CHAPTER 6
DATA SOURCES AND SAMPLE CHARACTERISTICS

Formal estimation of the models developed in Chapter 5 requires data on improved banana production management technology adoption, farm characteristics, a set of household characteristics, farm and market characteristics, social capital and other village-level factors. This chapter describes the sources of the data used in the empirical estimation, sampling procedures and methods of collecting data. A few basic characteristics of the sample are also summarized.

6.1. Data sources

The data used in the analysis come from two household surveys conducted in Uganda between 2003 and 2004. First, a survey was conducted among 800 households in 40 villages across the major banana-growing areas of the Lake Victoria region of Uganda and Tanzania[1]. Of these, 547 households are located in 27 villages in Uganda. This survey collected detailed information on household characteristics, market factors, landholdings and farm physical characteristics.

A second survey was conducted during the same period among 400 households to gather additional data required for the present study. Three hundred of the surveyed households were among the households selected for the first survey and 100 households were randomly selected from a domain defined by a project entitled "Reviving Bananas in the Central Region," implemented by NARO, the National Banana Research Programme. The 100 households had been surveyed in 2001 to collect baseline information for the same project.

The second survey collected data on social capital, the adoption of improved banana management practices and village-level social characteristics. The survey could not cover the entire household sample selected for the first survey because of budget constraints. Since the second survey was implemented as a sub-sample of the first

[1] The survey was conducted by the National Banana Research Program jointly with the Maruku Research Institution and in collaboration with the International Food Research Policy Institute (IFPRI) and the International Network for Banana and Plantains (INIBAP), as a baseline for the assessment of the socio-economic impact of banana improvement in East Africa. The study was also implemented in Tanzania.

survey, the next section will present a summary of the sampling frame used in the first survey. Following this, the sampling methods used to select the sub-sample and methods of data collection are presented.

6.2. Sampling frame

The domain was purposively selected to cover areas specializing in banana production, including those with declining, increasing and intermediate current levels of production. These correspond roughly to the Eastern, Central and south-western geographical zones in Uganda. The domain was stratified based on elevation (a physical environmental characteristic) and exposure to recently introduced new banana varieties (an institutional characteristic). Elevation is correlated with numerous factors that affect the incidence and severity of most pests and diseases affecting bananas in the Lake Victoria region (Speijer et al., 1994). Elevation is also related to soil quality, climate and the surrounding vegetation in these environments (Tushemereirwe et al., 2001). Two strata were delineated, defining low elevation as being below and high elevation as being above 1,200 masl (metres above sea level).

The domain was also stratified according to the previous exposure[2] to new banana varieties (exposed/not exposed) so as to compare the "factual" (i.e. the actual case) with the "counterfactual" (i.e. the situation in a comparable case where no adoption had occurred). Although the adoption of new banana varieties is not the major focus of the present study, exposure as a stratification variable is relevant to the study since improved banana management practices might have been disseminated along with the new banana varieties.

Geo-referenced data on banana production systems, a digital elevation model, maps of administrative units and information concerning the previous diffusion of banana planting material were used to disaggregate the domain into a total of four strata: (1) low elevation, with exposure; (2) low elevation, without exposure; (3) high elevation, with exposure, and (4) high elevation, without exposure. The domain and four strata

[2] Areas of "exposure" were defined as LC3s where researchers, extension services or other programmes had introduced improved plant material (in the form of banana suckers) into at least one community. Areas with no exposure are those where no organized programme designed to diffuse improved planting material had been conducted, according to personal consultation with NARO.

were then mapped onto the administrative level of a Ward in Tanzania and an LC3 (local council level 3, or sub-county) in Uganda. Wards and LC3s were designated as high or low-elevation based on a simple majority proportion of the unit being above or below 1200 metres above sea level.

An efficient allocation of sample to strata in formal sampling schemes is one that minimizes variation within the stratum and maximizes variation between the strata, in turn minimizing the overall sampling error (Hansen et al., 1953). When the variances in population parameters are known, the sample can be allocated optimally within and among primary sampling units (PSUs) by choosing their number and the number of households per PSU to minimize the survey (sampling and non-sampling) error, given a fixed budget (De Groote, 1996). In this case, the variances in the multiple population parameters of research interest were unknown.

