GeoJournal **61**: 353–363, 2004.

Temporal dynamics and spatial scales: Modeling deforestation in the southern Yucatán peninsular region

Jacqueline Geoghegan[1], Laura Schneider[2], & Colin Vance[3]
[1]*Department of Economics and George Perkins Marsh Institute, Clark University, 950 Main Street, Worcester, MA 01610, USA;* [2]*Department of Geography, Rutgers University, New Jersey, USA;*
[3]*German Aerospace Center, Institute of Transport Research, Berlin, Germany*

Key words: econometric models, spatially explicit land use change, tropical deforestation

Abstract

Two spatially explicit econometric land use change models are presented, focusing on tropical deforestation caused by agricultural expansion in the southern Yucatán peninsula, Mexico. The two models developed are both based on conceptually similar theoretical models of farmer behavior. However, there are different empirical specifications of this theoretical model according to the scale of the analysis as well as the availability of temporal data on the observation of deforestation. For both models, the unit of observation for the dependent variable of deforestation is the TM pixel from satellite data. However, the socio-economic explanatory variables are derived from different sources. The first econometric model links the satellite data for the entire study region with aggregate census data at the village level. This model is estimated using a discrete choice logit model over a single time period. The second econometric model uses individual household survey data for a small random sample of the region, linked to satellite data for the plots of each household over multiple time periods. This model is estimated using a dynamic hazard model that estimates the risk of a specific pixel converting from forest to agricultural use. Both estimated models are used to predict deforestation and the results of the two modeling approaches are compared.

Introduction

The spatial distribution of land use/cover change (LUCC) as a cause of other environmental change is well documented in the natural sciences. The *spatial modeling* of anthropocentric land use change within the domain of the social sciences has been much more limited. In order to model human-induced land use change, an understanding of the human causes of LUCC are crucial. Therefore, models of anthropogenic land use change need to start with an *explanatory* model of the human behavior that can address *where*, *when*, and *why* LUCC occurs. However, while there are many behavioral theories of land use change, there are enormous data limitations to test these theoretical hypotheses. While a complete coverage of the earth's surface by remotely-sensed data (RSD) greatly facilities the observation of *where* and *when* LUCC happens, observations from space do not give much insight into *why* human-induced land use change occurs.[1]

The spatially-explicit land use change models presented in this paper are from a large interdisciplinary project in the southern Yucatán peninsula, Mexico. The aims of the project are: to understand, through individual household survey work, the behavioral and structural dynamics that influence land managers' decisions to deforest and intensify land use (this includes geo-positioning individual plots of land through GPS); model these dynamics by linking their outcomes directly to satellite imagery through GIS; model at the regional scale using the RSD and available census data; and subsequently determine the robustness of the two modeling approaches.

The most extensive land use change for the study region is the transition of forest cover to agricultural use, therefore this paper will explicitly focus on this land use change. The majority of existing empirical methodologies used in models of deforestation fall short of simultaneously capturing both the temporal and spatial dimensions of land use change processes. Most current econometric models focus on explaining the economic causes of the landscape pattern and not on capturing the temporal dynamics from which these patterns emerge. This current literature that has focused exclusively on the *location* of land use change has analyzed deforestation in tropical countries: Belize (Chomitz and Gray, 1996); northern Mexico (Nelson and Hellerstein, 1997); Brazil (Pfaff, 1999); Thailand (Cropper et al., 1999, 2001); and Panama (Nelson et al., 2001). Models that include the temporal dimension as well as location of land use change have focused on urban fringe development in the United States (Geoghegan and Bockstael,

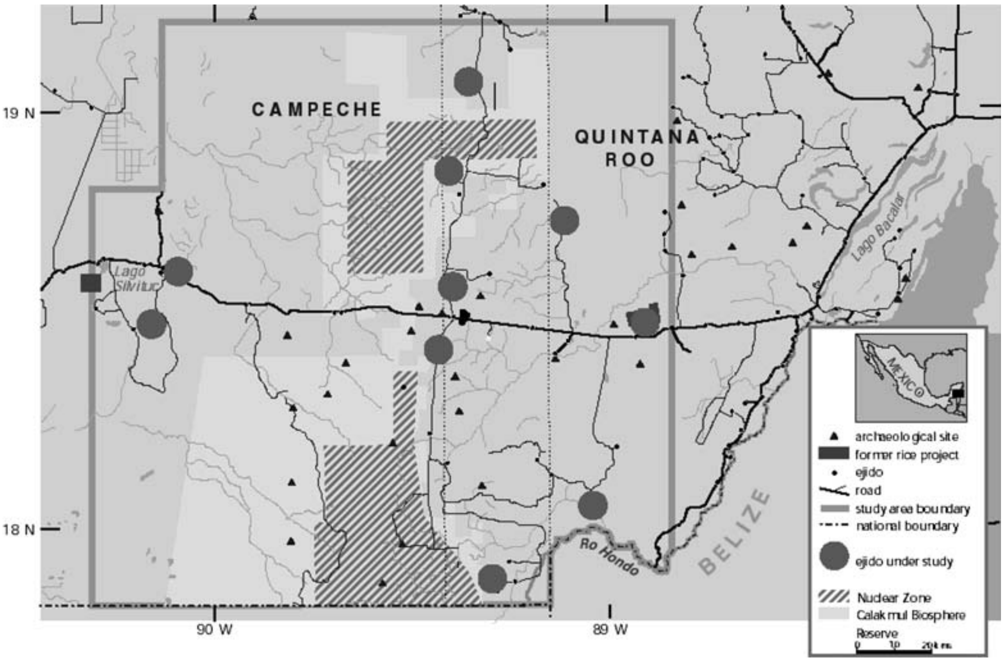

Figure 1. Southern Yucatán peninsular region study area.

2000; Irwin and Bockstael, 2001), but this latter approach can be modified to analyze tropical deforestation as well. The approach taken in this paper employs and then compares these two modeling approaches for the southern Yucatán region.

