ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Land Use Policy ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Land Use Policy

www.elsevier.com/locate/landusepol

Analyzing spatial hierarchies in remotely sensed data: Insights from a multilevel model of tropical deforestation

Colin Vance[a,*], Rich Iovanna[b]

[a]German Aerospace Center, Institute of Transport Research, Rutherfordstrasse 2, 12489 Berlin, Germany
[b]National Center for Environmental Economics, US Environmental Protection Agency, 1200 Pennsylvania Ave., NW (1809 T), Washington, DC 20460, USA

Received 5 July 2004; received in revised form 6 January 2005; accepted 17 February 2005

Abstract

This paper advances an empirical model assessing how changing economic and ecological conditions at different spatial scales affect land conversion decisions. We apply a multilevel econometric model to explore the implications for parameter estimates and their standard errors of ignoring hierarchical groupings in the data. The paper draws on a panel of agricultural-household data collected from a survey of Mexican farmers. A comparison of results obtained from a standard single level model reveals several stark distinctions in the estimated effects, some of which have immediate relevance for conservation policy. We conclude that the multilevel specification is warranted for alleviating issues associated with error structures inherent to spatial data.

Keywords: Tropical deforestation; Agricultural households; Multilevel models; Satellite data; Mexico

Introduction

Since the domestication of agriculture some 10,000 years ago, human-induced changes in land use and cover have played an increasingly important role in driving environmental change. The conversion and modification of the Earth's surface for agricultural and other uses has, of course, always affected local environs, but in the last several decades these effects have assumed global proportions, accounting for approximately 25% of human-caused emissions of heat-trapping gases (Houghton, 1994). Tropical deforestation is a particularly dramatic form of land cover change owing to the importance of these forests in regulating the biogeochemical states and processes that effect global climate. With upwards of a third of the earth's six billion inhabitants depending directly on tropical resources for many economic and environmental goods and services, deforestation also poses more immediate regional and local concerns, including soil erosion, habitat degradation, and increased frequency and severity of floods.

During the past two decades, growing scientific and public concern about these and other threats to global habitats has elicited increased efforts to model the exogenous drivers and associated land uses underlying tropical forest clearance in a variety of settings. Many of these efforts are based on econometric analysis of data obtained from surveys of semi-subsistence farm households, a form of economic organization that is commonly identified as an important proximate cause of tropical deforestation (e.g. Jones et al., 1995; Pichón, 1997; Walker et al., 2000; Vance and Geoghegan, 2004). While providing a rich source of information, such data is often collected in remote regions under conditions that pose challenges for ensuring a random, representative sample of observations. Whether highly complex or ad hoc survey designs are applied, the sample of

*Corresponding author. Tel.: +49 30 67055 147; fax: +49 30 67055 202.
E-mail addresses: colin.vance@dlr.de (C. Vance), iovanna.rich@epa.gov (R. Iovanna).

0264-8377/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.landusepol.2005.02.002

observations selected for analysis generally consists of clusters of households distributed across communities, which themselves may be embedded in higher-order administrative units or ecological zones.

The implications of this hierarchical structuring of the data for model estimation are potentially profound, but rarely acknowledged in the literature on land-use change. Clustered data may result in high sampling error in the estimated parameters resulting from the dependence of observations within the same cluster. For example, households located in the same village probably share more in common than households located in different villages. If there are correlated but unobserved characteristics across observations in a cluster, one of the fundamental assumptions of econometric analysis—that of uncorrelated disturbances—will be violated. The consequences of this violation are inefficient parameter estimates and downward-biased estimates of the standard errors, the latter of which can result in falsely ascribing statistical significance to effects that are due to chance. Despite their potential to seriously undermine the results of an analysis, these consequences are typically neglected, a practice that extends beyond the literature on deforestation to include the larger corpus of studies that analyze the economic behavior of agricultural households more generally.

The present paper analyzes the determinants of deforestation in an agricultural frontier in southern Mexico by drawing on a hierarchically structured database that integrates household survey data with Thematic Mapper satellite imagery. While a small but growing number of studies have used the linkage of remotely sensed imagery with household surveys to analyze land-use change (e.g. McCracken et al., 1999; Walker et al., 2000; Vance and Geoghegan, 2002), this paper moves beyond the existing literature through the use of multilevel modeling techniques as a way of avoiding the problems of biased estimation associated with autocorrelation. Multilevel models provide a method for systematically capturing the influence of hierarchical structures in the data that may emerge from either socially or environmentally determined groupings at various spatial scales. Compared with standard regression techniques, multilevel models offer four principle advantages: they correct for biases in the estimated parameters' standard errors resulting from clustering; they attenuate omitted variable bias by accounting for heterogeneity at the cluster level; they provide the ability to decompose the total variance in the dependent variable into portions associated with each level of the data; and, lastly, because of the treatment of higher level units as random samples from the population, the results from a multilevel regression can be generalized beyond the particular groups in the study (Bryk and Raudenbush, 1992; Goldstein, 1995; Guo and Zhao, 2000; Goulias, 2003). To assess these advantages, this paper estimates a binary multilevel model of forest clearance over a roughly 10-year period using the complementary log–log link. A comparison of these with results obtained from a standard single level complementary log–log model reveals several stark distinctions in the sign, significance and magnitude of effects, suggesting that the more flexible multilevel specification is warranted.

The paper begins with a brief overview of the study region and a description of the survey methods and data sources. Section 3 provides some background on multilevel modeling and discusses the model specification used in this research. Section 4 catalogs the estimation results, and a concluding section summarizes and interprets the findings.

Survey methods and data sources

Data collection for this study proceeded over an eleven-month period in 1996/97 in a roughly 22,000 km^2 agricultural frontier located in southern Mexico (Fig. 1). Spanning the states of Campeche and Quintana Roo across the base of the Yucatán peninsula, the study site is populated primarily by agricultural households, the majority of which are members of a communal form of land tenure referred to as the ejido.[1] Following the construction of a highway across the center of the frontier in 1972, the government extended a series of ejidal land grants to groups of petitioning farmers from other, primarily neighboring, regions of the country. While the region had been the site of isolated logging activities for much of the 20th century, the road and colonization policy instigated the first major population influx in its modern history. A prolonged period of deforestation ensued, with forest cut at a rate of between 0.32% and 0.39% per year between 1969 and 1997 (Turner et al., 2001). These land-use change dynamics have been captured by a time series of Thematic Mapper satellite imagery, the classified pixels from which serve as the dependent variable modeled in this study.