The minimum sample size for conducting hypothesis tests on variables measured at the community level (such as social capital and some physical capital) is 20 each in exposed and non-exposed areas (corresponding to a student's *t*-test). Although a larger sample of communities would have been preferred for the sake of statistical precision, the cost of conducting the research in more than 40 communities scattered across the domain exceeded the budget. The total number of PSUs was therefore fixed at 40, with half distributed through exposed areas and the other half through non-exposed areas. The 40 primary sampling units were then allocated between the two elevation levels and the two countries proportionate to the probability of selection. PSUs were drawn using systematic random sampling from a list frame with a random start. The sampling fractions for the primary sampling units among the four strata in the domain are shown in Table 4.

Table 4. Sampling fractions for primary sampling units (PSUs) in the survey domain

Population of PSUs				Sample of PSUs			
	Elevation				Elevation		
	Low	High	Total		Low	High	Total
Exposed	49	5	54	Exposed	18	2	20
Row pct	(91%)	(9%)	(100%)	s.f.	0.367	0.40	0.37
Non-exposed	155	49	204	Non-Exposed	15	5	20
Row pct	(76%)	(24%)	(100%)	s.f	0.097	0.10	0.098
Total	204	54	258	Total	32	8	40

Source: Smale et al. (Forthcoming)

S.f=PSU sampling fraction (n*ij*/N*ij*), where *i*=elevation (1,2) and *j*=exposure (1,0)

PSU sampling fractions (*s.f.*) vary by stratum, and are defined as the ratio of stratum-specific sample size (n_{ij}) and stratum-specific population size (N_{ij}), expressed as (n_{ij}/N_{ij}). The final sample for Uganda consists of 27 PSUs, of which 18 are located in non-exposed and 9 in exposed areas (Table 5).

Table 5. Survey sites (primary sampling units) as represented by elevation and diffusion status in Uganda

Low Elevation areas		High elevation areas		Total
Non-exposed	Exposed	Non-exposed	Exposed	
14	8	4	1	27

The spatial representation of the primary sampling units is shown in Figure 5 with sampled sites highlighted. Figure 6 shows only the sites surveyed.

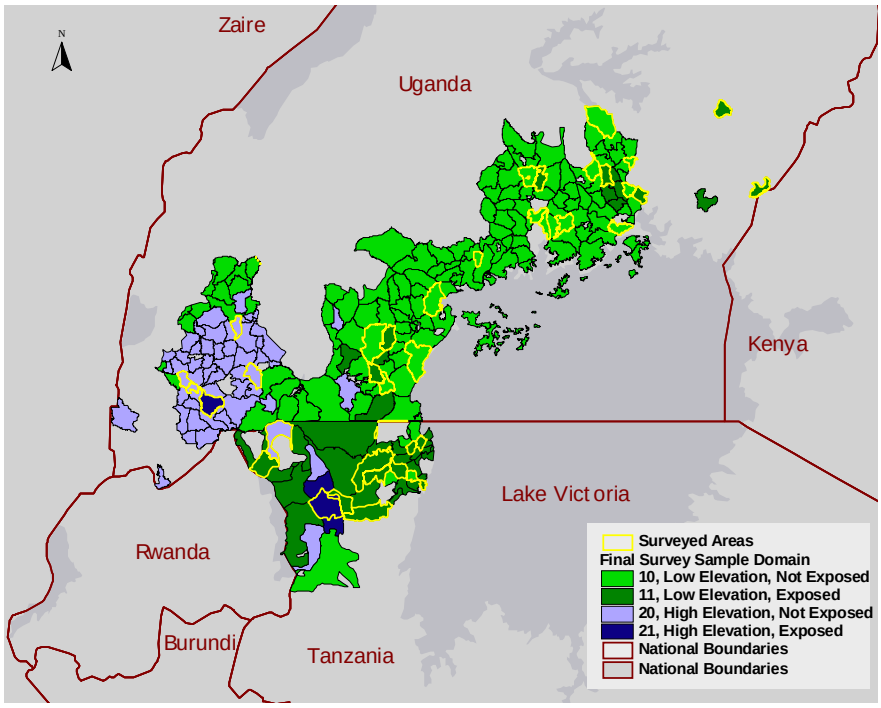

Figure 5. Sample Domain: elevation overlaid with exposure/non-exposure (Smale et al., Forthcoming)

The secondary sampling unit was a village. In Uganda, in each LC3, there are several parishes (LC2s), and each parish consists of several villages (LC1s). One SSU was

selected per PSU. The probability of selection (or sampling fraction) of an SSU varies by PSU and is denoted as ($1/M_p$), where M_p represents the number of villages in the p-th PSU (p = 1, …, 40 PSUs in the sample). For most exposed LC3s in Uganda, there is only one exposed LC1 per PSU. Where there is more than one exposed village per PSU, the SSU was drawn with a random number from the list of those villages with over 100 households according to the 1991 census. Whether or not a community selected in the sample had been properly classified as exposed or non-exposed was then verified at the site.