While both models take a theoretical approach of individual maximization, they differ in a number of ways, the most important of which, as suggested above, is the role of time in the decision making process. This is not to imply that *individual* behavior is necessarily different according to the spatial extent or temporal scale of the analysis. Rather, it is to suggest that the types and details of testable hypotheses derived from the theoretical models and the empirical methodologies used to test those hypotheses have to be supportable by the data that have been created and collected through this project. Indeed, in both models, the choice of explanatory variables and derivation of testable hypotheses concerning land use change are based on social science theory. For example, in a von Thünen model, the primary determinant hypothesized to affect choice is accessibility; in a Ricardian model, land quality; in a Chayanovian model, consumer-labor ratio; and in a Boserupian model, land pressures.

The role of time and the spatial extent are different in the two models presented in this paper. The first model links the satellite data for the entire study region with aggregate census data for use as the socio-economic explanatory variables in a discrete choice logit model of deforestation. This *regional* model focuses exclusively on the *location* of land use change, using as the dependent variable the results from a single GIS change analysis that was created from two different time periods of TM images. The second model uses the household survey data linked to the RSD for the plots of each household over an approximately 15-year time period. This *household* model investigates both the *location* and *timing* of land use change by tracking over time the land use associated with each particular pixel. The unit of observation for both models is the TM pixel and covers the same temporal span. The models differ in that the scalar unit for the socio-economic data for the regional model is at the *ejido* level, derived from government census data, while the scale for the household model is the individual land manager, from the household survey data.

The study region

The southern Yucatán peninsular region is located in the southern portions of the Mexican states of Campeche and Quintana Roo and encompasses approximately 22,000 km^2, of which 18,700 km^2 of *ejido* lands constitutes the main area of study (see Figure 1). A semi-evergreen tropical forest dominates this karstic region[2]. Northern Belize, El Petén, Guatemala, and the state of Chiapas, Mexico, combine with the study region to form the largest continuous expanse of tropical moist forest remaining in Central America and has been identified as a "hot spot" of forest and biotic diversity loss (Achard et al., 1998). For the first half of the 20th century, economic activity here was minimal and centered on the selective logging of tropical woods, particularly mahogany and cedar, as well as on the extraction of chicle, a tree resin used in the production of chewing gum. More extensive deforestation followed with the construction of a two-lane highway across the center of the region in 1967, which opened the frontier to agricultural colonization via the extension of *ejidal* land grants, as sanctioned by Article 27 of the Mexican constitution.

Current deforestation in the southern Yucatán peninsular region follows primarily from agricultural expansion. The four principle agricultural land uses represented currently within the region are: *milpa* (subsistence-oriented swidden agriculture, based on maize [*Zea mays* L.], squash [*Cucurbita* spp.] and legumes [*Phaseolus* spp.]), market-based chili (*Capsicum* spp.) cultivation; failed large-scale rice projects (several of which have been converted to pasture for livestock); and incipient, small-scale orchards. The research presented in this paper focuses specifically on the *ejido* sector of Mexican agriculture as this is currently the predominate form of land tenure in the study region. This sector was created following the Mexican Revolution (1910–1917), a political and social upheaval with roots in inequitable land distribution. Within *ejido* communities, land is communally regulated by an elected committee, but in this area of southern Mexico, *ejido* members (*ejidatarios)* typically enjoy usufruct access to individual parcels that are permanently allocated to their use. This tie between households and parcels permits land-use decisions to be linked to the geographical locations in which they have impact. In the recent past, most *ejidatarios* primarily engaged in subsistence production, but recently an increasing proportion of them engage the market, primarily through chili pepper cultivation. The region today is best considered one dominated by semi-subsistence small holders experimenting with chili and other vegetables, fruits, non-timber forest products, livestock, bee-keeping, and, where possible, engaged in modest off-farm employment.

Data methodology

The survey design and questionnaire

The survey data used in this study were collected by the project over an eleven-month period beginning in October of 1997. To accommodate the limitations of available sampling frames and control data collection costs, selection of respondents proceeded according to a stratified, two-stage cluster sample, with the first stage unit being *ejidos* (approximately 125 in the region) and the second stage unit being *ejidatarios* (Warwick and Lininger, 1975; Deaton, 1997). Completion of the first stage selections involved compiling an enumeration of *ejidos* in the region with the aid of 1990 cartographic sources and census records, both published by the Mexican government's National Institute of Statistics and Geography (INEGI). Using these maps, the region was partitioned into 11 geographic divisions – or *strata* – and randomly selected one *ejido* from each of the strata. Each *ejido* was assigned a probability of selection equal to the ratio of its population to the population of the stratum. The strata were drawn to ensure that *ejidos* from across the region were represented in the sample, thereby capturing variability in both ecological conditions and in the influence of market and road proximity. Ten out of the 11 *ejidos* selected have assigned parcels in which *ejidatarios* have usufruct access to a specific area of land. Settlers from some 23 states in Mexico have colonized the southern Yucatán peninsular region and the sample captures this regional and ethnic diversity.

In the second stage, the survey respondents themselves were randomly selected from each *ejido* after an inventory of households was enumerated. This inventory was secured after establishing contact with the *comisariado* (community leader) of each *ejido*, who, along with the *ejido* assembly, granted permission to proceed with the surveying. The number of households surveyed from each *ejido* was approximated such that the corresponding stratum was represented in roughly the same proportion as its share of the total population of the study zone. After eliminating 12 records collected during a pilot phase of the survey, the final sample size was 188 observations.

