Linking processes and pattern of land use change

Overmars, K.P.

4

Multilevel modelling of land use from fi eld tc village level in the Philippines

Abstract

In land use research regression techniques are a widely used approach to explore datasets and to test hypotheses between land use variables and socio-economic, institutional and environmental variables. Within land use science researchers have argued the importance of scale and levels. Nevertheless, the incorporation of multiple scales and levels and their interactions in one analysis is often lacking. Ignoring the hierarchical data structure originating from scale eff ects and levels may lead to erroneous conclusions due to invalid specifi cation of the regression model. The objective of this chapter is to apply a multilevel analysis to construct a predictive statistical model for the occurrence of land use. Multilevel modelling is a statistically sound methodology for the analysis of hierarchically structured data with regression models that explicitly takes variability at diff erent levels into account. For a land use study in the Philippines multilevel models are presented for two land use types that incorporate the fi eld, household and village level. The value of multilevel modelling for land use studies and the implications of multilevel modelling for data collection will be discussed. The results show that explanatory variables can account for group level variability, but in most cases a multilevel approach is necessary to construct a sound regression model. Although land use studies often show clear hierarchical structures, it is not always possible to use a multilevel approach due to the structure of most land use datasets and due to data quality. Potentially, multilevel models can address many important land use issues involving scales and levels. Therefore, it is important in land use change research to formulate hypotheses that explicitly take scale and levels into account and then collect the appropriate data to answer these questions with approaches such as multilevel analysis.

Based on: Overmars, K.P., Verburg, P.H. 2006. Multilevel modelling of land use from fi eld to village level in th Philippines. Agricultural systems 89, 435-456.

4.1 Introduction

In the past decade substantial advances has been made in land use and land cover chang
(LUCC) research by the development of a wide range of analytic tools to observe, explor
and model LUCC (Lambin *et al.*, 1999; Rindfuss *et al.*, 2004, Veldkamp and Verburg, 2004)
In general, LUCC is considered to be the result of the interplay between socio-economic,
institutional and environmental factors, the so-called 'driving forces' of land use change
These driving forces are oft en subdivided into proximate causes and underlying causes
Proximate causes are the activities and actions that directly aff ect land use. Underlying
causes are the fundamental processes that underpin the proximate causes, including de-
mographic, economic, technological, institutional and cultural factors (Geist and Lambin
2002). A widely used approach to explore the relations between land use (changes) and
the underlying causes are regression techniques of various kinds (e.g. Nelson *et al.*, 2001;
Chomitz and Thomas, 2003; Perz and Skole, 2003; Verburg *et al.*, 2004b). The approach in
this chapter makes use of a regression technique that explicitly can deal with issues of sc
and levels, which are characteristic for land use studie
Within the LUCC discipline as a whole and in reference to regression approaches in par-
ticular, LUCC scientists have argued the importance of scale and levels (e.g. McConnell
Moran, 2001; Veldkamp and Lambin, 2001; Walsh *et al.*, 2001; Nelson, 2002; Rindfuss *et al.*,
2004). Gibson *et al.* (2000) state that scale is the spatial, temporal, quantitative, or analytic
dimension used by scientists to measure and study objects and processes and level refer
to specifi c locations along a scale. For this chapter the following defi nitions are used: Le
refer to organisational levels originating from social context, for example, household lev
village level and municipality level and scale is used for artifi cial resolution and extent
originating from a geographic representation of reality in maps. The following issues re-
garding scales and levels that are important in land use (change) analysis can be identifi
(Gibson *et al.*, 2000; Verburg *et al.*, 2004d). First, land use is the result of processes that act
diff erent scales and levels, which ideally would be addressed simultaneously. The choic
that are made in a study about the extent and the unit of analysis determine to a large
extent what patt erns will be observed and which correlation will be found. Oft en, these
choices are diff erent between disciplines (Verburg *et al.*, 2003). Second, scale and levels ar
important in identifying relations, but the fact that a relation occurs at a certain scale or l
does not explain the phenomenon. Therefore, causal statements between variables shou
be made explicit and tested. Within these causal statements scale and level are importan
factors, because diff erent relations occur at diff erent scales and levels. Moreover, causa
relations can occur between diff erent scales and levels. For example, village level variab
like population or leadership capacity of the village head can infl uence land use at fi elc
level. Third, aggregation of processes to a higher level does not straightforwardly lead t
a proper representation of these higher level processes because relations identifi ed at th
micro-level (or fi ne resolution) does not automatically translate into the same relation a
the macro-level (or course resolution) (Robinson, 1950; Jones and Duncan, 1995; Easterli
1997). The other way around the same phenomenon occurs: Inferences made on higher
levels can oft en not be directly translated to lower level processes. Finally, all analyses,
therefore the insights from these analyses, are bounded by resolution or level of analysis
and extent, which are determined by data structure and choices made by the researche
Mostly, scale and level issues are identifi ed by comparing analyses at diff erent resoluti
and levels. Geoghegan *et al.* (2001) and Overmars and Verburg (2005) (Chapter 2) compar

an analysis of land use decisions based on a household dataset with an analysis using a spatial dataset. Walsh *et al.* (2001) and Veldkamp *et al.* (2001) analysed the relation between land use and its explanatory factors at different resolutions created by aggregating grid data. However, the incorporation of multiple scales and levels in the analysis and include interactions between levels is often lacking. So far, the statistical tools that explicitly deal with these issues are not often applied as noted by Pan and Bilsborrow (2005) and Polsky and Easterling (2001). Multilevel modelling, which is the approach used in this study, is one of the statistical tools that are potentially capable to integrate artificial scales and organisational levels and to include interactions between these scales and levels. Multilevel statistical modelling allows for the analysis of data with complex patterns of variability that originate from hierarchical structure (Snijders and Bosker, 199
Multilevel modelling has mainly been used in the social sciences, for example, in sociology, education, psychology, economics, criminology (Snijders and Bosker, 1999), and is becoming more popular in geographic applications (e.g. in studying transport and land values (Schwanen *et al.*, 2004; Polsky and Easterling, 2001)). In most of these applications multilevel modelling is used to study the effects of social context on the individual behaviour and to study the confusion between aggregate and individual effects. Land use studies can potentially benefit much from multilevel analysis, because land use data often has a very clear hierarchical structure (e.g. administrative levels, agro-ecological divisions and subdivisions, societal levels, artificial scales). Therefore, it is remarkable that multilevel modelling is not (yet) widely applied in land use studies. Some land use studies do incorporate data from multiple levels, but only few actually use multilevel modelling (Hoshino 2001; Pan and Bilsborrow, 2005)

