adequate resources exist to overcome them.

It is therefore recommended that, for any project involving the use of GIS in socio-economic studies, the following questions be asked:

- are data readily obtainable for all the socio-economic and locational variables required?
- are all these data already available in a GIS, or at least in an accessible computerised form?
- are the data for all the variables compatible, in terms of resolution, time reference, format, concepts and definitions?
- is there adequate documentation available to assess compatibility and data quality?
- if the answer to any of the above questions is negative, what resources are needed to obtain and make accessible the necessary information?
- are these resources available?

The application of these questions will enable projects which are not feasible in the current state of data availability under existing resource constraints to be filtered out at an early stage, so that resources can be used in the most effective manner.

References

FABBRI, A. (1992) Remote sensing, Geographic Information Systems and the environment: a review of interdisciplinary issues. *ITC Journal*, **2** : 119-126.

FOWLER, C. and BARNES, G. (1992) Modelling the relationship between deforestation and malnutrition in Zaire using a Geographical Information System (GIS) approach. *ASPRS/ACSM/RT 92 Technical Papers*, 3. GIS and Cartography, 24-33.

FOX, J. (1989) Spatial information for resource management in Asia: a review of international issues. *International Journal of Geographic Information Systems*, **5** (1): 59-72.

TOWNSLEY, P. (1990) A socio-economic appraisal of selected coastal fisherfolk communities in Nias Island, Indonesia. In: *The Fisheries and Fisherfolk of Nias Island, Indonesia.* Madras: Bay of Bengal Programme.

WALKER, D. R. F., NEWMAN, I. A., MEDYCKJ-SCOTT, D. J. and RUGGLES, C. L. N. (1992) A system for identifying datasets for GIS users. *International Journal of Geographical Information Systems*, **6** (6).

WALKER, P. (1989) *Famine early-warning systems: victims and destitution.* London: Earthscan.

Appendix 1

DATA REQUIRED TO MEASURE VARIABLES IN HYPOTHETICAL RELATIONSHIPS

LIVESTOCK DENSITY

Relationship

Livestock density = *f* (non-crop vegetation intensity, fresh surface-water availability, prevalence of livestock pests)

Data

Vegetation intensity

Proportion of area under non-crop vegetation cover, (weighted by fodder value?)

Livestock density

Standard animal equivalents (how defined?), per unit area.

Water

Proportion of area under fresh water, at maximum and minimum.

Resolution/precision

Feasible at any level, though the finer the better.

Exogenous

Cultural factors.

NON-FARM ACTIVITIES

Relationship

Importance of off-farm activities within household = *f* (seasonality of rainfall, population density distance to urban centre + size of centre (or other formal sector centre)

Data

Importance of off-farm activities (OFA)

Proportion of farm households with OFA,

or proportion of people with OFA,

or proportion of income attributable to OFA.

Seasonality of rainfall

Number of months with no rain,

or rain too little to allow crop production (define).

Population density

Persons per unit area (excluding urban settlements).

Settlement pattern

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Resolution/precision

Small enough unit area that urban centres not in same unit as large numbers of rural people - census enumeration area a possibility.

Rainfall seasonality fairly consistent over reasonable distances.

USE OF PURCHASED INPUTS

Relationship

Use of inputs =*f* (communication links, population density, settlement distribution, pest distribution, access to irrigation)

Data

Input use

Absolute quantities per farm household of fertiliser, purchased seed, herbicides, pesticides - weighted (?) and summed,

or ditto by cultivated area,

or expenditure on the above (absolute),

or ditto relative to area,

or score 1 for use of each category and sum.

Communication links

Distance units of road / rail (weighted by category) per unit area.

Settlement pattern

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Population density
Persons per unit area.

Pest distribution
Score of sum of prevalent crop pests.

Access to irrigation
Whether farmer or settlement is able to access existing irrigation system.

Resolution/precision
Any feasible, but the finer the better.

Exogenous factors
Cropping patterns, government policies, world economy.

CONSERVATION-RELATED ACTIVITIES

Relationship
Conservation-related activities = *f* (rainfall, topography, population density)

Data
Note: need to model each category (broad) of CRA separately?
e.g. *soil run-off control* (bunds) probably reflects incidence of periods of high rainfall (and steep slopes); control of *wind soil erosion* (windbreaks, etc.) probably reflects periods of low / nil rainfall (and few slopes), (and winds)
Difficult to model these activities together?

Conservation-related activities
Proportion of farm households undertaking conservation-related activity,

or proportion of labour inputs attributable to conservation-related activity.