A standardized questionnaire was used to elicit the socio-economic and land use data. The questionnaire was administered to the household head and was comprised of two sections. The first section elicited information on migration history, farm production and inputs, off-farm employment participation, land acquisition, and the demographic composition of the household. Completion of the second section involved a guided tour of the respondent's agricultural plot. Using a 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 use of these fields. The sketch map not only documented the spatial configuration of contemporary land uses, including forested areas, but also the land use transition histories of the principal areas of activity. This information was later used to calculate various indices of land use change, such as the area deforested since acquisition of the plot, as well as to inform the classification and interpretation of the satellite imagery.

Satellite data

The study region is covered by portions of three Landsat TM scenes, identified by the Landsat Worlwide Reference System as path 20, row 47; path 19, row 47 and path 19 row 48. TM images collected from late January to early April in 1984, 1985, 1987, 1988, 1994, 1995, 1996 and 1997 were used to create regional land cover maps. The individual scenes were first georeferenced to a latitude–longitude projection (Datum NAD27, Mexico) to under 0.5 RMS positional error. Following image registration, the TM scenes were subjected to haze removal using Tasselled Cap transformation. Using the de-hazed red and infrared bands, NDVI images were produced for each date for subsequent analysis. The de-hazed images were then processed using Principal Components Analysis (PCA) to complete noise removal and reduce data redundancy. PCA was run separately on the visible and near infrared bands, and the first three principal components from the separate analysis were retained for further processing. The higher order components captured random as well as systematic noise in the data, such as striping, and were dropped from further analysis.

The next steps consisted of performing texture analysis on the 3-band principal components image. By generating measures of spectral variance in immediate spatial neighborhoods of individual pixels, texture analysis produced three additional bands of information based on the original 3-band PCA image. The three texture bands were stacked with the three PCA bands to produce a six-band image. Next, the NDVI image produced from the originally de-hazed red and infra-red bands was added to the PCA and texture bands to generate a final seven band image for signature development and classification.

Training-site development involved ''ground-truthed'' data derived from GPS-assisted field visits, and topographic, vegetation and land-use maps. To this array of information detailed sketch maps on recent land-use history linked to household surveys was added. These maps were registered by GPS and used to focus on extant and past signals in the imagery (Klepeis and Turner, 2001). The availability of sketch maps provided additional ground truth data that augmented not only the number of training sites for land cover signatures, but also improved the geographic distribution of those training sites. The land-cover signatures were further refined by accepted measures of separability (e.g., Euclidean distance, divergence, transformed divergence, and Jefferies–Matusita distance). The final signatures were used in a maximum likelihood supervised classification to produce categorical land-cover maps used in modeling. All classes were subjected to accuracy assessments.

Two regional mosaics were created from the classified scenes, one for the mid-1980s (scenes for 1984, 1985, 1987 and 1988) and the second for the mid-1990s (scenes for 1994, 1996 and 1997). The dates for each mosaic were selected to match seasons (then reduce the effect of seasonal factors on the final classification) and minimize areas cover by clouds. The final classification scheme has seven classes: wetland forest (*bajos*); upland forest (*selva mediana*); seasonally inundated savanna; secondary vegetation from 4 to 15 years; agriculture which includes herbaceous secondary vegetation, cropland and pasture; water; and bracken fern *Pteridium aquilinum* (L.) Kuhn (a fern which apparently invades after overcropping *milpas*, pushing farmers to abandon the invaded plot). Change detection was estimated through spatial cross-classification, an overlay that provides the locations of the combinations of the two categories used in the merged classified mosaics.

Other data sources

Other spatial data collected or created for use in the models include: elevation and slope from a digital elevation model; soil types digitized form a 1:250,000 Mexican government (INEGI) map; digitized road network from INEGI 1:50,000 topographic maps; rainfall data (interpolated to cover the region) from the Mexican government (Secretary of Agricultural and Hydrological Resources); and socio-demographic data from the Mexican government 1990 population census. The soil maps for the region are complex, offering many geographical ''clusters'' of soils based on taxonomic order and type, plus other characteristics. For modeling purposes, the many types and subtypes have been aggregated into two simple categories that reflect the basic conditions facing farmers: rendzina and other mollisols (''upland soils'', preferred by farmers); litosols and vertisols (all other soils). With the use of the road layers, distance from each individual pixel to roads and markets were calculated. Only four precipitation-temperature stations exist in the region. To these, 16 others lying just outside the region were added to create 20 stations or precipitation points and used to create an interpolated map (ordinary kriging analysis, see Isaacks and Srivastava, 1989) of region rainfall (annual dry season, averaged over 30 years). Using a GIS map of *ejido* boundaries, the census data were linked with the pixels associated with each of the *ejidos* via a uniform distribution, so that each pixel in the *ejido* was assigned the value of the variable from the census data for that *ejido*. The data include information on number of households, language and literacy, and structural characteristic of houses, such as electricity and water service. While the level of detail is not ideal, the data do span the entire spatial extent of our study region, and constitute the only current source of socio-economic data for *ejidos* not covered in the household survey.

The models

Theoretical economic model and empirical specification

The simplest characterization of the land use decision for an individual land manager is based on profit maximization, applicable when the farmer is engaged in product and factor market. A more general characterization, that of utility maximization, is required to model the behavior of a subsistence-based farmer. While this difference has been studied previously for this region (Vance and Geoghegan, 2004), for this current work it is assumed that the land manager is maximizing profits. The first theoretical model considered focuses exclusively on the *location* of land use change. The land manager chooses land use in period T that maximizes expected net returns. Let $A(i,T)$ represent the present value of the future stream of returns to pixel i that is cleared in time period T. Let $F(i,T)$ represent the present value of the future stream of returns for leaving pixel i in forestry use in each time period, and let $C(i,T)$ be the one-time clearing costs associated with clearing the pixel in time period T. Then $V(i,a,T)$ represents the *net* returns to pixel i of deforestation at time T. Then the static conversion rule for the individual land manager is clear pixel i when the net future stream of returns to agricultural use are greater than the stream of returns to forest use. Given that all factors that affect V are not observable, this statement can be re-written in terms of the *probability* of parcel i being cleared, with an error term e, representing the unobserved random component to net returns. Let the characteristics of pixel i be $X(i)$, including the characteristics of the land manager of that pixel, and assume that the net returns are a linear function of these characteristics, then the decision rule for clearing is clear if: probability $(X(i,a,T)\beta_i + e(i,a,T))$ > probability $(X(i,f,T)\beta_i + e(i,a,T))$. By assuming an extreme value distribution for the error terms, the β parameters of this model can be estimated using a logit specification.