This chapter aims to use multilevel analysis as the methodology to construct a predictive statistical model for the occurrence of land use that is statistically sound and which integrates different scales and levels. On the basis of a case study from a municipality in the Philippines different multilevel models will be presented that explain the occurrence two major crops on individual fields in the area. In the discussion we explore and describe the (surplus) value of multilevel modelling for land use studies regarding the issues of scale and levels in LUCC research and describe the implications of multilevel modelling for data collection

4.2 Multilevel analysis

In this section a short introduction of multilevel models is given in respect to land use issues in a general manner regardless of the outcome variable. Specific differences exist between models with a continuous, binary or multinomial outcome variable regarding estimation, model formation and the interpretation of coefficients. For the case study the logistic approach was adopted and the model specification is given in Section 4.3
Multilevel analysis (e.g. Goldstein, 1995; Snijders en Bosker, 1999) is a methodology developed for the statistical analysis of hierarchically structured data. Multilevel regression models explicitly take the variability at different levels into account. Therefore, it is potentially a valuable tool in dealing with scaling issues in land use analysis. Multilevel modelling can address the scales and levels that are important to the land use system simultaneously, it can test hypothesis between scales and the modeller is not forced to aggregate or disag-

gregate data to one unit of analysis. Multilevel modelling can deal with nested data, suc
as hierarchically structured administrative units (e.g. farms in municipalities), as well as
handle cases with observations that are structured diff erently, like lower level observati
that are member of several groups at the higher level (e.g. farmers that have several buy
for their products)
Fundamental to multilevel modelling is "that the outcome variable Y has an individual
as well as a group aspect" (Snij ders and Bosker, 1999). This is refl ected in the model by
including explanatory variables at the individual level and at the group level, as well as
the way unexplained variation is modelled. Both unexplained variation within groups a
unexplained variation between groups is conceived as random variation and is expresse
in multilevel models as 'random eff ects'. Thus, multilevel models include an error term
every level in the model (Snij ders and Bosker, 1999). Multilevel models can be construct
by including random intercepts only or by including both random intercepts and rando
slopes. Furthermore, variables can be added to the model to explain variability at the
individual and group level, and also to explain the diff erences in slopes. For example, in
model with a household and a village level a random intercept can account for unobserv
structural eff ects between villages. These structural eff ects may be caused by diff erenc
in technology. Including an explanatory variable "technology" could explain part of the
structural eff ects. Random slopes actually incorporate diff erences between groups in th
rate of change in output per unit change in the explanatory variable (*i.e.* the regression
coeffi cients). For example, if you were predicting yields a random slope at the village le
for soil fertility would account for diff erences between villages in the relation between s
fertility and yield, which may be caused by an unobserved diff erence in use of chemical
fertiliser
Multilevel models are applicable to data with hierarchical structures of various origins.
Also for data that are acquired by using a multistage sampling scheme, and have therefo
a hierarchical structure, a conventional regression model may be incorrect and a multile
model would be a statistically sound method. In a multistage sampling design the select
of lower level observations depends on the choices made at higher levels. An example o
multistage sampling approach, when conducting a regional survey among land owners,
to fi rst sample villages and than sample people within these villages. In this case the da
at the lower level is not independent from the higher levels and therefore a conventiona
statistical approach might lead to underestimation of the standard errors (Rasbash *et al*.,
2000). In any case, having some kind of hierarchy in the data, a multilevel analysis will
model this hierarchy explicitly and prevent erroneous model inferenc
If the multilevel structure of the data is ignored the data will inevitably be analysed at ei
an aggregate level or a disaggregate level. Analysing aggregated data, like in the work o
Perz and Skole (2003), can only tell us something about the relation between macro-leve
variables. Analysing macro-micro or micro-level propositions with aggregated data may
result in gross errors (Jones and Duncan, 1995) because by aggregating the data the vari
changes in its meaning and cannot be used anymore to draw conclusions at the lower le
This phenomenon is called the ecological fallacy: A relation identifi ed between macro-le
does not automatically translate into the same relation at the micro-level (Robinson, 195
Jones and Duncan, 1995; Easterling, 1997). A drawback of aggregation is that it disables
examination of cross-level relations, for example, when a micro-level relation diff ers by
macro-level group or depends on a macro-level variabl
Disaggregation of macro-level data into micro-level data, by assigning the values of a fe

higher level observations to all lower level units, results in an exaggeration of the sampl size. Wrongly assuming that all these observations are independent leads to an over-cor dence in the estimated level of signifi cance (due to underestimation of the standard erro which in turn leads to elevated probabilities of a type I error when studying between gr diff erences (type I errors: concluding there is a relation while in reality there is none). W studying within group diff erences it can result in failing to detect a relation (Snij ders a Bosker, 1999; Rasbash *et al*., 2000; Polsky and Easterling, 2001).

4.3 Material and methods

4.3.1 Study area

The study area is situated in Cagayan Valley in the northeastern part of the island Luzo the Philippines (Figure 4.1). The study area includes 20 villages (*barangays*) in the munici- pality of San Mariano, in the province of Isabela, and comprises approximately 480 km2. It is situated between the town of San Mariano in the west and the forested mountains o the Sierra Madre in the east. The population is approximately 20,000 persons (about 4,0 households) of various ethnic groups, among whom the Ilocano, Ibanag and Ifugao, wh are all migrants or descendents of migrants that came to the area from the 1900s onward and the Kalinga and Agta, who are the indigenous inhabitants. Before immigration start the area was completely forested with tropical lowland forest. At present, the study area shows a clear land use gradient ranging from intensive agriculture (mainly wet rice and yellow corn) near San Mariano via a scatt ered patt ern of wet rice, yellow corn, banana, grasses, and (fruit) trees in the foothills to residual and primary forest in the eastern par the area a village unit actually consists of a group of sett lements (*sitios*). The people live in