Rainfall
Number of months during which rainfall exceeds (or falls below) X mm (to be defined).

Topography
Proportion of land on slopes X° or steeper.

Population density
Persons per unit area (excluding urban settlements).

Resolution/precision
Feasible at any level, but the finer the better.

Exogenous factors
Cultural factors. Dynamics (CRA more likely to be undertaken as survival strategy by populations exposed to above conditions for some time rather than those who have only started to face them in recent years).

CROP PROCESSING

Relationship
Crop processing activities = *f* (rainfall, temperature communication links, settlement distribution)

Data
Crop processing activities
Proportion of households undertaking crop processing activities,
or proportion of total crop production undergoing processing,
or proportion of labour inputs attributable to crop processing.

Rainfall
Number of weeks during which rainfall / relative humidity levels enable crop drying (how defined?)

Temperature
Number of weeks during which temperature levels enable crop drying.

Communication links
Distance units of road / rail (weighted by category) per unit area.

Settlement pattern

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Resolution/precision
Feasible at any level, though the finer the better.

Exogenous factors
Specific types of food processing undertaken influenced by cultural factors, though *general adoption* of processing as a strategy more closely linked to the above.

Cropping patterns will determine balance of durable and perishable crops and thus the need, though not necessarily the incentive, for processing.

MARKET ORIENTATION

Relationship
Market orientation = *f* (communication links, population density, settlement distribution, [surface water])

Data

Market orientation
Proportion of (household) income from off-farm sales (though proportion could be high even when market orientation is low).
Proportion of output sold off-farm (weighted by commodity?) but:
this measures reactive orientation?
i.e. high 'score' could reflect marketing of unexpectedly high output of otherwise 'subsistence' crop.
A measure of pro-active orientation might be:
proportion of land devoted to cash crop production,
or (better) proportion of labour devoted to cash crop production (and processing).

Communication links
Distance units of road / rail (weighted by category) per unit area.

Population density
Persons per unit area

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Surface water
(a) Sea: $\frac{1}{\text{Distance to coast}}$

(b) Fresh water: $\frac{1}{\text{Distance to shore}} \times \text{surface area of water}$

Assuming surface area of water provides approximation of (fisheries) resources within.

Resolution/precision
Feasible at any level, but the finer the better, as aggregation will mask variation in population density and communication links.

Exogenous factors
Presence of traders / middle-men, though this is also likely to be a function of the above factors.

Government policy (involving marketing constraints).

MARKETING PATTERNS

Relationship
Marketing patterns = *f* (communication links, settlement distribution)

Data

Marketing patterns
For example, proportion of marketed output sold at farm gate.

Communication links
Distance units of road / rail (weighted by category) per unit area.

Settlement distribution

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Resolution/precision
Feasible at any level, but the finer the better, as aggregation will mask variation in a key variable (communication links).

Exogenous factors
Government policy, e.g. marketing boards operating compulsory buying fixed price at farm gate thus removing options for marketing.

LANDHOLDING PRACTICES

Relationship
Landholding practices = *f* (topography, communication links, settlement distribution)

Data

Landholding practices
Proportion of land acquired by means other than inheritance during last X years. 'Acquire' is taken to mean 'have control over', though not necessarily own.

Other suggestions may be appropriate, but situation complicated by complexity of tenurial systems.

Possibly better to look for changes in practices associated with changes in location related variables (i.e. communication links and settlement distribution).

Topography
Perhaps only location-related in specific cases (e.g. Guatemala). It might be better to treat it as an exogenous factor in the general model. If not then:
proportion of land on slopes X° or steeper,
and / or proportion of land above X metres elevation.

Communication links
Distance units of road / rail (weighted by category) per unit area.

Settlement distribution

$$\frac{1}{\text{Distance to nearest urban centre}} \times \text{size of centre}$$

Possibly summed over centres if more than one within specified distance.

Resolution/precision
Feasible at any level, but the finer the better, as aggregation will mask variation in topography and communication links.

Exogenous factors
Many. Including government policy (re-settlement, land reform), cultural factors, proximity to international frontiers, etc.

INTENSITY OF CULTIVATION

Relationship
Intensity of cultivation = *f* (rainfall, topography, surface water)

Data

Intensity of cultivation
Proportion of growing seasons in which cultivated land is fallowed (excluding land devoted to perennial crops).

Rainfall (seasonality)
Number of months (or longer periods?) with no rain (or too little to allow crop production) to be defined.