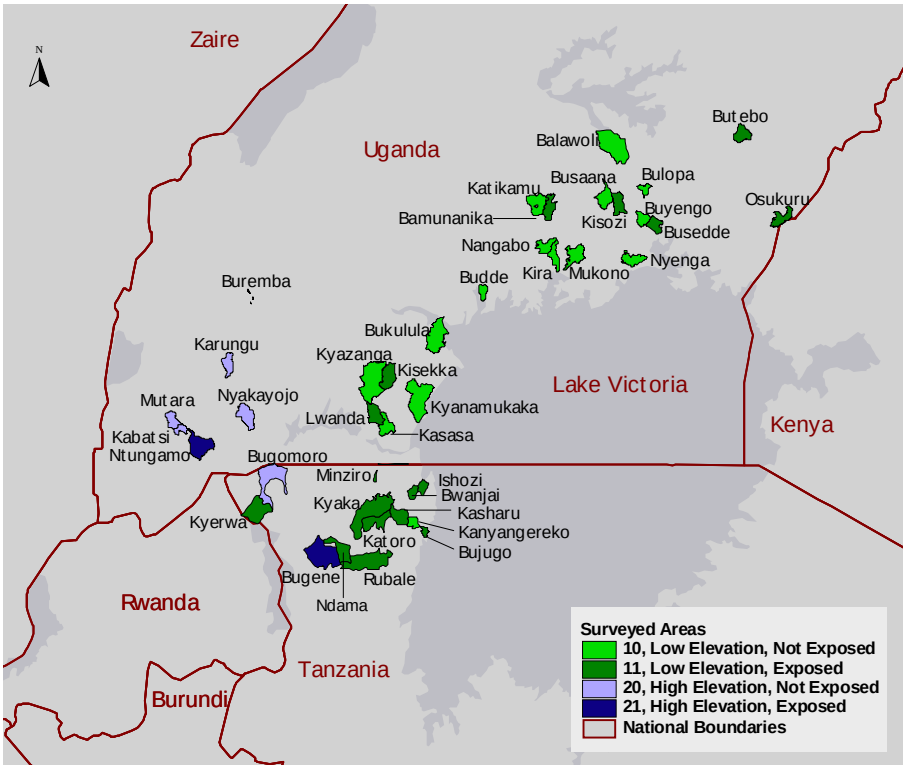

Figure 6. Sites sampled for the first survey (Smale et al., Forthcoming)

The sampled villages were visited and a current list of households in each village was requested from the chairman of the village (LC1). The total number of households selected per village was 20. The probability of selection (or a sampling fraction) of a household varies by village and is denoted as ($20/H_s$), where H_s is the number of households in the s-th village (s = *1* to *40* SSUs in the sample). If there was an order in the list of households, random numbers were used for selection. Otherwise, a random start with systematic random sampling from the list was employed.

The overall probability of selecting a household in the sub-sample (denoted as PSH) is a unique number, and it is defined as the product of the sampling fractions at each level. PSH = $[(n_i/N_i)x(1/M_p)x(20/H_s)]$. For descriptive analysis, survey weights (w) for each household were computed as the inverse PSH.

6.3. Selection of the sub-samples for the present study

The study was conducted in 20 villages selected from the Central, Eastern and Western regions. These regions were selected purposively because they represent major banana growing areas in Uganda and yet differ in their cultural dynamics and social context. This was intended to increase the variation in both the adoption of banana management practices and the social capital variables measured at the village level. The sample selected for the first survey was stratified according to region. Five villages were randomly selected from each region. Five villages were also randomly selected from a domain defined by a project entitled "Reviving banana productivity in central Uganda through introducing improved banana management technologies" implemented by NARO.

In each selected village, 20 households participating in the ongoing research projects were interviewed. The basic unit of observation for the sample survey is the farm household. A farm household is defined according to the culture of which the household is a part, and includes female-headed and child-headed (orphaned) households, as well as male-headed households with more than one wife.