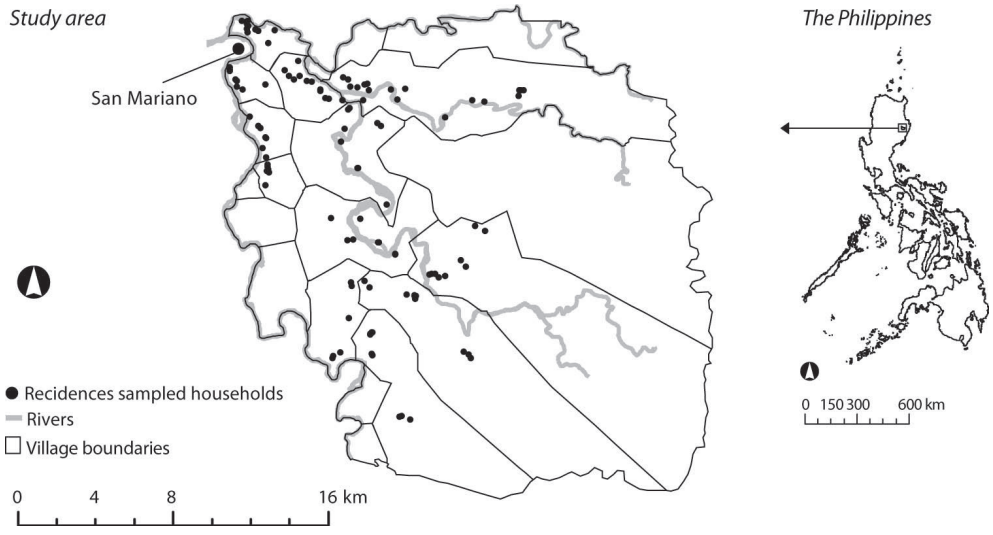

Figure 4.1: Location of the study area in the Philippines and the location of the households' homes within the area

these sett lements, while their fi elds are oft en located in the surroundings of the sett ler at an average distance of about 30 minutes walkin

4.3.2 Data

Data were collected in 13 of the 20 villages between June and November 2002 by interviewing households about their land use practices and household characteristics using a structured questionnaire. The questionnaire was designed to create an exhaustive list of variables that might explain land use decisions. This list of variables was based on litera ture, theories from a range of disciplines and expert knowledge of the area (see Overma and Verburg, 2005 (Chapter 2) for more information). For the analysis in this study a sul of variables was used

The selection of households to be interviewed was based on systematic random samplin using population data available at the POPMAT (POPulation Manipulation Action Tean member in the village. In all 13 villages every twentieth household was selected (system atic random sampling with sampling interval 20) from the POPMAT's list. From a total approximately 3150 households in the 13 villages, 151 households were interviewed. Th number of interviews per village ranges from 6 in the least populated village to 20 in the most populated. For the selected households the relevant characteristics were recorded all fi elds (where a fi eld is defi ned as a piece of land of a single owner used for one crop type). A household oft en owns or uses a number of fi elds at diff erent locations and wh are cultivated with diff erent crop

The most detailed (nested) hierarchy in the area, relevant to the land use system, could b constructed as follows (from the lowest level to the higher level): fi elds - plots (where a plot consists of a number of adjacent fi elds from the same owner) - household *sitios* (the sett lements) - villages - municipality. For the analysis only the fi eld, household and vill level were used (see Figure 4.2). This is the most functional grouping, because the plots consist mostly of only one fi eld and the dataset does not contain enough observations to discriminate between *sitio* and household level. Most *sitios* have only one or two households within the sample, which is insuffi cient for a proper multilevel analysis. Each of variables was collected at its corresponding level, e.g. soil characteristics and slope at fi level and household structure at the household level. Village level variables were derive from census data of 1997 (data about ethnicity and the percentage of the population that born in the municipality of San Mariano)

Records with missing data were omitt ed from the dataset. Table 4.1 presents the dataset as it was used in the analysis, which is a subset of the original dataset and includes the most relevant variables based on preceding research and fi eld experience (Overmars an Verburg, 2005 (Chapter 2))

4.3.3 Multilevel model specifi cation

Multilevel models can be constructed in various forms with diff erent levels of complexi In this section we start with the description of a simple model to explain how we arrive at the model that we will use to explain the occurrence of land use. The description of th models is based on Snij ders and Bosker (199

Since we will estimate a binary response variable (land use choice) we start with a conve tional multiple logistic regression model (Equation 4.1

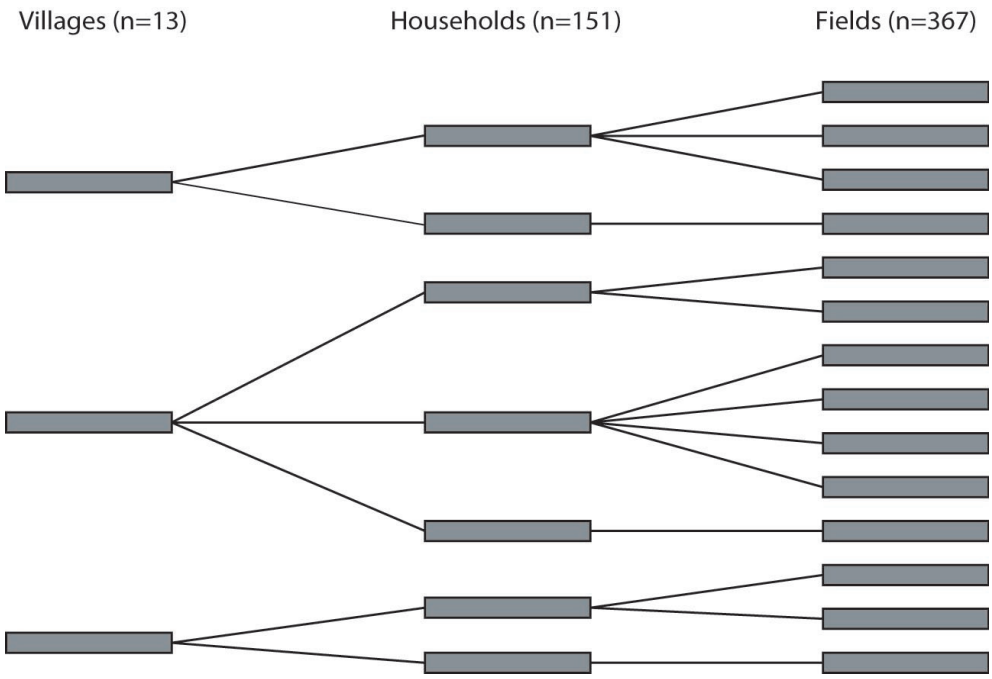

Figure 4.2: Schematic representation of the hierarchical structure of the dataset