Topography
Proportion of land on slopes X° or steeper,
and / or proportion of land above X metres elevation.

Surface water

$$\frac{1}{\text{Distance to nearest permanent surface water}}$$

Resolution/precision
Feasible at any level, but the finer the better, as aggregation will mask variation in topography. Measure of surface water only relevant over short distances?

Exogenous factors
Land availability. Demand for output. Government policies.

TREE ESTABLISHMENT/REMOVAL

Relationship
Tree establishment = *f* (rainfall, topography, surface water, pest distribution)

Data

Tree establishment
Trees established per unit of land during last X years.

Model separately for removal or 'net out', i.e. trees established - trees removed.

Rainfall (seasonality)

Number of months with no rain, or too little to allow crop production to be defined.

Topography

Proportion of land on slopes X° or steeper, and / or proportion of land above X metres elevation.

Surface water

$$\frac{1}{\text{Distance to nearest permanent surface water}}$$

Pest distribution

Score 1 for each tree pest present and sum.

Resolution/precision

Feasible at any level, but the finer the better.

Exogenous factors

Cropping patterns may influence tree establishment strategies, e.g. establishing shade for coffee / cocoa.

Livestock density may influence numbers of trees successfully established, as opposed to planted.

Appendix 2

CASE STUDIES ON DATA AVAILABILITY FOR SELECTED RELATIONSHIPS

RURAL NON-FARMING ACTIVITIES

Kenya, Rural Labour Force Survey, 1988/89

Locational spread of survey, and number of locations

Six provinces in Kenya: Coast, Eastern, Central, Rift Valley, Nyanza, Western. No information on number of locations, but large sample size and several locations in each province.

Measure used to indicate its importance

Average weekly hours spent on eight economic activities, including 'labour in non-farm own profit-making activities'. Derived measure: proportion of working population by hours worked in eight categories.

Unit of study

Individual rural residents (working age population).

Date to which information relates

Two phases to capture seasonality, 1988 / 89.

How can the information be obtained?

Presumably the full data set is available from the Long Range Planning Division and Central Bureau of Statistics, Ministry of Planning and National Development, Kenya. April 1991 report at University of London, Institute of Commonwealth Studies.

What form is it in?

Methodological issues which have a bearing on the usefulness of the data

Seasonality: phase 1 – data on hours in last week (and the day before); phase 2 – data on hours the day before. In fact seasonality may be hidden because interviews were spread over 2-3 months for each phase.

Western Kenya, Lavrijsen

Locational spread of survey and number of locations

One village in Bungoma District and two in Kakamega, covering three ecological zones.

Measure used to indicate its importance

Percentage of income derived from farm and off-farm income. Absolute income data also available.

Unit of study

Household; n = 155 (1977 phase, when off-farm income data collected, village split - 50, 50, 55).

Date to which information relates

Baseline studies 1974/75; off-farm income data Feb.-Oct. 1977; qualitative phase 1978-80.

How can the information be obtained?

Perhaps from the researcher or from the Geographical Institute of Utrecht State University

What form is it in?

Probably suitable for rudimentary computer systems—but should be analysed by village.

Methodological issues which have a bearing on the usefulness of the data

Nothing immediately apparent.

The Integrated Rural Surveys, Kenya, 1976-79

Locational spread of survey and number of locations

Six provinces in Kenya; sample frame and questions modified over time (survey 'run' four times); several locations in each province covered. Eighty-six (1977) and 92 (1978) primary sampling units (geographically distinct areas in different provinces, further divided into 'chunks' and 'clusters').

Measure used to indicate its importance

Whether or not involved in various, listed, types of non-farm activity (i.e., apparently not by percentage of income). Also hours spent in different activity (labour force survey).

Unit of study

Household (1977 n=3020; 1978 n=3563).

Date to which information relates

Rural non-farm activities introduced in IRS3 (March 1977). IRS4 (March 1978) as IRS3 though geographical coverage changed a little. Also IRS3 respondents participated in a labour force survey, indicating hours of off-farm employment, for all household members (day or week prior to the interview). Non-farm activity was the subject of a special 'module' of IRS4.

How can the information be obtained?

Presumably from the Central Bureau of Statistics, Ministry of Economic Planning and Development, Kenya.

What form is it in?