The independent variables used in the analysis were obtained from a field survey that focused specifically on the ejido sector. Data collection proceeded on the basis of a stratified two-stage cluster design. In the first stage, the region was geographically divided into several strata based on proximity to the highway, with one ejido

[1]The ejido sector was created by the land reform that followed the Mexican Revolution of 1910 as part of a national agenda to redistribute land to the rural poor. Article 27 of the Mexican constitution provided the legal sanction for this redistribution by vesting the state with the authority to re-appropriate land from the large haciendas. In 1992, amendments to the constitution were promulgated that, in addition to terminating the continuation of ejido land grants, allowed ejido members to privatize and sell their lands. At the time of this survey, no ejidos in the sample had privatized.

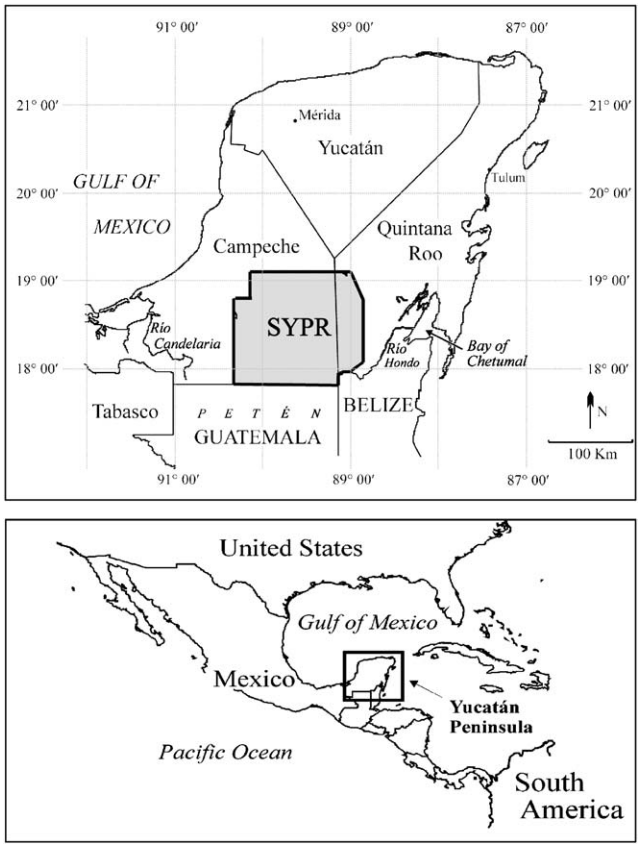

Fig. 1. The southern Yucatán Peninsular region.

selected from each stratum. In selecting the ejido, the principle of probabilities proportional to size was applied, whereby each ejido was assigned a chance of selection that was given by the ratio of its population to the population of the strata. In the second stage, the survey respondents themselves were chosen from each ejido via simple random sampling techniques after an enumeration of households was created. The number of respondents selected from each ejido was based on the application of a compensating selection rate to ensure that the stratum's representation in the sample was roughly proportional to its representation in the overall population (Warwick and Luinger 1975, pp. 104–115). It bears noting that although ejidos are managed communally and held in usufruct, ejido members in this region of Mexico generally maintain exclusive access to particular parcels of land throughout their residence. Thus, the selection of ejidatario respondents afforded the linkage of specific household data with specific land parcels. The final sample used for model estimation comprises the parcels of 135 households across ten ejidos.

The data collection protocol divided each household survey into two stages. In the first stage, an interview was conducted at the home during which socioeconomic data on demographics, farm capital, credit, purchased inputs, off-farm employment, yields, and sales of agricultural produce was collected and recorded onto a standardized questionnaire. The questions primarily elicited information on the household's current circumstances, but information was also collected on ownership of farm capital (e.g. vehicles and chain saws) for years extending back to 1986. By collecting information on the births, deaths, and out-migration of any sons or daughters of the head, it was additionally possible to reconstruct—at least for biological family members—the household's age and gender composition through time.

Table 1
Dates of the satellite imagery

Zone I	Zones II & III	Zone IV
Nov. 11, 1984	Apr. 01, 1987	Jan. 14, 1985
Feb. 21, 1993	Oct. 29, 1994	Nov. 07, 1994
Jan. 31, 1997	Feb. 05, 1996	Jan. 31, 1997

The second stage involved a guided tour of the farmer's plot. Using a geographic positioning system (GPS), the interviewer created a geo-referenced sketch map detailing the distinct fields within the plot as the farmer-respondent provided an interpretation of the contemporary and past use of these fields. An attempt was made to record several GPS points throughout the plot, particularly along the borders. Following the interview, these points were plotted onto a backdrop of the most recently available satellite imagery, making it possible to digitize the boundaries of the plot. The digitized plot borders were then extracted and superimposed on available images from prior years, thereby yielding a longitudinal database of land-use change associated with individual farm households.[2]

To construct the dependent variable, satellite images from four contiguous zones over three dates for each zone were classified into seven land-use classes: mature lowland forest, mature upland forest, secondary forest (7–15 years old), agriculture, fern, inundated savannah, and water. The dates and zones of the satellite imagery are presented in Table 1. As the goal of this paper is specifically to model the determinants of forest clearance, a binary dependent variable was defined that assumes a value of 1 if an individual pixel underwent a transition from mature upland or lowland forest to agriculture over an interval and 0 otherwise. Only those pixels classified as mature forest (older than 15 years) at the earliest date of the imagery were included in the analysis. While other work has explored a broader definition of deforestation that includes the clearance of vegetation over 7 years of age (Vance and Geoghegan,

[2]In order to maintain the direct link between the household and the parcel, satellite data from past years was only used if the household reported having access to the land as of the date of the imagery.

2002), we use the narrower definition of 15 years in order to purge the dependent variable of fallow cycle dynamics to the extent allowed by the data.