6.4. Data collection methods

Data were collected by face-to-face interviews with the primary decision makers and their spouses using pre-tested questionnaires, complemented by interviews with key informants and actual observations where necessary. The questionnaires were pre-tested on at least 20 households selected from different locations across the three regions. Twenty enumerators were selected and trained in data collection, the importance of the study and data quality management. They were regularly supervised to minimize measurement errors. Survey instruments used to collect data used in the present study are summarized below.

A set of ten structured, pre-tested questionnaires (schedules) were used as instruments for data collection, and each questionnaire was designed to address a different aspect of the study. Six of the instruments were designed and implemented in the first survey. They collected data on household characteristics, farm characteristics, banana plot characteristics and market characteristics. Four were designed and implemented in the second survey. These included banana management schedules, associations and social networks. The format and structure of the banana management and social capital instruments depart in some ways from the more typical household and plot surveys often conducted in studies of technology use by smallholder farmers. These instruments are described briefly below.

Banana management schedules

The banana management schedules elicited information on the farmers' management of the natural resources on their banana plantations as well as sanitation practices (mat management) for pest and disease control, including the use and awareness of recommended practices, and sources of information by management practice. The extent of use of the organic fertilizers for a single production cycle (January to December 2003) was measured by counting the number of mats under each type of organic fertilizer. To minimize measurement errors, the interview was conducted on the plot and the farmer showed the enumerator the parts of the plantation that had been treated with the organic fertilizers as the enumerator counted mats in the area.

For mat management practices, coloured photographs were used to enhance the farmer's recognition of the practice. This helped to overcome the problem of the different names used by farmers and reduced confusion. The use of mat management technologies that could be measured as a continuous variable, such as stumping and pseudo-stem management, was measured by counting the residues that were managed and those that were not managed. The counting was conducted at the end of the production cycle so that the measurement captures the cumulative use of the technique over a period of a year. The idea was that if the farmer used these techniques extensively, unmanaged residues would not be found at the time of counting. In the few cases where neither managed nor unmanaged residues were

found[3], the enumerator identified the cause from interviewing the farmer and the observation was excluded from the analysis.

Social capital schedules

Social capital data were collected through discussions with key informants and the sampled households. Key informants (the local leaders and village elders) were interviewed about village social homogeneity (in terms of ethnicity and religious affiliation), and formal and informal organizations. Information generated through key informants helped in identifying social structures and improving the formal questionnaires.

The schedule regarding associations recorded information about household membership, the level of household participation, the major activities of associations, the benefits to the members, as well as the composition, function and leadership quality of associations. These were measured following the work of Narayan (1997). The social network schedule elicited information on the household social network, recorded information on bilateral transfers disaggregated by type of item (specified as labour, cash, food and other durable goods), the relationship between the receiver and the giver and the number of people for each item received or given. To keep the interviews confidential and encourage a high level of cooperation by the respondents, no mention of names was required. The social network schedule also elicited information on the characteristics of social network members (education, ethnicity, religion and location of residence), the major economic activities of the individuals in the social network and the intensity of social interactions and places where interactions often took place.

6.5. Basic characteristics of the sample

Household-selected demographic and socio-economic characteristics of interest are summarized in Table 6. The demographic characteristics (age, gender and education) described are those of the primary banana production decision maker. This could be the household head or another household member. Although sometimes the primary

[3] Banana pseudo-stems may be missing when they have been fed to livestock or used for vanilla or split to make mats for domestic use. However, cases of missing pseudo-stems (managed and unmanaged) were too rare to cause any selection bias in the data.

production decision maker is taken to be the household head, this measure was considered unrealistic for banana producers, given that in some locations banana is predominantly managed by women (the spouses) while the husbands (the household heads) play a minor role. During the survey, the enumerator first identified the primary production decision maker, who then became the primary respondent. The age of the primary decision maker is lower in the higher elevation areas, although this difference is not meaningful. The size of the household is, on average, five persons, with dependents (children below 15 years and adults above 64 years of age) constituting about half of the household.

In about half of the surveyed households the primary production decision maker is female, though there are significant differences in the two major production regions defined by elevation. About half of the primary decision makers in low elevation areas are women, while 68 per cent of the primary decision makers in high elevation areas are male. This may reflect the subsistence orientation in the low elevation areas, relative to the more commercial orientation of banana production in the high elevation areas. Other household characteristics do not differ across the two production regions.