Methodological issues which have a bearing on the usefulness of the data

Problem with 'blurred' definition of employment: 'employed', 'employed but not at work' (the day before the interview), 'unemployed' (if no work and looking for work); those not falling into any category were considered to be outside the labour force. Data reported (and recorded?) in unweighted annual averages, therefore not picking up any seasonal variation in non-farm activities. Cluster boundary problems during enumeration. Spurious population estimates led to over- or under-representation of some of the clusters.

Kenya, Hebinck

Locational spread of survey and number of locations

Farming community of Nandi, different agro-ecological zones, six 'sub-locations' (but no indication of proximity).

Measure used to indicate its importance.

Percentage of income 'off-farm'

Unit of study

Household, n=90.

Date to which information relates

August-October 1987.

How can the information be obtained?

From the researcher at Nijmegen.

What form is it in?
Recent, therefore should be computerized at least.

Methodological issues which have a bearing on the usefulness of the data
Some participant selection problems.

Kenya, Kisii, Orvis, 1989

Locational spread of survey and number of locations
Most data from one community, but some comparative data over three different administrative divisions in Kisii.

Measure used to indicate its importance
Income derived from different sources.

Unit of study
'House' - defined as a subset of household. n = 50 for case study in one place; comparative study n = 305.

Date to which information relates
Probably mid-80s; written up in 1989.

How can the information be obtained?
Perhaps from the researcher at University of Wisonsin, Madison.

What form is it in?
PhD thesis and presumably a computerized database.

Methodological issues which have a bearing on the usefulness of the data
Some problem with the definition of household.

Kenya, Paterson

Not discussed because covers only one village, though could perhaps combine with information from other surveys. (Litala village, Esianda, East Bunyore). 188 households surveyed in 1978-79, giving information on whether or not engaged in certain (listed) activities. PhD dissertation, University of Washington.

Guatemala, Smith

Locational spread of survey and number of locations
143 communities of Western Guatemala, across four agro-ecological zones.

Measure used to indicate its importance
Household income derived from different activities. Time allocated to different activities.

Unit of study
Household, n=2 544.

Date to which information applies
1984 occupational survey; other data for 1976-78.

How can the information be obtained?
From researcher at University of North Carolina, or International Labour Organization.

What form is it in?
Computerized, but no indication of form. Analysis broken down by agro-ecological zone.

Methodological issues which have a bearing on the usefulness of the data
None immediately evident.

Mexico, Cook

Locational spread of survey and number of locations
Twenty-one villages and three towns in the Oaxaca Valley.

Measure used to indicate its importance
Involvement of household and individuals in the household in certain named activities.

Unit of study
Household, n=1005.

Date to which information relates
1978-80.

How can the information be obtained?
Perhaps from the researcher at the University of Connecticut.

What form is it in?
No indication given.

Methodological issues which have a bearing on the usefulness of the data
Interesting discussion in the journal article supplied of the influence of local natural resources on non-farm activities. For example, more irrigation - less craft; local supply of raw materials used in craft industries; some relationship between on-farm and off-farm income.

FOOD PROCESSING

Collaborative Study of Cassava in Africa (COSCA)

Locational spread of survey, and number of locations

Initially the study involved Ghana, Côte d'Ivoire, Nigeria, Tanzania, Uganda and Zaire. Other countries have subsequently started similar programmes, but these are at an earlier stage - the information below relates to the initial group of countries. Locations (villages) are spread randomly across the cassava-growing area of the whole country except for Zaire where the survey is limited to a restricted area. For most of the countries there are 35 - 40 locations, and for all of them at least 30, with a double size sample for Nigeria.

Unit of study, and measure used to indicate importance of cassava processing

Phase I of COSCA consisted of a group interview in each sample village. The group was asked what processed cassava products were produced in the village. There was no information gathered on what proportion of households processed, or on what degree of importance this had in the individual households or the village as a whole.

In Phase II of COSCA individual households were surveyed, randomly selected from the Phase I villages. Phase II concentrated on production, and the sample frame consisted of households growing cassava. Phase III, which included a processing component, used the same frame. In Phase III, the sample households were asked whether they processed cassava. Thus Phase III can be used to derive an indicator of importance of cassava processing defined as the proportion of households growing cassava who process it.

Information on locational variables

The initial sampling of villages was carried out by defining grid squares of size 12 minutes of latitude and longitude, stratifying these according to agro-climatic zone, accessibility and population density, and then sampling grid squares, and selecting a sample village from each sample grid square. The categories of the stratification variables are shown below:

Agro-climatic zone	**Accessibility**
Lowland humid	Poor
Lowland semi-hot	Good
Lowland continental	
Lowland semi-arid	**Population density**
Highland humid	Low
Highland continental	High

More details of these categories are contained in a project working paper, which should be reasonably accessible.