The complete dataset has a nested, three-tiered structure, with each level having its own suite of variables. Level 1—the unit of analysis—consists of a sample of 87,945 satellite pixels having dimensions of roughly 30×30 m. Explanatory variables measured at this level include soil quality, elevation, and slope. Each pixel is associated with a parcel, which defines the level two unit of analysis. The dataset contains a total of 135 parcels, corresponding to each farmer interviewed, with an average of 777 pixels per parcel. Variables measured at level 2 consist primarily of the socioeconomic characteristics of the household managing the parcel, including farm capital, demographic composition and access to government farm support. Finally, level 3 is defined by the 10 ejidos within which each parcel is located, with an average of 13.5 parcels per ejido. One variable is measured at level 3, the ejido's population. A multilevel approach is thus ideally suited to this data structure, as it enables capturing the effects of correlation within the levels that could otherwise undermine the validity of traditional regression techniques.

Multilevel modeling of land use

Changes in land use result from ecological and socioeconomic determinants that operate across multiple spatial, temporal, and political-hierarchical scales (Geoghegan et al., 1998; Walker and Solecki, 2001; Klepeis and Vance, 2003). Although it is generally recognized that natural and social systems are "functionally or operationally layered in their relationships with one another" (IGBP-HDP, 1995, p. 44), disentangling these relationships in land-use models has proved a vexing empirical challenge. The spatial econometrics literature has made considerable headway in accounting for what Anselin (1988, p. 7) has described as the "peculiarities caused by space in the statistical analysis of regional science models" (cited in Anselin and Florax, 1995). These peculiarities may emerge from either the existence of functional relationships between the values of observations in space or from instability of the model parameters across locations, conditions referred to as spatial dependence and spatial heterogeneity, respectively (Irwin and Geoghegan, 2001). Various methods have been developed for handling these issues, including the construction of spatial weight matrices, the inclusion of spatially lagged variables as regressors, and the specification of coefficients that are allowed to vary over space. Despite an increasing sophistication of techniques, however, higher-order effects are still often handled in an ad hoc way, typically through the inclusion of regional dummy variables. The development of a "truly hierarchical approach in both observation and explanation of the processes of land-use change" remains a largely uncharted frontier in the development of regional models (IGBP-HDP, 1995, p. 45).

Multilevel modeling techniques can contribute toward this development by offering a systematic means of controlling for and quantifying spatial effects in model estimation. Although multilevel models can be regarded as complementary—rather than alternative—to the set of tools used in spatial econometrics, their application in land-use studies is rare. This is despite the fact that these models are particularly well suited to analyzing data with a spatially organized hierarchical structure. In the present application, the spatial clustering of pixels within parcels and parcels within ejidos creates the potential for the data to be characterized by spatial dependence, spatial heterogeneity, or both. The specification of a multilevel model makes it possible to handle each of these features, the former through the specification of random intercepts at each level of the data and the latter through the specification of random covariate effects across spatially organized units. Goulias (2003) provides an introductory review of multilevel models for continuous data and Guo and Zhao (2000) provide a review of multilevel models for binary data, with a particular focus on the logit link.

The multilevel model estimated in this research uses the complementary log–log link, which is conceptually similar to the logit but more appropriate for the data at hand. Noting that the dependent variable assumes a value of 1 if a pixel is converted between two dates and zero otherwise, the standard complementary log–log model is expressed as

$$\log[-\log(1 - P_{it})] = \beta_0 + \beta_1 X_{it1} + \cdots + \beta_z X_{itz}, \qquad (1)$$

where P_{it} is the probability that pixel i is converted in interval t given that the pixel was not converted in an earlier interval, β are the parameters to be estimated 1 through z including the constant term β_0, and the X_{it} are exogenous covariates. By deriving the model from a latent response reflecting the unobserved utility from conversion, it can be conceived as having a level one random error, ε_{it}, which is assumed to follow a standard extreme value distribution (with variance set equal to $\pi^2/6$). As with the logit function, the complementary log–log transforms the modeled probability to range between minus infinity and plus infinity. Unlike the logit, the distribution function is not symmetric around zero but skewed to the right, making the model particularly applicable when the probability of an event is very large, or, as in the case of the present data, very small. An additional advantage of the model is that the β coefficients have a relative risk interpretation, as in Cox's proportional hazards model, thereby reconciling the temporal continuity of the conversion process being

modeled with the coarseness in the measurement of timing (Allison, 1995; Hosmer and Lemeshow, 1999).[3]

Eq. (1) is a single-level regression model, the binary variants of which are ubiquitous in the agricultural and land-use literatures. Underlying the model is the assumption that the co-variances among all the n observations are zero. Imposing this restriction would be reasonable were the data to have been drawn from a simple random sample having no likely sources of autocorrelation. Pixels located in the same vicinity and parcels located in the same village can be expected to be more alike than a random sample, however, in which case important relationships may be neglected by specifying a single-level model.

A multilevel structure can be introduced to Eq. (1) by expanding the constant term to include a random effect among parcels, denoted by the subscript j. This randomness at level two can be reflected in the model as

$$\log[-\log(1-P_{ij})] = \beta_{0j} + \beta_1 X_{it1} + \cdots + \beta_z X_{itz},$$
$$\beta_{0j} = \beta_0 + u_{0j},$$
$$[u_{0j}] \sim N(0, \Omega_u) : \Omega = [\sigma^2_{u0}]. \quad (2)$$

In the above system, the constant β_{0j} has a mean of β_0 and a variation around this mean among parcels depicted by u_{0j}. The term u_{0j} is a random error component that captures variation across parcels in their intercepts. It is assumed to be normally distributed with $E(u_{0j}) = 0$, and with $\mathrm{Var}(u_{0j}) = \sigma^2_{u0}$ to be estimated.