The second theoretical model is a slightly more sophisticated model than is given above and focuses on both the *location* and *timing* of land use change. The empirical specification of this model is used to estimate the factors affecting deforestation at the individual plot level, using the household survey data for information on socio-economic factors. Now, let $A(i, t)$ be the net benefits to agricultural use for *each time period* after pixel i is cleared in time period T. Let $F(i, t)$ be the net benefits to the farmer for leaving pixel i in forestry use in *each time period*, and let $C(i,T)$ be the one-time clearing costs associated with clearing the pixel in time period T and δ is the discount rate. Then for T to be the optimal time to clear, two conditions must hold: (i) the net returns to clearing are positive, similarly to the first model; (ii) there are no expected benefits to waiting because of the potential for even higher benefits at some future date. Such a circumstance could arise, for example, in anticipation of improved technologies that reduce clearing costs. In this model, the optimal time for clearing this pixel then is the first time period in which the following holds: $A(X(i),T) - \mathrm{F}(\mathrm{X}(i),\mathrm{T}) - \delta\mathrm{C}(\mathrm{X}(i), \mathrm{T}+1) \geq 0$. By adding a similar error term to this model, the parameters of this dynamic model can be estimated using a hazard model.

Hazard models typically estimate the conditional probability of exiting a state given that the state has been occupied for some length t. The dependent variable, the duration, is the length of time that elapses from the beginning of the state until its end or until measurement is taken and therefore truncates the observation. For this paper, the duration of interest is the length of time that an individual pixel remains in forest before being converted to cropland or pasture. Let P_{it} be the probability that deforestation occurs to pixel i in interval t given that the pixel was not deforested any earlier periods and let t_i be the time period in which pixel i is deforested. Then, if it is assumed that the error terms are distributed extreme value, as in the logit model above, the empirical specification can be derived as a complementary log–log model: $\log[-\log(1-P_{it})] = X(i,T)\beta_i$ where the β_i terms are parameters to be estimated.

Variables and hypotheses

As suggested in the theoretical models presented above, the land clearance decision will be based on a comparison of the net benefits from forest and non-forest land uses. There are several testable socio-economic and environmental factors, the $X(i)$ variables, that could influence this comparison. For both the regional logit model and the household hazard model, the variables in the $X(i)$ vector are qualitatively similar consisting of socio-economic, biophysical and spatial variables. While both models use the TM pixel the unit of observation, the information associated with each of the pixels differs. For the regional model, all that is known about the management unit of each individual pixel is that it is in a particular *ejido*, while for the household model, the land manager associated with that pixel as well as the total amount of land available to that land manager are known. This important difference leads to some extent to different specifications for the socio-economic and spatial variables in the two models. For example, the total area of the *ejido* is known (from digital land tenure maps described above), for the regional model, but it is not known how much of that land is actually available for agricultural use. Therefore, it is assumed that all pixels have a potential for agricultural use. However, some of the contiguity of agricultural land uses is controlled for by included variables on the total number of agricultural pixels in a five by five pixel window around each pixel as well as a measure of the distance from each pixel to the nearest agricultural pixel. The hypothesis is that the greater and closer the number of agricultural pixels, the greater the likelihood for a pixel to be cleared for agricultural use.

For the household model, the number of males older than 11, number of females older than 11 and total number of children under 12 are included as exogenous variables, this information having been collected during the field survey. These indices are measured as the average number of members in each age/sex category over the corresponding time interval of the imagery and accordingly vary across households and through time. For the regional model, by contrast, only total number of males and females for the *ejido* in 1990 can be included given that the census data only contains information on total population size by gender but not by age. The expectation in both cases is that population pressure will increase the probability of deforestation. For the household model, other socio-economic variables include education of household head as well as the number of household members with more than eight years of education, while for the regional model, education is measured by the population older than 15 who has attended post-elementary school. The *a priori* expectation on the education variables is that greater education increases the off-farm employment opportunities, thereby leading to a decrease in deforestation. In both models, native speakers of Spanish are controlled for either by a dummy variable if the household head is a native speaker, as in the household model, or by a measure of the total number of native speakers of Spanish in the *ejido*, as in the regional model. It is hypothesized that native speakers of Spanish have greater off-farm employment opportunities than the indigenous Maya population.

A variable measuring credit received is included in both models but with different specifications. In the household model, this variable is measured by the percent of the time interval that the household received government agricultural credit, while in the regional model it is measured by the percent of the *ejido* population that received government agricultural credit in 1990. The expectation in both models is that a greater level of agricultural credit will lead to expansion of agricultural output, thereby increasing deforestation. Because of data limitations, other potential wealth effects and physical capital are included in different ways. In the household model, the percent of each observed time interval that the household owned a chain saw and a vehicle are included, both of which are expected to increase deforestation. For the regional model, the only vaguely similar variables available were the percent of households in the *ejido* that had either electricity or running water in their houses, with the hypothesis that wealthier communities have more off-farm opportunities, thereby decreasing the probability of deforestation. The variability in land tenure is controlled for in the regional model, by including a dummy variable to indicate if a pixel is within a forest extension, where agricultural cultivation is generally prohibited, but is occasionally observed in the data. To control for unobserved differences across *ejidos*, a dummy variable for each is included in the household model.