For each sample village, the category of these stratification variables is recorded. Also recorded are the map reference of the village location, the distance to the nearest city, and the state of the road to the city.

Date to which information relates

Phase I was conducted in 1989. Phase III was conducted in 1992 in some countries, but may be continuing in 1993 in others.

How can the information be obtained, and what form is it in?

The Phase I data sets are held in dBase files at COSCA headquarters at the International Institute for Tropical Agriculture in Ibadan, and also at various collaborating institutions including NRI. The Phase III data are unlikely to be processed as yet, but it is likely that they will be held in the same format and locations.

Methodological issues having a bearing on the usefulness of the data

The indicator of the importance of cassava processing is a fairly crude one.

The questionnaires used were very long, and this may have affected data quality. The variables of interest for this study, however, are fairly simple and well-defined, and hence less likely to be affected. It is thought likely that the best data may be those for Ghana and Côte d'Ivoire.

FISH PROCESSING

Bay of Bengal Programme, Nias Island, Indonesia

Locational spread of survey, and number of locations

Study based on Nias island, which covers an area of 4800 km^2. Eleven sites selected for study, distributed all round the coast - a map showing locations is available.

Unit of study, and measure used to indicate importance of fish processing

Eleven villages selected as unit of study from a sample of approximately 70. Fish processing indicated through measures of employment by type, including fish curing and ice making.

Information on locational variables

Not specified.

Date to which information relates

1989/90.

How can the information be obtained, and what form is it in?

Key finding of the survey have been published (Townsley 1990). Disaggregated data available at project office, form unknown.

Methodological issues having a bearing on the usefulness of the data

Information gathered from a single, key informant. Extent to which proxy indicator of employment sufficiently quantifies relative importance of processing unclear.

MARKET ORIENTATION AND MARKETING PATTERNS

Social Dimensions of Adjustment Priority Survey

Survey objectives

The Social Dimensions of Adjustment (SDA) Priority Survey was designed to be used for the identification of target groups for the implementation of socio-economic policies, and the production of socio-economic indicators describing the well-being of different groups of households.

The survey is household-based with the household head as the principal respondent. Most questions are based on the household as an economic unit, with a small number addressed to individual household members. The data are collected on a single visit and the interview lasts approximately one hour. Coverage has been national in most countries, but results may not be available for all areas.

Countries

The survey, or an associated survey, with some common types of question and/or sample design has been carried out in 12 African countries. These include:

Country	*Current status*
Chad, 1990	Report available for Ndjamena area
Guinea Conakry, 1990	Report available
Guinea Bissau, 1990	Report available
Senegal, 1990	Report available
Gambia, 1991	Statistical abstract in production
Kenya, 1991/2	Data processing in progress
Zambia, 1991	Report available

Associated surveys

Rwanda, income and expenditure, 1991
Malawi, income and expenditure, 1990
Zimbabwe, household budget survey, 1990/91

Sample design

The sample design is a two-stage probability sample. First-stage units are census enumeration areas, with large areas split and small areas combined. Stratification is used at first stage units to separate country-specified socio-economic groups. Selection at first stage unit is by systematic sampling using probability proportional to size, with sampling rates held constant across strata as far as possible. Household listing is done in each sampled cluster, with typically 20 households selected by simple random or linear systematic sample. Sample size is typically in the region of 8000 households.

Questions

Market and income related questions include the following:

Crop production and marketing

(Named) crop production last season
Quantity of (named) crop sold
Outlet for sales: classified into roadside stall, village market, large market, trader, co-operative, marketing board, other
Unit price for (named) crop sale

Household income

Sale of export crops (named), sale of food crops (named), livestock and livestock products, fishing, other farming income, non-farm enterprises, salaries, rent, remittances, transfer payments, and other sources

Location-specific variables

No location-specific information was collected on the survey questionnaire. To the extent that records exist for the enumeration areas where the survey was conducted, it would be possible to obtain that information by additional fieldwork.

Ownership and access

The data sets from the SDA surveys have been stored on computer and processed using a combination of IBM-PC-compatible software,typically dBase IV, PC-Focus, SAS-PC and SPSS. Hierarchical databases were considered for data management, but it is thought that most data are stored in flat files. Details are country-specific.

Access to the data is at the discretion of the responsible national agency. No examples are known of instances where the data have been made available.