This variation in the intercepts could alternatively be captured by the inclusion of parcel dummies as exogenous regressors, but there are several possible drawbacks to such an approach: First, if the number of parcels is large relative to the number of observations, then the associated loss of degrees of freedom will compromise the model's efficiency, possibly to the point of precluding their inclusion. Second, those parcels that are relatively small in size and have few associated pixels will have poorly estimated effects. Third, specifying parcel level effects as random rather than as dummies enables extrapolating the results to the larger population of parcels rather than just those included in the model. Last, in contrast to the difficulty in interpreting a slew of fixed effects, a random effects treatment also provides a ready means by which to assess the overall influence of parcels on deforestation, i.e., by decomposing the variance (see below).

The model can be elaborated in a number of directions, including through the introduction of higher levels of nesting and by allowing the β-coefficients to vary. For example, a two-level model that allows random variation over parcels and that additionally allows the coefficient on the covariate X_1 to vary over parcels would be expressed as:

$$\log[-\log(1-P_{ij})] = \beta_{0j} + \beta_{1j} X_{it1} + \cdots + \beta_z X_{itz},$$
$$\beta_{0j} = \beta_0 + u_{0j},$$
$$\beta_{1j} = \beta_1 + u_{1j},$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_j = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u10} & \sigma^2_{u1} \end{bmatrix}. \quad (3)$$

In addition to incorporating random variation in the intercept among parcels, random variation in the slope coefficient estimate β_{1j} is captured by the term u_{1j} in the above model. The variances for the error terms are represented by σ^2_{u0} and σ^2_{u1}, while the covariance of the slope and intercept at the parcel level is represented by σ_{u10}. By referencing the standard errors of these variance estimates, significance tests for behavioral heterogeneity can be conducted for each level of the data. Moreover, random coefficients can be specified for other variables measured at the different levels of the data, which will introduce additional variance and covariance terms that reflect interactions across levels. The multilevel framework can thereby relax several potentially onerous restrictions imposed by single-level methods through the capturing of interdependencies that may exist both within and between levels.

Models of the form of Eqs. (2) and (3) can in principle be estimated using maximum likelihood techniques, but the procedure is computationally intensive since the random effects necessitate numerical integration over a high dimensional integral. A limited number of alternative approximation procedures have been developed, the most prevalent of which are marginal quasi likelihood and penalized quasi likelihood (Guo and Zhao, 2000). Both techniques rely on either a first or second order Taylor expansion to linearize the model, estimating it with an Iterative Generalized Least Squares approach that alternately estimates the fixed and random parameters until convergence is achieved. We explored the application of these techniques using the specialized multilevel modeling software, MLWIN, but encountered convergence problems particularly when specifying models with random coefficients. We therefore opted to employ an alternative method using SAS' GLIMMIX macro. As noted by Guo and Zhao (2000, p. 451), this procedure uses a pseudo likelihood developed by Wolfinger and O'Connell (1993) that is similar to the penalized quasi likelihood. While one drawback of the technique is that computational demands typically require constraining the covariance of the slope and intercepts to equal zero, it has the advantage of taking into account extra-dispersion, a feature that could otherwise lead to unreliable estimates of the standard errors.

[3]When applied to interval data the logit also has a proportional hazards interpretation, but unlike the complementary log–log, it is more appropriate for modeling events that can only occur at discrete points in time. As an example of such events, Allison (1995) cites tenure decisions, reviews of which generally occur once a year.

Table 2
Explanatory variables used in the model[a]

Explanatory variable	Definition	Mean	Standard deviation	Static or time-varying
Upland soil	1 if pixel classified by soil superior for farming, 0 otherwise	0.75	0.43	Static
Slope	Slope of pixel in degrees	1.25	2.73	Static
Elevation	Elevation of pixel in meters	164.04	69.10	Static
Plot size	Number of pixels in plot (in 100s)	7.77	6.79	Static
Distance to parcel	Distance separating the household from the parcel in kilometers	7.31	6.84	Static
Household members > 11	Number of household members older than 11, averaged over the interval	1.82	1.80	Time-varying
Household members < 12	Number of household members younger than 12, averaged over the interval	3.36	1.70	Time-varying
Education of head	Years of schooling of household head	2.84	3.17	Static
Native Spanish speaker	1 if the mother tongue of the household head is Spanish, 0 otherwise	0.77	0.422	Static
Duration of occupancy	Years of occupancy in ejido	18.53	10.63	Time-varying
Vehicle	Percent of interval length (in years) owning vehicle	0.09	0.25	Time-varying
Chain saw	Percent of interval length (in years) owning chain saw	0.18	0.31	Time-varying
Farm support	Number of hectares registered for PROCAMPO support[b]	3.88	4.41	Time-varying
Ejido population	Population of the ejido (in 100s)	5.82	9.24	Time-varying

[a]The means presented here are measured at the level of the variable's aggregation (pixel, parcel/household, or ejido) rather than at the observational unit of modeling, the pixel.

[b]The PROCAMPO program was initiated in 1994 and extends an annual per hectare payment on a fixed area of land. In 1996, these payments were fixed in real terms at 464 pesos (US $64) per hectare until the program's expiration, scheduled for 2010. Farmer's must maintain land registered in the program under a 'productive' use in order to receive continued disbursements of the payment.

Results

The first model presented is a standard single-level model, which is used as a benchmark for assessing the results from the multilevel analysis. The specification is similar to that found in Vance and Geoghegan (2002), where a theoretical model of the land conversion decision is developed and a detailed accounting of hypothesized effects is presented. For present purposes, it suffices to note that a suite of time-varying and static covariates were selected that capture the socioeconomic and biophysical factors expected to affect the utility derived from the land in forested and cleared states and hence the hazard of forest conversion. The specification additionally includes time period dummy variables indicating the beginning of each interval of the satellite data to control for the fact that they are of differing lengths (Allison, 1995), as well as ejido dummy variables to control for fixed effects at the ejido level. Summary statistics and variable definitions are presented in Table 2. Column one of Table 3 presents the coefficient estimates for the single level, to be referred to as Model I.

Interpretation of the coefficient estimates generated from the complementary log–log model is complicated by the log–odds transformation of the dependent variable. The figures presented in the table are therefore the transformed coefficients in terms of "risk ratios." The risk ratio is interpreted as the percent change in the hazard rate from a unit increase in the covariate, which is obtained by subtracting one from e^{β} and multiplying the resulting value by 100 (Allison, 1995).