Household wealth is composed of livestock, landholding and other consumer durables (radios, bicycles, motorcycles and motor vehicles). Livestock include cattle, chickens, goats, sheep and pigs and are aggregated using both physical measures (livestock units) and market values (value of livestock). There is an average livestock value of Ugsh 234 845, with 30 per cent of the sample owning no livestock. Wealth measured in terms of other consumable durable goods was even smaller (an average of Ugsh 85 859), which is about US$ 46 of wealth.

On average, each household owns 1.5 hectares of land, with the landholdings much smaller, less than one hectare, in the highlands (Table 6), but the mean cropped areas in the two production areas do not differ significantly (Edmeades et al., forthcoming). Significant differences regarding banana production between lowland and highland farmers are also evident.

Table 6. Descriptive statistics of the surveyed sample

Variable name	Aggregate sample (N= 380)	Low altitude (N=285	High altitude (N=95)	P-value
Demographic characteristics				
Age	43.230	44.084	40.653	0.0589
	(15.35)	(15.574)	(14.419)	
Gender (1=male)	0.558	0.5157	0.684	0.0049
	(0.507)	(0.501)	(0.511)	
Education	4.832	4.885	4.674	0.6408
	(4.170)	(4.342)	(3.619)	
Household size	5.746	5.822	5.5157	0.3409
	(2.715)	(2.795)	(2.458)	
Value of Assets (Ugsh)				
1) Value of Livestock	246200	237948.1	270957.9	0.8002
	(526061)	(447107.4)	(71484.4)	
2) Farm land (hectares)	1.539	1.868	0.963	0.000
	(1.731)	(2.583)	(0.977)	
3) Durable consumer goods	85858.68	82175.44	96868.42	0.7692
	(338670.2)	(289265.8)	(457501.6)	
Non-labour income (Ugsh)				
1) Private assets	106087.1	7439.76	480512.8	0.256
	(1667922)	(43811.29)	(413332)	
2) Social networks	4330.46	4602.42	3594.05	0.811
	(1873.41)	(2194.06)	(3614.03)	
Farm characteristics				
Number of banana mats (total count)	283.43	208.495	464.905	0.000
	(334.677)	(230.863)	(456.49)	
Age of banana plantations (years)	16.490	9.946	34.672	0.000
	(20.246)	(10.238)	(27.59)	
Distance from tarmac roads (Km)	10.689	10.880	10.474	0.615
	(7.050)	(7.128)	(6.669)	

The scale of banana production is higher in high elevation areas than in low elevation areas (Table 6). Furthermore, households in high elevation areas have tended their banana plantations more than three times as long (a mean of 35 years, as compared to 10 years), allocate more than twice the cropped area share to banana production, and are more likely to grow the bananas in pure stands (Edmeades et al., forthcoming). Overall, access to infrastructure development is not significantly different in the two production areas.

There are potentially two household sources of non-labour income: (1) transfers to the household from private assets (i.e. rent from buildings, land) and (2) bilateral transfers. Credit from formal institutions plays a very minor role, as noted earlier. The

descriptive statistics by Edmeades et al. (forthcoming) show that very few farmers in the sample area seek credit from formal sources (12.8%). Nevertheless, a significant number of households had access to bilateral transfers from their social networks (i.e. friends, relatives or acquaintances) that come in form of remittances, gifts or labour exchange, but the value of such transfers was small compared to the income received from private assets, such as rent from buildings, land or other assets.

Household participation in banana market transactions is described in Edmeades et al. (forthcoming). Some households choose not to participate, while others participate only as sellers, only as buyers, or as both sellers and buyers. The majority of households in the survey domain report some involvement in banana markets, although roughly a quarter of the households remain autarkic. In high elevation areas, about 90 per cent of market participation is associated with the selling of banana bunches at the farm gate. Buying banana bunches is a more common practice among households in low elevation areas (accounting for 32 per cent of market participation) than in high elevation areas. Such disparities in market participation between the two production areas reflect the geographical shift in the locus of banana production from the Central region to the south-western highlands, where bananas are transported and distributed across markets in the lowlands.

6.6. Summary

Knowing the source of data and collection procedures is important for the interpretation of any research results. This chapter has indicated that the research for this study used primary data collected from households from the villages selected to represent the domain. The sampling frame used and methods of data collection have been discussed. The data collection and survey instruments were designed taking into account the problems of recognition and the low literacy of farmers. Basic characteristics of the sample are also highlighted. The chapter indicates that the majority of households interviewed for this study had only a primary level of education, which is also the national average level of education for rural farmers, thus confirming that the sample is representative of banana producers in Uganda. Both female and male farmers were interviewed.