In both models, the biophysical variables are the same: dummy variables indicating whether the pixel began the time period of observation in primary forest and whether the pixel had upland soils; the slope and elevation of the pixel; and the average annual rainfall associated with the pixel. The potential effect of primary forest is unclear. Primary forests are associated with higher fertility soils, but have much higher clearing costs than secondary forests. Farmers prefer upland soils, so it is expected that this variable will have a positive effect on the probability of deforestation. It is hypothesized that higher slope and elevation, decrease the probability of deforestation, while an increase in average rainfall will increase the probability of clearing for agricultural land use. Land availability is controlled for by the land area associated with each household in the survey in the household model and by the total land area of the *ejido* in the regional model. The expectation in both cases is that with a larger land endowment, the probability of deforestation for an individual pixel will decrease. The spatial variables associated with each model are slightly different. For the household model, the distance measures include distance from the household to the plot and distance from the *ejido* that the household is in to the nearest market. For the regional model, these measures are the distance from each pixel to the nearest road as well as distance to nearest market. In each case, it is expected that an increase in accessibility will lead to an increase in deforestation.

To capture the effect of unobserved, inter-temporal factors affecting land-use choices, the household model includes a variable that measures the duration of the household's occupancy as of the end of the time

Table 1. Summary statistics for regional aggregate census logit model.

Variable	Definition	Mean	SD	Min	Max
Exten	Forest extension (dummy)	0.20	0.40	0	1.00
Male	Male population (per ejido)	316.49	368.91	9	1529.00
Female	Female population (per ejido)	209.13	192.26	11	725.00
Span	Spanish speaking (per ejido)	97.91	157.94	0	636.00
m_prim15	Higher education (per ejido)	24.73	31.47	0	109.00
num_pixl	Ejido size (# pixels)	284,549.60	252,115.8	8063	791,775.00
Primary	Primary forest (dummy)	0.89	0.32	0	1.00
Elevation	Elevation (m amsl)	170.63	65.87	0	350.00
Rain	Rainfall (mm)	1123.27	86.44	0	1315.00
Slope	Slope (degrees)	1.11	2.25	0	67.00
freq_ag	Number ag pixels (5×5)	0.43	1.80	0	24.00
d_roads	Distance to roads (m)	5198	3991.32	0	25,750.00
Agdist	Distance to nearest ag land (m)	860.99	966.73	30	7252.00
d_market	Distance to market (m)	30,458.98	15,751.22	30	77,420.00
pwatr	Percent with water (hh/ejido)	31.72	39.32	0	100.00
pelec	Percent with electricity (hh/ejido)	47.37	37.63	0	95.65
p_credit	Percent receiving credit (hh/ejido)	20.03	16.01	0	76.79
Goodsoil	Good soil (dummy)	0.68	0.47	0	1.00

Table 2. Summary statistics from household survey.

Variable	Definition	Mean	SD	Min	Max
Male	Average # of males > 11 over interval	1.61	1.03	0	7.00
Female	Average # of females > 11 over interval	1.44	1.04	0	7.00
Kids	Average # of children < 12 over interval	1.60	1.79	0	6.75
Primary	Primary forest	0.79	0.41	0	1.00
Soil	Upland soil	0.76	0.43	0	1.00
Dem	Elevation	168.04	71.09	0	310.00
Slope	Slope	1.23	2.78	0	58.67
Dry	Precipitation	56.10	4.29	49	72.00
totpixl	Plot size (# of pixels)	1321.92	1051.02	6	4819.00
Saw	Percent of interval owning chain saw	0.25	0.37	0	1.00
vehic	Percent of interval owning vehicle	0.13	0.30	0	1.00
edhead	Education of household head	3.31	3.87	0	14.00
ed9	# of household members w/ > 8 years education	1.05	1.57	0	7.00
Span	Native Spanish speaker	0.83	0.38	0	1.00
Credit	Percent of interval receiving government credit	0.34	0.36	0	1.00
Dis	Distance household to plot	9.07	7.62	0	28.60
Nearmkt	Distance ejido to nearest market	21.23	19.22	0	82.63
yr	Duration of occupancy	20.23	12.82	0	58.00
yrsq	Duration of occupancy squared	573.54	674.30	0	3364.00

Table 3. Econometric results from regional logit model.

Dependent variable: Deforestation or not	Estimated coefficient	*t*-statistic
Primary	−0.8773	−150.74
Elevation	−0.0059	−91.57
Rain	0.0012	22.24
Slope	0.0145	13.15
Goodsoil	0.3281	53.37
freq_ag	0.0592	67.11
Agdist	−0.0021	−201.97
d_roads	−0.0001	−109.03
d_market	−4.45e−06	−12.72
Female	−0.0004	−2.22
Male	0.0035	21.73
num_pixl	−2.01e−06	−88.26
m_prim15	−0.0098	−39.26
Span	−0.0041	−69.04
pwatr	−0.0048	−33.82
pelec	−0.0006	−4.49
Exten	−0.0795	−5.96
p_credit	0.0038	14.60
Constant	−2.099	−30.35
Pseudo R^2	0.2264	
Number of observations	3,944,875	

Table 4. Econometric results from individual household hazard model.

Dependent variable: Deforestation or not	Estimated coefficient	*t*-statistic
Male	0.0145	1.63
Female	0.0365	3.88
Kids	0.0270	5.40
Primary	−0.6861	−39.62
Soil	0.2566	10.62
Dem	−0.0104	−21.92
Slope	−0.0253	−6.33
Dry	0.1453	19.99
totpixl	−0.0002	−16.77
Saw	0.1366	4.55
Vehic	0.4522	13.09
edhead	0.0261	9.81
Ed9	0.0322	5.40
Span	−0.1005	−3.93
Credit	−0.2248	−7.81
Dis	−0.0452	−27.56
Nearmkt	−0.0723	−36.72
yr	−0.0413	−13.20
yrsq	0.0005	8.499
Constant	−3.5278	−8.70
Likelihood ratio $\chi^2(31)$	18,641.66	
Number of observations	115,017	

interval. The square of this variable is also included to allow for non-linearities. A similar variable does not exist from the census data for the regional model. Finally, both models include dummy variables for each TM scene used to create the dependent variable to control for the fact that they are of differing lengths (Allison, 2000). Summary statistics for the regional model can be found in Table 1 and for the household model in Table 2.