The results from the model suggest several statistically significant determinants of deforestation among agricultural households. Most of the variables are significant at the 5% level and have signs that are generally consistent with intuition. With regard to the ecological variables, improved soil quality is seen to increase the hazard of deforestation, while steeper slopes and higher elevations discourage conversion. The land endowment and travel distance from the household to the parcel also have negative effects on the hazard, both plausible findings to the extent that increased travel time and increased area for cultivation reduces the likelihood that any given plot within the parcel is deforested.[4] Human, physical, and financial capital, as measured by the number of household members over 11 years of age, the education of the household head, the duration of occupancy, vehicle ownership, and government farm support all have positive effects on the hazard, although in the case of the duration of occupancy the effect is nonlinear. Specifically, the hazard is seen to be initially decreasing with each year of occupancy but eventually increasing, with the transition occurring after roughly 26 years. The positive effect of education is somewhat unexpected, particularly given that higher education tends to increase wage-earning ability and hence the opportunity costs of farm related activities. Two other unexpected findings are the coefficients on the number of household members under age 12 and chain saw ownership, both of which are seen to be negative and

[4]A variable measuring the ejido's distance to the nearest market is not included in the model as this influence is controlled for by the ejido dummy variables.

Table 3
Model results

	Model I	Model II	Model III	Model IV
Sample size	87,945			
Fixed effect risk ratios (p-values)				
Upland soil	47.39		17.94	21.88
	(0.00)		(0.00)	(0.00)
Slope	−2.21		−3.42	−3.38
	(0.00)		(0.00)	(0.00)
Elevation	−1.10		−1.11	−1.14
	(0.00)		(0.00)	(0.00)
Plot size	−3.24		−3.74	−5.92
	(0.00)		(0.07)	(0.03)
Distance to parcel	−4.76		−5.44	−5.58
	(0.00)		(0.05)	(0.11)
Household members > 11	8.53		−1.06	13.43
	(0.00)		(0.63)	(0.00)
Household members < 12	−1.46		0.71	10.30
	(0.02)		(0.76)	(0.00)
Education of head	2.35		−3.41	6.66
	(0.00)		(0.46)	(0.25)
Native Spanish speaker	−24.31		5.88	−6.83
	(0.00)		(0.89)	(0.88)
Duration of occupancy	−5.24		−6.84	−3.02
	(0.00)		(0.03)	(0.42)
Duration of occupancy squared	0.10		0.20	0.16
	(0.00)		(0.00)	(0.01)
Vehicle	93.79		−49.16	−85.79
	(0.00)		(0.00)	(0.09)
Chain saw	−6.45		−34.85	93.40
	(0.10)		(0.00)	(0.40)
Farm support	3.05		0.081	−2.21
	(0.00)		(0.84)	(0.00)
Ejido population	−2.37		10.32	19.23
	(0.12)		(0.00)	(0.00)
Joint test chi square p-values				
Ejido dummies	0.00		0.00	0.00
Time period dummies	0.00		0.00	0.00
Occupancy duration quadratic	0.00		0.00	0.00
Random effects estimates (p-values)				
σ^2_{parcel}		4.17	2.34	2.78
		(0.00)	(0.00)	(0.00)
$\sigma^2_{chainsaw}$				18.98
				(0.00)
$\sigma^2_{vehicle}$				16.15
				(0.01)
Deviance	53,247	45,913	44,842	44,182

highly significant. With regard to the latter, Vance and Geoghegan (2002) attribute a similar result from the same data set to a shortcoming of the model to control for wealth effects, noting that those farmers who own chain saws tend to be more diversified in their incoming earning activities and less reliant on the natural resource base. The remaining measure of human capital, the dummy indicating Spanish as a first language, has a negative effect, suggesting that indigenous farmers, who may have less access to land-saving inputs, are more reliant on the deforestation of older tracts for cultivation. The only variable measured at the ejido level, ejido population, has a negative effect on the hazard of deforestation, though the coefficient estimate is out of the range of significance at the 10% level. Finally, both the ejido and time dummies are found to be jointly significant on the basis of χ^2 tests.

Our application of the multilevel model involved three specifications of a two-level model. We initially explored the specification of a three level model with pixels at level one, parcels at level two and ejidos at level three, but ultimately opted to incorporate ejido effects through the inclusion of dummy variables. This decision was informed by a recent simulation study by Maas and Hox (2004) suggesting a minimum of 50 groups to avoid biased statistical tests of the variance estimates at level two. We found that the substitution of ejido dummies for a variance component term had little effect on the magnitude of the coefficient estimates and no effect on the qualitative findings.

Column 2 presents estimates from a null model or fully unconditional model that includes only the intercept term and the variance components. The null model is useful for decomposing the variance in the data across levels before any account is taken of exogenous determinants. From this decomposition, it is possible to generate estimates of the intraclass correlation. When the model contains two levels, the intraclass correlation is calculated as

$$\rho_{parcel} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}, \tag{4}$$

where $\sigma^2_e = \pi^2/6 = 1.645$ is the variance of the standard complementary log–log distribution. The intraclass correlation can be interpreted as the portion of total variance explained by the grouping of the data or, alternatively, as the expected correlation between two randomly chosen elements in the same parcel in terms of the latent variable representing forest conversion (Guo and Zhao, 2000; Hox 2002). According to the estimates from Model II, the value for ρ_{parcel} is 0.72, suggesting that the bulk of the variation in the dependent variable is due to factors that play out at the parcel level.