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (4.1)$$

In this model p is the probability for the occurrence of the event, which in this study is th occurrence of a land use type on a fi eld $ß_0$ is an intercept, $ß_n$ are regression coeffi cients t be estimated, and the x_n are exogenous explanatory variables
The simplest imaginable way to incorporate levels would be to identify explanatory var ables at the lower and the higher level (Equation 4.2

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \alpha_1 z_1 + \ldots + \alpha_m z_m \qquad (4.2)$$

Here, $ß_n$ and α_m are regression coeffi cients to be estimated, and the x_n are exogenous explanatory variables at the lower level (e.g. fi eld) and z_m are explanatory variables at the higher level (e.g. a household)
Actually, many studies apply this approach by including variables from diff erent levels the regression model (e.g. Müller and Zeller, 2002; Overmars and Verburg, 2005 (Chapte 2)) but do not report this explicitly. This model is typically called a fi xed eff ect model si it lacks the random eff ects corresponding to higher levels in a multilevel model (Snij de and Bosker, 1999). The assumptions that belong to this model are that the residuals are mutually independent and have a zero mean. An additional assumption that is oft en m

Table 4.1: Description of the variables in the dataset used in this study

Variable name	Description	Min.	Max.	Mean	St. dev.
Dependent variables (field level: level 1, n=297					
Yellow corn	1 if yellow corn, 0 otherwise	0	1	0.532	
Banana	1 if banana, 0 otherwise	0	1	0.215	
Independent variables at field level (level 1, n=297					
Slope1	1 if slope category is flat, 0 otherwise	0	1	0.380	
Slope2	1 if slope category is flat to rolling/moderate, 0 otherwise	0	1	0.229	
Slope3	1 if slope category is rolling/moderate, 0 otherwise	0	1	0.283	
Slope4	1 if slope category is rolling/'moderate to steep/hilly, 0 otherwise	0	1	0.081	
Slope5	1 if slope category is steep/hilly, 0 otherwise	0	1	0.027	
Creek	1 if there is a creek or spring trough or bordering the plot, 0 otherwise	0	1	0.593	
Plot distance	Hours walking from the residence of the household to the plot (hrs)	0	10	0.511	
Independent variables at household level (level 2, n=115)					
Ethnicity Ilocano	1 if male household head is Ilocano (or Tagalog speaking), 0 otherwise	0	1	0.539	
Ethnicity Ifugao	1 if male household head is Ifugao, 0 otherwise	0	1	0.087	
Ethnicity rest	0 if ethnicity is Ilocano or Ifugao, 1 otherwise	0	1	0.374	
Transportation cost	Cost to transport a bag of corn from the residence to San Mariano (pesos)	7	45	22.652	12.214
Municipality of origin 0	1 if both male and female were not born in San Mariano, 0 otherwise	0	1	0.244	
Municipality of origin 1	1 if male or female head is born in San Mariano, 0 otherwise	0	1	0.322	
Municipality of origin 2	1 if both male and female were born in San Mariano, 1 otherwise	0	1	0.435	
Independent variables at village level (level 3, n=12)					
Ethnicity Ilocano (village)	Fraction of the population of the village that is Ilocano (or Tagalog speaking)	0.021	0.900	0.573	0.259
Ethnicity Ifugao (village)	Fraction of the population of the village that is Ifugao	0.000	0.404	0.076	0.147
Municipality of origin (village)	% of the population of the village born in San Mariano	64.899	99.007	84.479	9.748

is that all groups have the same variances (homoskedasticity assumption). Implicitly the assumption is made that all group structure is represented by the explanatory variables. If this is not the case the residuals will be heteroskedastic. A second problem with this approach is that the higher level data is oft en disaggregated to the lowest level. As said before, this will lead to type I errors. The following models describe how the eff ects of the diff erent levels can be incorporated into the regression model. With these models th assumptions stated above can be tested

The model in Equation 4.3 incorporates group eff ects but as yet without any explanator variables. Besides the general intercept a random tern U_{0j} is introduced, which is a group dependent intercept, in other words, an error term at the group level. With this random term the variance that exists between groups is modelled explicitly. The eff ect of being a 'member' of a specifi c group is taken into account. Introducing this term will help to prevent the residuals from being heteroskedasti
For reasons of clarity indices mark the diff erent levels: *i* for level 1, *j* for level 2 (and *k* for level 3) and a zero indicates that a parameter is not variable at that lev

$$\log\left(\frac{p_j}{1-p_j}\right) = \gamma_{00} + U_{0j} \qquad (4.3)$$

In Equation 4.3 γ_{00} is the general intercept and U_{0j} is the group dependent deviation. The deviations U_{0j} are assumed to be independent and normally distributed with a zero mea and a variance of τ_0^2 (Snij ders and Bosker, 1999
This model is called the 'pure random eff ects model', 'empty model' or 'unconditional model'. The empty model is a random intercept model without explanatory variables. W this model the variance of the dependent variable can be decomposed in a part caused b the individual level and a part caused by the group level (Snij ders and Bosker, 1999; Po and Easterling, 2001). We will use this model as the base model to estimate if the group level variance in the dependent variable is signifi cant. In the case study this is called mo 1.