Estimation results

The results from the regional logit model can be found in Table 3 and the results from the household hazard model can be found in Table 4. The size of the estimated coefficients for the two models cannot be easily tested to determine if they are statistically significantly different from each other, as they are different statistical models using different datasets. However, it is possible to compare the sign of the estimated coefficients of similar variables between the two models. Specifically, it is of interest to identify how each similar explanatory variable increases or decreases the probability or risk of deforestation in both of the model specifications.

With respect to the demographic variables in the hazard model, two of the three indices, females over 11 and children under 12, are seen to be positive and statistically significant determinants of the hazard of deforestation. The coefficient on males is not statistically significant. However, the estimated results from the regional model indicates that male population has a positive and statistically significant effect on increasing the probability of deforestation while female population is negative and statistically significant. While the estimated coefficients on the gender variables differ between the models, the overall results are similar, in that population pressures lead to greater levels of deforestation. This is an unsurprising result given that the majority of households in the region are semi-subsistent producers, for whom which family members simultaneously represent a source of labor as well as demand for outputs from the agricultural plots. Future research will investigate further this gender difference between scales of analysis.

All of the biophysical variables are statistically significant and all but one have the same sign across the two models. The primary forest variable is negative in both models, suggesting that farmers prefer to clear secondary forest, all else being equal. This is perhaps because clearing costs are much lower, even though the weeding costs associated with agricultural use on secondary growth is higher (Dvorak, 1992). The superior upland soils as well as increased average rainfall increase the hazard or probability of deforestation in both models, as expected. Higher elevation is estimated to the lower the likelihood of deforestation in both models[3]. The only biophysical variable that differs between the two models is the slope of the pixel. The logit model yields the counterintuitive result that increased slope increases the probability of deforestation.

With respect to the variables measuring the land endowment, each model yields negative and statistically significant coefficients, so that all else being equal, the larger the land endowment, the less the pressure on forested lands for agricultural use. In the household hazard model, the time-varying parameters of ownership of a chain saw and vehicle, both increase the hazard of deforestation, which is consistent with the idea of lower labor costs in forest clearance and in access to the plot. For the regional logit model, the percent of houses with running water and electricity, which attempt to measure a level of *ejido* wealth or connectivity with political forces, both decrease the probability of deforestation. These results may reflect a correlation with unobserved variables such as access to more off-farm opportunities, as would characterize the *ejidos* closer to the market centers where electricity and running water are more common.

The education variables are all statistically significant in each model, but with different signs: in the hazard model education has a positive effect on deforestation, while in the logit model it has a negative effect. To the extent that higher education implies a higher opportunity cost of on-farm labor due to increased wage-earning potential, the expectation is for these variables to carry negative coefficients. However, higher education could also imply better agricultural management skills, leading to an increase in potential returns to agricultural use. Perhaps it is this difference in the interpretation of the education variables that leads to this difference in the empirical results from the two scales of the model. Being a Spanish speaker decreases the hazard and probability of deforestation in both models, which may result from a greater reliance of indigenous farmers on the resource base as opposed to off-farm wage earning opportunities.

The credit variables are statistically significant in both models, but have differing signs. In the household model, the estimated coefficient is negative, while in the regional model, it is positive. This outcome could be the result of the different way that the variable is defined in each of the models. However, if it can be argued that the two variables are the same basic measure, then perhaps the results show that at the household scale, the estimated negative coefficient on credit could be capturing the increased ability to purchase land-saving inputs such as chemical fertilizers, thereby intensifying land use and decreasing the pressure on forests, while at the *ejido* scale, the greater the amount of government credit, the greater the ability to, for example, hire labor and clear more forest land for agricultural use. In the regional model, the negative coefficient on the forest extension tenure dummy variable indicates that all else being equal, land is less likely to be cleared there for agricultural use.

For the location and access variables in both models, the estimated coefficients are negative and statistically significant. This is consistent with the hypothesis that parcel location affects the land under cultivation for both profit-maximizing and subsistence farmers since each will be operating under a time constraint that is partly determined by travel time to and from the parcel. Moreover, higher travel costs reduce the farm-gate price of output for the market-oriented farmer.

In the hazard model, different specifications of the variable measuring duration of occupancy were explored by means of a nested likelihood ratio test, where it was determined that the quadratic functional form was optimal in terms of fit and parsimony. The estimates indicate that the conditional probability of forest conversion decreases with the passage of time at a decreasing rate. This result may reflect a confluence of factors, including fallow cycle strategies and adaptation to local market opportunities and ecological constraints. A similar variable is not available at the regional level, although future research will attempt to include a vari-