Further insight into the distribution of the data can be gleaned from inspection of the parcel level shrunken residuals generated from Model II. Following the discussion in Rasbash et al. (2004), these residuals are calculated by multiplying the raw residual for parcel j, r_j,

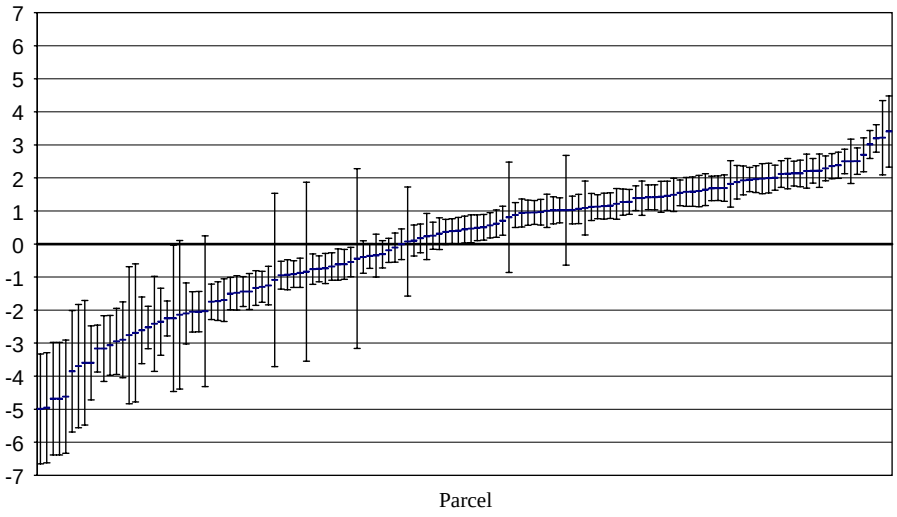

Fig. 2. Parcel level residuals.

by a shrinkage factor, written as

$$u_{0j} = \left[\frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e/n_j}\right] r_j, \tag{5}$$

where n_j is the number of pixels in parcel j, and the term in brackets, which is always less than one, is the shrinkage factor. This factor reflects the extent of uncertainty in the mean estimate of the parcel level effect. As this uncertainty increases, due to either a small number of pixels in the parcel or to a high degree of pixel variability, the shrinkage factor will approach zero, thereby pulling the estimate of the shrunken residual toward zero and the estimate of the parcel mean toward the center of the population. The shrunken residual thus represents a weighted combination of the estimated parcel mean and the estimated population mean, with the latter playing a stronger role as the shrinkage factor tends toward zero (Rasbash et al., 2004).

Fig. 2 presents a "caterpillar" plot in which the shrunken residuals are arranged by magnitude. The bars represent the associated 95% confidence intervals, which are obtained by multiplying the standard error of the residual by 1.96. The residual plots enable testing for significant deviations from the average, level-specific intercept predicted by the model. Thus, in Fig. 2 it can be seen that the majority of the parcels differ significantly from the overall average as evidenced by the fact that their error bars extend either entirely below or above the horizontal line at zero.[5]

As with Model II, Model III incorporates random effects at the parcel level and additionally includes the same set of covariates as in Model I. To gauge the model's improved explanatory power, we can conduct a likelihood ratio test by computing the difference in the deviance statistic relative to Model II. As the two models are nested, this difference has an approximate χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters (Guo and Zhao, 2000). The calculated χ^2 is 1,071 ($df = 22$; $p < 0.0001$), providing clear evidence for a significant improvement in fit.

In comparing the fixed effects estimates from Model III with those of Model I, several distinctions emerge with respect to the significance, signs, and magnitude of the coefficients. Five of the variables identified as statistically significant in Model I—the number of adults, the number of children, education, the Spanish language dummy, and government farm support—are insignificant in Model III. The most striking discrepancies, however, are the sign shifts on the variables vehicle ownership and ejido population, which in Model III are statistically significant and negative and positive, respectively. The former finding comports with the observation that owners of vehicles have better access to labor markets and a wider range of options for diversifying income earning activities through, for example, the transport of farm produce. That population pressure increases deforestation is also consistent with intuition, particularly in an area where land managers may have an incentive to deforest as a means of solidifying usufruct-based tenure arrangements in the face of population growth. In addition to the sign reversals, there are also some notable differences in the magnitudes of the coefficient estimates. For example, the estimate on the variable soil quality in Model I would lead to the conclusion that upland soils have a 47% higher hazard of deforestation than lowland soils.

[5]Figure 2 as presented does not enable testing for significant pairwise parcel differences, as this would require multiplying the standard error by 1.4 instead of 1.96 (see Goldstein and Healy, 1995 for further details).

The corresponding estimate in Model III is 18%, still statistically significant but less than half the magnitude of the estimate in Model I. Finally, we note that Model III retains the negative coefficient for chain saw ownership, though the magnitude of the effect increases substantially.

Further work was conducted to explore whether the estimated coefficients on the covariates varied significantly by parcel. Several specifications were tested, many of which registered no significant improvement in fit or failed to converge. We were particularly interested in pursuing the counter-intuitive sign on chain saw ownership and the sign reversal on vehicle ownership between Models I and III. The final column of the table presents a model in which both these variables are specified as random effects at the parcel level. This increased flexibility is seen to both significantly improve the fit of Model IV over Model III, as indicated by the reduction in the deviance statistic ($\chi^2 = 660$; $df = 2$; $p < 0.0001$), and to affect the estimates on several of the other coefficients.[6]

While the vehicle ownership fixed effect retains the negative and significant coefficient, the coefficient for chain saw ownership flips to positive, though it is insignificant. Interestingly, the random parameters of both variables are highly significant. With respect to the influence of chain saw ownership, this suggests a high degree of heterogeneity among parcel managers, the behaviors of whom evidently have a counter-veiling effect on the sign, and hence significance, of the coefficient estimate. Contingent on the normality assumption of the random component, we can use the fixed effect estimate for the variable and the corresponding standard deviation of the random effect to make a rough calculation of the percent of respondents for which the effect is positive and negative. Referring to the standard normal distribution, this breakdown is approximately 56–44%, respectively, for the chain saw variable, and 31–69% for the vehicle variable. Thus, the result from Model IV suggests that for over half of the parcel managers in the sample, the effect of chain saw ownership on deforestation is positive, a markedly different conclusion from that obtained by the other models.

Also of note in Model IV is that both demographic variables are positive and highly significant, in contrast to Model I, in which the number of members under 12 is negative and significant, and to Model III, in which both variables are insignificant. The positive coefficients of Model IV are in agreement with other studies in the region (Geoghegan et al., 2001; Vance and Geoghegan, 2004; Vance, 2004), which have uncovered evidence for—at best—partial engagement of households in labor and output markets and consequent significant effects of household demographic composition in land-use decisions. To the extent that households maintain a strong focus on consumption production, this finding suggests that traditional policy interventions to influence production behavior (e.g. price supports, taxes and subsidies) may elicit a sluggish, if not perverse, supply response (Singh et al., 1986; de Janvry et al., 1991; Medellin et al., 1994).