Living Standards Measurement Study

Study objectives

The Living Standards Measurement Study (LSMS) was set up by the World Bank in order to help countries improve data collection efforts to understand poverty and the determinants of living standards. The household survey methodology was first implemented in 1985. Enumeration is spread over one year, although individual households are visited just twice, at a two-week interval. In addition to a household questionnaire, additional data are collected about prices and local conditions at a community level, including transport, communications and other infrastructure data.

The LSMS was modified by the World Bank's SDA unit and has been implemented as the SDA Integrated Survey in some countries. Methodology and content are comparable.

Countries

The survey or an associated survey with some common features of question specification and/or sample design has been carried out in nine countries. They are listed below with the years of data collection.

Country	*Years*
Côte d'Ivoire	1985, 1986, 1987, 1988
Peru	1985-86
Ghana	1987-88, 1988-89
Mauretania	1987-88
Bolivia	1988, 1989
Jamaica	1988, 1989
Laos	1990
Morocco	1990
Pakistan	1990

Sample design

Not known in detail, but understood to be national multi-stage cluster samples of communities and households.

Questions

Not known in detail, but due to the close linkage with the Priority Survey, it is assumed that similar data, plus the community level data would be available.

Location-specific variables

Potentially available from the community survey.

Ownership and access

The data are controlled by the national statistical agency in each country.

Pakistan Left Bank Outfall Drain (LBOD) Project - Socio-economic Impact Evaluation Study

Study objectives

The LBOD impact study is a longitudinal investigation taking up to 10 years, and designed to identify the socio-economic impact of an irrigation, drainage and on-farm water management project in Sind Province of Pakistan.

The study consists of a number of surveys based around irrigation watercourses. One survey deals with the watercourse as an operational entity; a linked survey is based on individual farmers, stratified by type of land tenure, at the watercourse; a third investigation is a one-off sociological survey at a subsample of watercourses.

The study has been conducted twice each year (during the Rabi and Kharif seasons) since 1988, and is planned to continue until 1996.

A potential disadvantage is that the agrarian economy of the study area is essentially feudal. The labour and financial relations among landowners, tenants, sharecroppers and labourers is complex and determined by the complicated agreements for sharing risks and returns to farming. It is not clear to what extent these relationships would distort ordinary market activities, but such distortion is possible.

Geographical location

The study is taking place in three districts of Sind: Nawabshah, Sanghar and Mirpurkhas. A group of six control watercourses are located close to, but outside these districts.

Sample design

Watercourse survey: the watercourse survey is a multi-stage stratified random sample. The sample was stratified by project component area, in proportion to cultivable command area (CCA). One district was further stratified by drainage method. Within each district a systematic sample was drawn of irrigation channels, with probability proportional to CCA (PPC). Within the selected channels, a further systematic PPC selection was made of watercourses. A total of 50 watercourses were sampled.

Farmer survey: within each watercourse, a group of farmers were purposively selected to represent sharecroppers, tenants and landowners.

Sociological survey: two phases of sociological investigation have been carried out. In 1988, nine villages located at sampled watercourses in the study area and three from the control sample were purposively selected. The villages were divided equally among the three component areas. At each village, six male heads of household and six women were interviewed in depth. In 1990, in a second-phase study, virtually all the households in six of the twelve villages (two from each component area, none from the control) were investigated. Total households numbered 245.

Questions

The watercourse survey is concerned with physical and aggregate measures, not household data.

The farmer survey records the area, total output and selling price for named crops, and the amounts sold, consumed or disposed in other ways. Particular emphasis is given to crop division between tenant and landlord and use for payment in kind. Other sections record the on-crop on-farm income and income from off-farm sources. Expenses are covered in more detail.

The sociological study collected detailed information about incomes and expenditure, but the summary reports do not distinguish between the value of farm output and the value of output sold. Further investigation would be needed in Sind to explore the original data sets.

Location-specific variables

The study itself has not been concerned with information about communications, transport and other aspects of settlement patterns and market orientation. However, the study area is mapped, and full locational details could readily be obtained from the resident survey team.

The sampled watercourses are spread more widely than 10 km. It is not known how far apart the sociological subsamples are located.

Ownership and access

The study is being undertaken by the Sind Development Studies Centre, on contract to the Planning and Development Department, Government of Sind. SDSC receives technical assistence for the study from Information Technology and Agricultural Development Ltd in collaboration with Wye College. The data are in dBase format.

Printed and bound in the United Kingdom by SPR International
The Gresham Press, Old Woking, Surrey, GU22 9LH