Including explanatory variables leads to the following mode

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \gamma_{10}x_{1ij} + ... + \gamma_{q0}x_{qij} + \gamma_{01}z_{1j} + ... + \gamma_{0r}z_{rj} + U_{0j} \qquad (4.4)$$

where x_{qij} are q explanatory level-1 variables and z_{rj} are r explanatory level-2 variables. Again, the deviations U_{0j} are assumed to have zero mean (given the values of the explana tory variables) and a variance of τ_0^2 (Snij ders and Bosker, 1999). This model (Equation 4.4 a random intercept model: a model where the intercept varies randomly between group The fi rst part $\gamma_{00} + \gamma_{10}x_{1ij} + \ldots + \gamma_{q0}x_{qij} + \gamma_{01}z_{1j} + \ldots + \gamma_{0r}z_{rj}$ is called the fi xed part of the mode and the second part U_{0j} is the random part of the model. In the case study analysis mod 2, 3, 4 and 5 are based on this model (note that in the case study the model is extended t a model with 3 levels)
The interpretation of the regression coeffi cients is similar to ordinary logistic regressior and is facilitated by the odds ratio $exp(\gamma)$). The odds ratio can be interpreted as the chang in odds for the considered event upon an increase of one unit in the corresponding facto

while the other factors are considered to be unchanged. This means that the odds $p/(p-1)$, are multiplied by $exp(\gamma)$ for every unit increase of the variable corresponding to γ (Neter *et al.*, 1996).

Starting from the empty model variables can be added at all levels. Variables at the individual level can explain part of the individual level variability as well as part of the group level variability, in the case when the values of the level one variable are consistently higher or lower than the general mean. For example, the slopes of the fields can be consistently higher in some of the villages and lower in others. Incorporating this field level variable can account for village level variability detected with the empty model.

Variables at the higher level(s) can be grouped in higher level variables that can only be observed at the higher level (e.g. the presence of a secondary school in a village) and aggregates of lower level variables (e.g. the average income of the inhabitants, which is an aggregate of observations at a lower level). Including these aggregates allows for the separation of the effect at the individual level and the effect at group level, which gives insight in the way a variable influences the outcome. In a model with only the level 1 data of that variable included the effect at both levels is forced to be equal (Snijders and Bosker, 1999). This difference is important while interpreting the regression coefficient. As described in Section 1 processes at the aggregate level can be substantially different from processes at the individual level. The village level variables in this study are of the aggregated type. Although they were calculated from census data, they have their equivalent at the household level.

The random intercept model (Equation 4.4) can be expanded by introducing random slopes. In a model with random slopes the regression coefficient (γ_{q0}) that act on the explanatory (level 1) variables are subdivided in a fixed and a random part. The addition of random slopes allows specific variables to differ by group. Even more complexity can be modelled by introducing level 2 variables in these slopes to explain (part of) the differences in slope. This is actually the same as a cross-product with an explanatory variable from level 1 and an explanatory variable from level 2. In multilevel modelling this cross-product is called a cross-level interaction (Snijders and Bosker, 1999). In this study random slopes and cross-level interactions were not included in the models. This will be explained in greater detail in the final discussion.

In the case study models with three levels were applied (Equation 4.5), which is just an expansion of the model in Equation 4.4. The first model in the analysis is a pure random intercept model (empty model) with 3 levels. The subsequent models (models 2, 3, 4 and 5) are random effect models with three levels (Equation 4.5).

$$\log\left(\frac{p}{1-p}\right) = \gamma_{000} + \gamma_{100}x_{1ijk} + \ldots + \gamma_{q00}x_{qijk} + \gamma_{010}z_{1jk} + \ldots + \gamma_{0r0}z_{rjk} + \gamma_{001}a_{1k} + \ldots + \gamma_{00s}a_{sk} + R_{0jk} + U_{00k} \quad (4.5)$$

In Equation 4.5 the a_{sk} are s explanatory level 3 variables, the R_{0jk} is the level 2 random part and U_{00k} is the level 3 random part. In this model fields are the unit of analysis at level 1, level 2 consists of households and level 3 are the villages. The dependent variable Y is land use. If Y = 1 the land use occurs, if Y = 0 the land use does not occur and p is the probability that the land use is found on that field.

Two analyses will be presented: one explaining the occurrence of yellow corn and one explaining the occurrence of banana. These are the most dominant crops in the study ar (53 % of the fi elds were cultivated with corn and 22 % with banana). In the analysis we present fi ve diff erent random intercept models per land use type. The fi rst model is th empty model, which informs about the variability at the diff erent levels. In the subsequ models variables will be added per level to see the infl uence of these groups of variable the variance component of the higher level
The variables included were selected by studying prior analyses (Overmars and Verbur 2005 (Chapter 2), Overmar *et al.*, 2006 (Chapter 3)) and fi eld experience. For the corn mo variables from the following list were added in diff erent compositions: slope, creek and plot distance at the fi eld level; transportation cost, ethnicity and municipality of origin a the household level; and averages of municipality of origin and ethnicity at the municip level. For the banana model the same variables were used except for presence of creeks, because this was considered to be of no infl uence to the occurrence of banar
The analysis is performed with HLM soft ware (Raudenbus *et al.*, 2004). All models were estimated using the PQL (Penalized Quasi likelihood) routine. In HLM6 all 3-level hiera chical generalised linear models are estimated by full PQL by default (Snij ders and Bosl 1999; Raudenbush *et al.*, 2004).

To indicate the proportion of variance that is accounted for by the group level the intrac correlation coeffi cients ρ_R and ρ_U, for the household and village level, respectively) are calculated. Equation 4.6 shows the calculation of the intraclass correlation coeffi cient fo the household level. (Snij ders and Bosker, 1999; Brow *et al.*, 2005).

$$\rho_R = \text{var}(R_{0jk}) / (\text{var}(R_{0jk}) + \text{var}(U_{00k}) + \pi^2 / 3) \qquad (4.6)$$

Where ρ_R is the intraclass coeffi cient for the household lev $var(R_{0jk})$ is the variance of the random intercept at household level and $var(U_{00k})$ is the variance of the random intercept at village level. A logistic distribution for the level one residual implies a variance o $\pi/3$, which appears as the level 1 variance in Equation 4.6 (Snij ders and Bosker, 1999). In an linear multilevel model this would be the level 1 varianc σ^2.
To assess the goodness-of-fi t of the models the ROC (Relative Operating Characteristic) (Swets, 1988) was used . This measure is capable to assess the quality of the predictor an can be compared between diff erent models. The ROC summarises the performance of a logistic regression model over a range of cut-off values classifying the probabilities. The value of the ROC is defi ned as the area under the curve linking the relation between the proportion of true positives versus the proportion of false positives for an infi nite numb of cut-off values. The ROC statistic varies between 0.5 (completely random) and 1 (perfe discrimination).