A final notable result of policy relevance is the negative and statistically significant effect of farm support, which is positive and significant in Model I and insignificant in Model III. The finding of Model IV in this case counters much of the regional evidence gathered to date on the program through which the support is administered, referred to as PROCAMPO. Vance and Geoghegan (2002) and Klepeis and Vance (2003) present results suggesting a positive effect of PROCAMPO on forest loss. In interpreting this finding, these authors argue that the program, by stipulating the continuous cultivation of a fixed area of land for receiving the per hectare payment, effectively removes that land from the fallow cycle and hence places increased pressure on remaining forested tracts to maintain yields. The result presented in Model IV, by contrast, appears to support the position found in governmental literature, which states that the program encourages intensification and thereby aims to "slow down environmental degradation, promote conservation and reforestation to help reduce soil erosion and water pollution caused by excessive use of non-organic pesticides, and to promote sustainable development" (SARH, 1993, cited in Klepeis and Vance, 2003). Specifically, the result suggests that each hectare registered in the program decreases the hazard of deforestation by 2.21%.

Concluding discussion

The results of the foregoing analysis reveal marked differences between the single and multilevel models, and to a lesser extent, between the two multilevel specifications themselves. Given that the models specify an identical array of explanatory variables, it is reasonable to conclude that the identified incongruities can be largely attributed to differences in the treatment of unobserved heterogeneity. As elaborated by Vance and Geoghegan (2002) in their initial analysis of the data, the single-level model estimated in theirs and the present

[6]It is noted that the random intercept coefficient at level two increases substantially in Model IV. As discussed in Fielding (2004), interpretation of changes in intercept variances is difficult for the case of generalized linear models. With the addition of new covariates to such models, there is an implicit scale change as the model is re-standardized to maintain the level 1 variance (equal to $\pi^2/6$ in the case of the complementary log–log). To the extent that changes in the residual variances result from these scale changes, it is not possible to observe directly the results of real changes in level 1 variation.

paper essentially relies on the use of ejido dummies as fixed effects to control for unobservable characteristics that are correlated across space within ejido boundaries. This approach clearly has a serious limitation insofar as it fails to control for the effects of unobservable variables at the parcel level. The multilevel model, by contrast, explicitly models the manner in which pixels are clustered in parcels.

Controlling for this clustering has several consequences: The standard errors are in many cases more conservative—i.e. larger—because the model no longer assumes complete independence of observations. This recognition of the reduction in the effective sample size means that the multilevel model is not prone to the single-level model's elevated probabilities of a Type-I error (Snijders and Bosker, 1999; Polsky and Easterling, 2001) and, in the case of the variables education and the Spanish language dummy, accounts for why the coefficient estimate is now statistically insignificant. Pinpointing the source of the sign reversals in vehicle ownership, ejido population, the number of household members under 12 and government farm support is less straightforward, but here again it is plausible that the explanation lies in the accounting of unobserved heterogeneity. One possible source of this heterogeneity is spatial autocorrelation at the level of the parcel, which could induce omitted variable bias as a result of unobserved household level characteristics. Whether the bias is positive or negative will largely depend on both the effect of these characteristics on the dependent variable and on their correlation with the included independent variables. Vehicle ownership, for example, is likely to be positively correlated with other forms of unobserved farm capital (e.g. managerial talent), which themselves may positively affect the hazard of deforestation on net. The overall impact of these factors would thus be to induce upward bias on the vehicle variable, which would account for the positive coefficient in Model I. To the extent that the parcel-level error term in models III and IV control for these factors, this source of bias is attenuated, resulting in a negative coefficient estimate. Modeling of this same variable and chain saw as random effects pointed to another possible source of spurious results arising from the single-level model's restrictive assumption of homogeneous preferences among farmers. In this regard, the results from Model IV strongly supported what was observed anecdotally during the field survey: that farmers having access to such capital relied to varying degrees on crop cultivation for their livelihoods.

In conclusion, we find that the exploitation of hierarchies in the data enables the estimation of more flexible models that provide unique insights into both contextual effects through the specification of random intercept terms, and behavioral heterogeneity through the specification of random coefficients. Given the prevalence of hierarchically structured data in land-use change science, particularly in studies using household level data collected through the application clustered survey designs, the estimation of multilevel models to control for autocorrelation would seem warranted, at least as a check for robustness. As is demonstrated by the results presented here, the simplifying assumptions of single-level techniques can otherwise lead to highly inaccurate inferences regarding the effects of explanatory variables.

Acknowledgments

This research is part of the southern Yucatán peninsular region (SYPR) project. The SYPR project has core sponsorship from NASA's LCLUC (Land-Cover and Land-Use Change) program (NAG 56406) and the Center for Integrated Studies on Global Change, Carnegie Mellon University (CIS-CMU; NSF-SBR 95-21914), as well as sponsorship from various sources for specific elements of project study. Additional funding from NASA's New Investigator Program (NAG5-8559) supported the specific research in this article. SYPR is a collaboration of El Colegio de la Frontera Sur (ECOSUR), Harvard Forest—Harvard University, the George Perkins Marsh Institute—Clark University, and CIS-CMU (see http://earth.clarku.edu/lcluc/).

References

Allison, P.D., 1995. Survival Analysis Using the SAS System: A Practical Guide. SAS Institute Inc., Carey.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.

Anselin, L., Florax, R., 1995. New directions in spatial econometrics: introduction. In: Anselin, L., Florax, R. (Eds.), New Directions in Spatial Econometrics. Springer, Berlin, pp. 3–15.

Bryk, A., Raudenbush, S., 1992. Hierarchical Linear Models. Sage, Newberry Park.

de Janvry, A., Fafchamps, M., Sadoulet, E., 1991. Peasant household behavior with missing markets: some paradoxes explained. The Economic Journal 101 (409), 1400–1417.