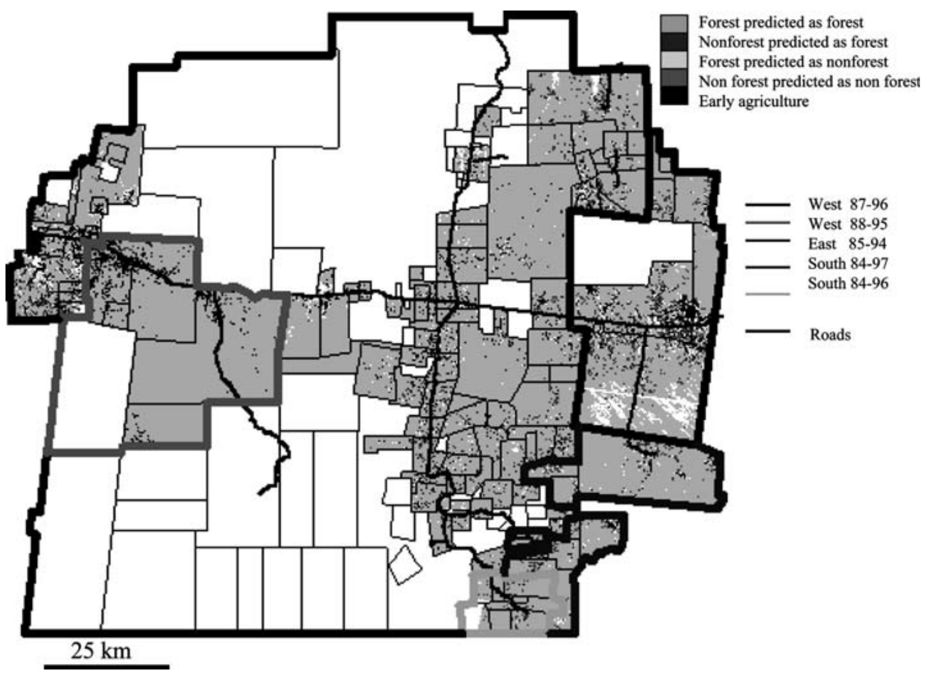

Figure 2. Regional model: correct and incorrect predictions.

able on the date of *ejido* formation to test similar hypotheses concerning the role of duration of occupancy on land use change.

Prediction

Prediction in models where the dependent variable is a zero or one, as in both models used in this paper, is different than in a regression model, in that the prediction for any observation is a *probability* of an event or choice occurring. In each of the models here, the probability of a particular pixel becoming deforested was estimated. In order to compare the predicted results with what actually occurred, some method of comparison must be developed. In the "actual" event, either a pixel was deforested or not, so there is no clear way of comparing predicted value versus actual value. That is, the prediction is a probability of change, potentially ranging from 0% to 100%, while the "actual" is either "yes" the pixel did change, or "no" it did not. The question is then, what percent predicted probability of deforestation is the critical value for a pixel to be considered a "yes" deforestation pixel?

Often researchers take 50% as the critical threshold, that is, if the probability of a change occurring is greater than 50%, then that observation is calculated as a change, in order to compare to the actual changes. However, a different approach, developed in previous work (Geoghegan, et al., 2001) is taken here. This approach can conceptually be divided into two steps for each model. Instead of imposing an ad hoc percentage to consider a pixel a "deforested" pixel, first the total number of actual pixels, n, that were deforested was identified. Secondly, the spatial distribution of highest probability pixels were sequentially identified until n have been selected. That is, each model does not attempt to predict the actual amount of deforestation, but takes that amount as given, and predicts the spatial distribution of that deforestation over the landscape. For the regional logit model, this critical threshold was 21%, so that any pixel with a predicted probability of deforestation of 21% or higher was considered deforested for comparison with actual deforestation. A similar procedure was completed for the household hazard model. The critical threshold for the hazard model is 29%.

This difference in probabilities between the regional census logit model and the household hazard model could be considered a measure of the improved fit of the household model, given the higher cut-off probability. Another way to evaluate the relative fits of the models is to compare the total number of correct and incorrect predictions. The logit model correctly predicts 33% of the deforested pixels, while the hazard model correctly predicts 46% of the deforested pixels. For the pixels that remain forested, the logit model correctly predicts 96% of those pixels, while the hazard model correctly predicts 89% of those pixels. The improvement of the hazard model for predicting deforestation reflects the value of adding the richer household level survey data that better captures the individual causes of deforestation than is possible at the aggregate level with the census data in the regional model. An additional possible reason for the better predictions on remaining in forest for the logit model over the hazard model is that for the latter, the agricultural plot boundaries for the individual household are known from the sketch map work associated with the household survey. However, for the regional model, only the boundaries of each *ejido* are known, not the boundaries of the potential agricultural land within

Figure 3. Household model: correct and incorrect predictions detail.

each *ejido*. As much *ejido* land is designated as communal forest, and is therefore ostensibly off limits for cultivation, the model coincidently achieves a relatively high accuracy in correctly predicting such land to remain in forest.

The maps of these comparisons of predicted deforestation to actual deforestation for the regional logit model are found in Figure 2 for the entire region[4]. The overall pattern of predictions in Figure 2 shows the large number of pixels that were correctly predicted to remain in forest throughout the region. The correctly predicted deforested pixels are mainly found along the road network and contiguous to previous agricultural land use, showing the influence of these important variables on the predictions. However, the incorrectly predicted deforested pixels (i.e. the pixels that were predicted to be deforested, but actually remained in forest) also show a similar spatial pattern, suggesting that these two variables, distance to roads and distance to nearest agricultural land use, is over-predicting in the model. The incorrect forest predictions (i.e. those pixels that were predicted to remain in forest use, but actually were deforested) are much more randomly distributed throughout the landscape, suggesting that there are further spatial processes occurring in these areas that are not currently explained by the model. The most important information that is currently lacking at the regional level is further information on the location of allowed land uses within an *ejido*, such as the forest communal lands, as mentioned previously, or individual agricultural plot boundaries, as is available for the survey data.

As the spatial scale of analysis for the hazard model, the individual parcels from the household survey, is small and scattered over the entire region[5], it is not possible to make one easy to read map of the predicted results, so instead what is included is a ''zoom in'' on one of the regions on the eastern edge of the study region. In Figure 3, the correct and incorrect predictions are shown for a small area of the household model. Visual inspection of these results suggest that in the southeastern portion of this area, which is near the main road, the model predicts correctly the deforested pixels, while in the remainder of the southern region, the model predicts correctly a large portion of the remaining forest lands. However, in the northwest part of this region, the results appear to be more mixed and spatially scattered, suggesting that the spatial pattern of the agricultural behavior of individual *ejidotarios* is not well captured by the current model. It is hard to visibly discern much more of a spatial pattern of these results, except to comment that the pixels where both models are incorrectly predicting do appear to be clumped together, so that for particular farmers in particular places, both models are incorrectly predicting behavior, and future research will focus on further investigation of what modeling features are driving these results.