4.4 Results

4.4.1 Corn models

This section presents various multilevel models, with diff erent sets of explanatory varia predicting the occurrence of yellow corn on a fi eld. Model 1 is the empty model, which

does not include any explanatory variables, but only includes random eff ects at the higl
levels. Model 1 (Table 4.2) shows that the variance is signifi car$p<0.05$) at both level 2 anc
3. The intraclass correlation coeffi cients ρ_R and ρ_U, Table 4.2) indicate that 10 percent of
the variance can be att ributed to the household level and 4 percent to the village level. T
remaining variance is in level 1, which is fi xed in this modelling approach to $\pi/3$. Thus,
both the households and the villages show signifi cant clustering of the occurrence of co
The variance detected in this model might be accounted for by explanatory variables. Th
is studied with the models 2, 3, 4 and 5
Model 2 introduces a set of geographic and biophysical variables that are known explan
tory variables for the occurrence of corn in the study area. These are slope, presence of
a creek, hours walking from the residence of the household to the plot, and the cost to
transport a bag of corn from the residence to San Marianc
Table 4.2 (model 2) shows that almost all explanatory variables (the fi xed eff ects) have
signifi cant coeffi cients. Corn is more likely to occur on fi eld that are fl att er, not close
creek, close to the household's residence and close to the market town of San Mariano. T
random part of level 3 turns out to be lower. So, these variables explain some of the vari
at the village level detected in the empty model. This might be caused, for example, by t
fact the transportation costs vary on average per village because the villages are situatec
at diff erent distance from the market place. By introducing this variable (or perhaps one
the other variables) the variability disappeared from the village level. Although the leve
3 random part is not signifi cant and, theoretically, level 3 could be excluded, the structu
with three levels is maintained in order to study the level 3 behaviour in the following
models. At the household level the variance component is still signifi cant and similar tc
variance component of the empty model. Thus, the variables included do not account fo
any of the household level variability
Model 3 adds household variables to model 2. Additional to the relations in model 2 cor
turned out the be negatively related with households where both the male and female a
born outside the municipality and negatively with people of Ilocano origin. Aft er incluc
the household level variables the random part of level 2 (the household level) decreased
substantially. Apparently, the variance at level 2 is captured by the included variables. T
geographical/biophysical variables are still signifi cant. The level 3 variance increased in
comparison with model 2
Model 4 adds the village level variables to the model 3 confi guration. This model inves-
tigates if there is a fi xed eff ect of the village level variables besides the variables includ
in model 3. For example, one can imagine that a village dominated by one ethnic group
has an extra village level eff ect besides the eff ect of ethnicity at household level for the
whole study area. The village level variables are aggregated values of variables at level 1
(ethnicity and municipality of origin). Instead of using the survey data to derive these le
3 variables census data of the complete population was used. Table 4.2 shows that there
are no signifi cant eff ects for the variables at village level. Including these variables resu
in a similar random part at the village level as model 3. Thus the level 3 variables did nc
explain any of the variance in level 3
Model 5 was constructed to see if including the household variables at village level inste
of at the household level would be a good alternative. This would be convenient becaus
census data at village level is oft en more easily available then household level data. Ho
ever, like in model 4, none of the village level variables are signifi cant in model 5. Besid
that, the variance component at household level is the same at in the models 1 and 2. Th

Table 4.2: Multilevel models for yellow corn

Yellow corn	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects					
Level 1					
Intercept	0.201	-0.546	-0.217	0.330	-1.083
Slope1		3.108**	3.688**	3.572**	3.113**
Slope2		3.446**	4.091**	4.007**	3.397**
Slope3		2.274	2.768*	2.657*	2.281
Slope4		-0.469	-0.560	-0.629	-0.603
Creek		-0.833**	-0.759*	-0.742*	-0.843**
Plot distance		-0.586*	-0.599*	-0.616*	-0.569*
Level 2					
Transportation cost		-0.050**	-0.050**	-0.055**	-0.038*
Ethnicity Ilocano			-0.929*	-0.973*	
Ethnicity Ifugao			-0.997	-1.094	
Municipality of origin 0			-1.313**	-1.347**	
Municipality of origin 1			0.233	0.239	
Level 3					
Municipality of origin village				-0.744	0.845
Ethnicity Ilocano village				0.314	-0.384
Ethnicity Ifugao village				0.283	-0.781
Random effects					
Level 2					
var (R_{0jk})	0.395*	0.441**	0.001***	0.004***	0.494**
ρ_R	0.103	0.115	0.000	0.001	0.130
Level 3					
var (U_{00k})	0.143*	0.103	0.187*	0.177*	0.018
ρ_U	0.037	0.027	0.054	0.051	0.005
ROC	0.855	0.881	0.864	0.863	0.882

** p<0.05, ** p<0.01, *** p<0.001*

shows that the aggregated variables do not capture any of the variability at the household level. As theory suggests (Robinson, 1950; Jones and Duncan, 1995) the effect of the aggregate variable is quite different than that of its lower level equivale

The ROC value of the corn model 1 is 0.855. The ROC value of model 2 is about the same as model 1. This indicates that including the variables at field level does not lead to better predictions, because then the ROC would be higher if they did explain field level variar However, the variables included in model 2 do explain part of the village level variance, which is showed by a lower variance component at the village level and significant regression coefficient