Fielding, A., 2004. Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. Quality and Quantity 38 (4), 425–433.

Geoghegan, J., Pritchard Jr., L., Ognava-Himmelberger, Y., Chowdhury, R.R., Sanderson, S., Turner II, B.L., 1998. Socializing the pixel and "pixelizing the social" in land-use/cover change. In: Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P.C. (Eds.), People and Pixels: Linking Remote Sensing and Social Science. Committee on the Human Dimensions of Global Environmental Change, National Research Council. National Academy Press, Washington, DC, pp. 51–69.

Geoghegan, J., Cortina Villar, S., Klepeis, P., Macario Mendoza, P., Ogneva-Himmelberger, Y., Roy Chowdhury, R., Turner II, B.L., Vance, C., 2001. Modeling tropical deforestation in the Southern Yucatan Peninsular region: comparing survey and satellite data. Agriculture, Ecosystems, and Environment 85 (1–3), 26–46.

Goldstein, H., 1995. Multilevel Statistical Models. Edward Arnold, New York.

Goldstein, H., Healy, M.J.R., 1995. The graphical presentation of a collection of means. Journal of the Royal Statistical Society, A 158 (Part 1), 505–513.

Goulias, K.G., 2003. Multilevel statistical models. In: Goulias, K.G. (Ed.), Transportation Systems Planning: Methods and Applications. CRC Press, Boca Raton, pp. 1–22 (Chapter 9).

Guo, G., Zhao, H., 2000. Multilevel modeling for binary data. American Review of Sociology 26, 441–462.

Hosmer Jr., D.W., Lemeshow, S., 1999. Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, New York.

Houghton, R.A., 1994. The worldwide extent of land use change. BioScience 44 (5), 305–313.

Hox, J., 2002. Multilevel Analysis Techniques and Applications. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

IGBP-HDP (The International Geosphere-Biosphere Programme, The Human Dimensions of Global Environmental Change Programme), 1995. Land-Use and Land-Cover Change Science/Research Plan.

Irwin, E.G., Geoghegan, J., 2001. Theory, data, methods: developing spatially explicit economic models of land use change. Agriculture, Ecosystems and Environment 85 (1–3), 7–23.

Jones, D.W., Dale, V.H., Beauchamp, J.J., Pedlowski, M.A., O'Neill, R.V., 1995. Farming in Rondonia. Resource and Energy Economics 17 (2), 155–188.

Klepeis, P., Vance, C., 2003. Neoliberal policy and deforestation in southeastern Mexico: an assessment of the PROCAMPO program. Economic Geography 79 (3), 221–240.

Maas, C.J.M., Hox, J.J., 2004. Robustness issues in multilevel regression analysis. Statistica Neerlandica 58 (2), 127–137.

McCracken, S.D., Brondizio, E., Nelson, D., Siqueira, A., Rodriguez-Pedraza, C., 1999. Remote sensing and GIS at the farm property level: demography and deforestation in the Brazilian Amazon. Photogrammetric Engineering and Remote Sensing 65 (11), 1311–1320.

Medellin, M.A., Apedaile, L.P., Pachico, D., 1994. Commercialization and price response of a bean-growing farming system in Columbia. Economic Development and Cultural Change 42 (4), 795–816.

Pichón, F.J., 1997. Colonist land-allocation decisions, land use, and deforestation in the Ecuadorian Amazon frontier. Economic Development and Cultural Change 45 (4), 707–744.

Polsky, C., Easterling III, W.E., 2001. Adaptation to climate variability and change in the US Great Plains: a multi-scale analysis of Ricardian climate sensitivities. Agriculture, Ecosystems and Environment 85 (1–3), 133–144.

Rasbash, J., Steele, F., Browne, W., Prosser, B., 2004. A User's Guide to MLwiN Version 2.0. Centre for Multilevel Modelling, Institute of Education, University of London, London.

SARH (Secretaria de Agricultura y Recursos Hidraulicos), 1993. PROCAMPO: Vamos al grano para progresar. SARH, Mexico City.

Singh, I., Squire, L., Strauss, J., 1986. The basic model: theory, empirical results, and policy conclusions. In: Singh, I., Squire, L., Strauss, J. (Eds.), Agricultural Household Models: Extensions, Applications, and Policy. The Johns Hopkins University Press, Baltimore, pp. 17–47.

Snijders, T.A.B., Bosker, R.J., 1999. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, New York.

Turner II, B.L., Cortina Villar, S., Foster, D., Geoghegan, J., Keys, E., Klepeis, P., Lawrence, D., Macario Mendoza, P., Manson, S., Ogneva-Himmelberger, Y., Plotkin, A., Pérez Salicrup, D., Chowdhury, R.R., Savitsky, B., Schneider, L., Schmook, B., Vance, C., 2001. Deforestation in the southern Yucatán peninsular region: an integrative approach. Forest Ecology and Management 154 (3), 353–370.

Vance, C., 2004. The semi-market and semi-subsistence household: the evidence and test of smallholder behavior. In: Turner, II, B.L., Geoghegan, J., Foster, D.R. (Eds.), Integrated Land-Change Science and Tropical Deforestation in the Southern Yucatán: Final Frontiers. Oxford University Press, Oxford, pp. 221–243.

Vance, C., Geoghegan, J., 2002. Temporal and spatial modeling of tropical deforestation: a survival analysis linking satellite and household survey data. Agricultural Economics 27 (3), 317–332.

Vance, C., Geoghegan, J., 2004. Semi-subsistent and commercial land-use determinants in an agricultural frontier of southern Mexico: a switching regression approach. International Regional Science Review 27 (3), 326–347.

Walker, R.T., Moran, E., Anselin, L., 2000. Deforestation and cattle ranching in the Brazilian Amazon: external capital and household processes. World Development 28 (4), 683–699.

Walker, R.T., Solecki, W., 2001. South Florida: the reality of change and the prospects for sustainability. Ecological Economics 37 (3), 333–337.

Warwick, D., Luinger, C., 1975. The Sample Survey: Theory and Practice. McGraw-Hill, New York.

Wolfinger, R., O'Connell, M., 1993. Generalized linear models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48, 223–243.