Conclusions

This paper has demonstrated some of the methodologies developed in this research project for combining different types of spatial data for modeling land use change. Elsewhere, this kind of effort has been referred to as ''socializing the pixel'' and ''pixelizing the social'' (Geoghegan, et al., 1998), a centerpiece of proposed work by the larger LUCC community of researchers. The aim of this body of work is to improve understanding and explanation of the variation in the regional dynamics of land change, including improvements in spatially-explicit models that move beyond the magnitude of changes within a region to the locations of that change.

There remains much data and modeling work to be done to achieve these aims. As previously discussed, further information on the location of within-*ejido* settlement locations and *ejido* forest reserves, and variations in *ejido* usufruct rules would likely greatly improve the accuracy of the regional model. Further modeling work includes adding more temporal observations to the regional model, so that a hazard model approach could be used for those data. Then a more direct comparison between the regional model and the household model will be possible. As of now, it is likely that at least some of the differences between the two models estimated in this paper are artifacts of the different empirical specifications used and not necessarily a feature of the spatial or temporal scales of the data or the models.

However, this preliminary comparison of the two modeling approaches leads to some tentative conclusions. The higher critical value for calculation of "deforested or not" for the household hazard model versus the census logit model (28% versus 21%), and the increased precision in the total number of correctly predicted deforested pixels (46% versus 33%) implies an overall better fit of the former model over the latter model for the modeling of the factors that affect deforestation for agricultural use in the region. This suggests the overall "value added" of using the richer individual specific data made available through the household survey work.

Acknowledgements

This work was undertaken through the auspices of the Southern Yucatán Peninsular Region project with 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). Additional funding from NASA's New Investigator Program in Earth Sciences (NAG5-8559) also supported the specific research in this article. This project is a collaborative project of El Colegio de la Frontera Sur (ECOSUR), Harvard Forest-Harvard University, the George Perkins Marsh Institute-Clark University and CIS-CMU. We would like to thank all of our project colleagues, especially Peter Klepeis, Yelena Ogneva-Himmelberger, and Rinku Roy Chowdhury for their help with the survey and satellite data.

Notes

1. Recent advances in land use change modeling are discussed in Veldkamp and Lambin (2001).
2. Karst refers to lands dominated by limestone and typified by erosional process of solution, paucity of surface water (e.g., streams), and abundant sinks and cave systems.
3. Higher elevated lands tend to be related to more rugged the terrain, shallow soils and caprock, and less soil moisture, all factors inhibiting yields.
4. This map also includes the borders of each different TM scene that was used in the analysis and their associated dates.
5. See Figure 1 for the spatial distribution of the *ejidos* sampled in the household survey.

References

Achard F., Eva H., Glinni A., Mayaux P., Richards T. and Stibig H.J. (eds), 1998: *Identification of Deforestation Hot Spot Areas in the Humid Tropics*. Trees Publ. Series B. Research Report No. 4. Space Application Institute, Global Vegetation Monitoring Unit. Joint Research Centre, European Commission. Brussels.

Allison P.D., 2000: *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute Inc., Carey, NC.

Chomitz K.M. and Gray D.A., 1996: Roads, land, markets and deforestation: a spatial model of land use in Belize. *World Bank Economic Review* **10**: 487–512.

Cropper M., Griffiths C. and Mani M., 1999: Roads, population pressures, and deforestation in Thailand, 1976–1989. *Land Economics* **75**: 58–73.

Cropper M., Puri J. and Griffiths C., 2001: Predicting the location of deforestation. *Land Economics* **77**: 172–186.

Deaton A., 1997: *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. The Johns Hopkins University Press, Baltimore.

Dvorak K., 1992: Resource management by west African farmers and the economics of shifting cultivation. *American Journal of Agricultural Economics* **74**: 809–815.

Geoghegan J. and Bockstael N.E., 2000: Smart growth and the supply of sprawl. Paper presented at the Association of Environmental and Resource Economists Workshop on the Effectiveness of Resource and Environmental Regulation. La Jolla, CA.

Geoghegan J., Cortina Villar S., Klepeis P., Macario Mendoza P., Ogneva-Himmelberger Y., Roy Chowdhury R., Turner II, B.L. and Vance C., 2001: Modeling tropical deforestation in the southern Yucatán peninsular region: comparing survey and satellite data. *Agriculture, Ecosystems, and Environment* **85**: 25–46.

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

Irwin E.G. and Bockstael N.E., 2001: Interacting agents, spatial externalities, and the endogenous evolution of residential land use pattern. *Journal of Economic Geography* forthcoming.

Isaacks E.H. and Srivastava R.M., 1989: *Applied Geostatistics*. Oxford University Press, Oxford.

Klepeis P. and Turner II B.L., 2001: Integrated land history and global change science: the example of the southern Yucatán peninsular region project. *Land Use Policy* **18**: 272–39.

Nelson G., Harris V. and Stone S., 2001: Deforestation, land use, and property rights. *Land Economics* **77**: 187–205.

Nelson G.A. and Hellerstein D., 1997: Do roads cause deforestation? Using satellite images in econometric analysis of land use. *American Journal of Agricultural Economics* **79**: 80–88.

Pfaff A., 1999: What drives deforestation in the Brazilian amazon? Evidence from satellite and socioeconomic data. *Journal of Environmental Economics and Management* **37**: 26–43.

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

Veldkamp A. and Lambin E.F., 2001: Predicting land-use change. *Agriculture, Ecosystems & Environment* **85**: 1–6.

Warwick D.P. and Lininger C.A., 1975: *The Sample Survey: Theory and Practice*. McGraw Hill, Inc., New York.