4.4.2 Banana models

The analysis of the occurrence of banana shows a diff erent result (Table 4.3). In the emp
model (model 1) there are no signs of signifi cant between-group variances. Model 2, wh
incorporates the geographical/biophysical variables, shows a signifi cant relation betwee
slope of a fi eld and the choice to cultivate banana and a signifi cant random eff ect at the
village level. Thus, including variables results in a large and signifi cant random part at
the village level. This could not be explained completely. Part of the explanation is that i
general changes in the fi xed eff ects part can cause big changes in the random part whil
changes in the random part usually do not cause big changes in the fi xed eff ect part. Th
mean structure the model can change the stationarity of the mean causing a shift in vari
ance and making the random eff ects part signifi ca
Model 3 introduces household level variables ethnicity and municipality of origin, mode
includes also village averages of these variables and model 5 includes only the village le
and fi eld level variables. All the coeffi cients of these variables do not diff er signifi can
from zero, neither do they infl uence the level 2 and level 3 random parts signifi cantly.
Therefore, we conclude that these variables do not infl uence banana cultivation signifi c
and that this is predominantly determined by slope. To fi nd out what process might cau
the diff erences between villages the random intercepts of the village level were examine
This did not show a clear patt ern. Furthermore, models with additional explanatory var
ables and models with random slopes were tested , but this did not result in a satisfying
explanation of the variability at the village level in model
The ROC of banana model 1 is 0.694. Model 2 has an ROC of 0.906. This indicates that th
slope of the fi elds does explain part of the variance at the fi eld level. Including variable
model 3, 4 and 5 does not produce a higher ROC than model 2, which is obvious, becaus
in the model 2 the random part is included in the predicted values and no additional lev
1 variables are included. The two random parts accounts for all variance at level 2 and 3
The diff erence between model 2 and models 3 ,4 and 5 is that the variables at household
and village level can explain part of the variance. However, in this model the explanator
factors at household and village level are not signifi cant and the variance of the random
part is similar throughout models 2, 3, 4 and 5

4.5 Discussion and conclusions

4.5.1 Multilevel statistics for land use studies

In this section the main fi ndings of the multilevel analysis are discussed for the case stu
Then, these fi ndings are used to evaluate the advantages and disadvantages of multilev
analysis for land use studies in general
For corn cultivation the empty model indicated signifi cant between-group variability at
higher levels (household and village). Explanatory variables at the household level turn
out to account for that variability at that level (Table 4.2, model 3). Replacing some of th
household level variables with their village level aggregates did show a signifi cant vari
component at the household level. From this it can be concluded that the household leve
variables cannot be substituted by village level aggregates in this case. The explanatory
variables at the household level can explain a signifi cant part of the occurrence of corn a

Table 4.3: Multilevel models for banana

Banana	model1	model2	model3	model4	model5
Fixed effects					
Level 1					
Intercept	-1.289***	-3.430***	-3.260**	-4.050	-4.120
Slope3		2.389***	2.435***	2.432***	2.397***
Slope4		5.022***	5.006***	4.975***	4.970***
Slope5		5.634***	5.971***	5.765***	5.461***
Plot distance		0.006	0.007	0.023	0.020
Level 2					
Transportation cost		0.015	0.012	0.017	0.019
Ethnicity Ilocano			-0.297	-0.236	
Ethnicity Ifugao			-0.613	-0.500	
Municipality of origin 0			0.532	0.528	
Municipality of origin 1			-0.242	-0.247	
Level 3					
Municipality of origin village				1.253	1.198
Ethnicity Ilocano village				-0.380	-0.414
Ethnicity Ifugao village				-0.055	0.114
Random effects					
Level 2					
var (R_{0jk})	0.003	0.006	0.003	0.004	0.006
ρ_R	0.001	0.002	0.001	0.001	0.002
Level 3					
var (U_{00k})	0.107	0.546**	0.724**	0.637***	0.487**
ρ_U	0.031	0.142	0.180	0.162	0.129
ROC	0.694	0.906	0.909	0.908	0.903

** p<0.05, ** p<0.01, ***p<0.001*

fi eld level. The exploratory procedure applied has revealed at which levels important v ables explain land use decisions. Model 3 and 4 show that within these models a signifi part of variability is left at the village level, which is left unexplained in these mod The empty model with banana as dependent variable did not show any signifi cant variance component. However, aft er introducing the variables slope and transportation costs as explanatory variables (Table 4.3, model 2) the village level variance component signifi cant *p<0.01*). The variability at village level could not be accounted for by any of t explanatory variables used in this study. The question what causes the village diff erenc in this model will therefore remain unanswered. The village level variability might be caused by diff erences in soils or geomorphology, though these variables were not inclu in this study. The results for both the corn and banana models indicate that a conventio regression model would not be correct because the residuals would be heteroskedastic. The multilevel structure accounts for the unobserved eff ects between villages and provi a statistically correct model

The multilevel analysis of the land use system in the study area provided additional information to previous analyses based on conventional regression models (Overmars and Verburg, 2005 (Chapter 2)). In the case of corn the analysis confirmed the hypothesis that the household level plays an important role. In this analysis the municipality of origin of the household (which is proxy for migration history) in combination with the ethnicity variables turned out to be a significant explanatory variables that account for the variability at the household level. For the case of banana the analysis confirmed the idea that household level characteristics did not play an important role in the decision to cultivate bananas. Bananas occur mostly on sites that are less productive or too steep for arable crops like corn, rice or vegetables. Significant village level variance indicates the importance of village level conditions in explaining the decision to cultivate banana. The ROC values of the analyses of Chapter 2 and this analysis cannot be compared straightforwardly because the multilevel analysis incorporates a random part that contributes to the value of the ROC without actually explaining the dependent variable since this random part is fitted.
Like in any other statistical analysis, drawing conclusions about the causality of the relations from the regression analysis should be done with care. For example, the positive relation between slope of a field and bananas results from the fact that the flatter areas are devoted to arable crops, not because bananas perform better on the steep slopes. Like any other regression analysis multilevel models can only reveal associations between variables and partition variance. Additional research is needed to study the causality of the relations. An example of such a method for the case study area is described in Overmars *et al.* (2006) (Chapter 3).
In this chapter random slopes were not incorporated in any of the models. Although there were no strong arguments suggesting that the coefficients for the explanatory variables were different, some experiments were carried out to study the behaviour of models that include random slopes. This resulted in either insignificant random slopes or models that did not converge. Most likely the data structure and the amount of observations made the estimation of the random slopes complicated. The number of observations (fields) per household is low and this complicates the determination of the random slopes.

The results indicate that the household level can be crucial in explaining land use at the field level. However, in many studies household level data are not available because in many regional studies the analysis is based on remote sensing, maps and census data (e.g. Nelson *et al.*, 2001; Walsh *et al.*, 2001; Müller and Zeller, 2002). As this study shows simply substituting household level variables with their village level equivalents, which can be calculated from widely available census data, will most often not account for the household level variability because of errors due to aggregation. Disaggregating higher level variables to the level of analysis can lead to erroneous conclusions. In any case, disregarding the household level variables while explaining land use at field level ignores the conclusion of Rindfuss *et al.* (2003) that the household level is the central level to be included in explanations of land use
Data availability and data structure play an important role in land use studies. As illustrated in this chapter, data availability determines at what level land use can be studied, and therefore at what level one can draw conclusions. If the hierarchical structure of the data is important to the land use system under study and the research questions that arise from this, this structure should be considered in the sample design to take full advantage of the multilevel modelling technique. Ideally, at every level a sample is drawn that is

representative for the population at that level. For the highest level, one should keep in mind that a small sample size cause the same diffi culties as an ordinary regression with that sample size (Snij ders and Bosker, 1999) *i.e.* small sample sizes have less power than larger samples. For the lowest level, which is the unit of analysis, the number of observations per group (e.g. the number of fi elds per household) should be enough to estimate parameters that are included in the model

Datasets that were not designed for multilevel modelling oft en appear to be inadequate This is a serious constraint for applying multilevel modelling in land use studies, because many studies use available datasets. In studies with levels other than farmers and fi elds for example including country and sub-country level, the data structure can be more favourable to multilevel modelling

In the dataset used in this study the number of observations (fi elds) per household was very low, but this is inherent to the structure of the land use system, because the farmers have only a few fi elds. At the village level only 12 observations were present, but this is the complete population in the study area (*i.e.* one village was kept out of the analysis due to missing data). This data structure provides relatively few degrees of freedom for multilevel modelling and may have hampered the estimation of random slopes, which were therefore not included in the models presented. Polsky and Easterling (2001) have similar experience in estimating a multilevel model based on 446 counties nested within districts. To deal with small sample sizes one might consider to use bootstrap or MCMC (Markov Chain Monte Carlo) approaches, which are available in MLwiN (Rasbash *et al*., 2000), for example

Verburg *et al.* (2004d) emphasise the importance of multi-scale approaches and cross-scale dynamics and name multilevel modelling as a potential approach that can deal with scale issues in land use studies. Multilevel modelling can address a variety of these issues. First of all, the multilevel approach explicitly includes diff erent levels. These levels can be, for example, organisational levels of the land use system or nested administrative units, but can also be artifi cial aggregations of a grid. Where in other studies the eff ects of scale on the observed relations between land use and driving factors were studied by the separate analysis at diff erent organisational levels or by (dis)aggregating grids to one level of analysis (e.g. Verburg and Chen, 2000; Walsh *et al*., 2001; Overmars and Verburg, 2005 (Chapter 2)), the multilevel approach is capable of incorporating diff erent levels of aggregation within one model and exploring the contributions of the various levels

Secondly, within the multilevel approach cross-scale dynamics can be modelled as cross-level interactions. A cross-level interaction can be defi ned as dependence of a relation between two micro variables on a macro-level variable (Snij ders en Bosker, 1999). A difference with conventional models is that when including the cross-level interaction the slope parameters also have a random eff ect. An additional option in a multilevel approach is to include group level aggregates of variables. This clearly separates level 1 eff ects from higher level eff ects, which can be completely diff erent

Another important aspect to consider in land use studies is spatial dependency, which refers to the geographic law that nearby things are more related then distant things (Tobler 1970). Spatial dependency in land use patt erns can be caused by dependence of the land use patt ern on an explanatory factor that is spatially structured (trend) or a spatial interaction process of the land use variable itself, like competition or imitation (Anselin, 1988; Irwin and Geoghegan, 2001; Overmars *et al*., 2003; Polsky, 2004). Both Polsky and Easterling

(2001) and Pan and Bilsborrow (2005) mention that multilevel modelling can partly redu
the eff ect of spatial autocorrelation when neighbouring observations are nested within
group. If the spatial dependency is only related to the nested hierarchy this might even
correct for all spatial autocorrelation. However, oft en spatial dependency is structured
ferently than the nested hierarchy of the dataset. In this case the neighbourhood eff ects
be incorporated in the multilevel model as cross random eff ects (where lower level obse
tion can be member of diff erent groups at the higher level). For example, each observati
can be part of a group with all its neighbours. This approach would correct for spatial
autocorrelation but is not yet studied in land use research. In this study this approach w
not applied because the observed fi elds are relatively far apart due to the relatively sma
sample size and spatial autocorrelation is therefore assumed to be minima

4.5.2 Conclusions

The case study has shown that multilevel analysis can be applied statistically to model
the occurrence of land use. We consider multilevel modelling to be a relevant tool for la
use studies because organisational levels and spatial and temporal scale dependencies a
characteristic for land use data. Multilevel modelling off ers a method to study the infl u
of these levels and scales as well as great fl exibility in testing hypothesis on explanatory
variables and their cross-level interactions and spatial dependencies. Multilevel regressi
modelling is considered to be a statistically sound method to create regression models
when analysing hierarchically structured data. Including random parts in the model en-
sures correct estimates of the regression parameters and their signifi cance levels. Howe
so far, few scholars have applied this approach in land use studies. This might have to d
with data quality and data availability. Another cause can be that the methodology is or
recently developed. Currently, multilevel soft ware is becoming more generally availabl
(see Centre for Multilevel Modelling (2005) for a detailed review) which might promote
use of multilevel models in land use change studie
In recent LUCC literature many have advocated for explicit att ention for scale issues in
LUCC research (e.g. McConnell and Moran, 2001; Veldkamp and Lambin, 2001; Rindfus
et al., 2004). From this study it can be concluded that it is indeed important explicitly to
identify and report on the levels that are present in the study. Levels that are crucial in
explaining the land use system should be included in modelling exercises. Moreover, th
propositions that are studied should indicate more explicitly to which scales and levels
they apply. Potentially, a multitude of propositions can be formulated that involve scale
and level, like micro-micro, macro-macro, micro-macro, macro-micro and multi-level
propositions. To be able to test these hypotheses it is important to collect adequate data
to enable the application of a multilevel approach in order to answer questions that are
inherently hierarchical in reference to land use studies. Multilevel modelling is a useful
addition to the land use research toolbox that allows the exploration of a number of cros
scale propositions