Environ Ecol Stat
DOI 10.1007/s10651-007-0012-5

A regression tree-based method for integrating land-cover and land-use data collected at multiple scales

Jeffrey A. Cardille · Murray K. Clayton

Received: 1 November 2003 / Revised: 1 September 2005

Abstract As data sets of multiple types and scales proliferate, it will be increasingly important to be able to flexibly combine them in ways that retain relevant information. A case in point is Amazonia, a large, data-poor region where most whole-basin data sets are limited to understanding land cover interpreted through a variety of remote sensing techniques and sensors. A growing body of work, however, indicates that the future state of much of Amazonia depends on the land use to which converted areas are put, but land use in the tropics is difficult to assess from remotely sensed data alone. An earlier paper developed new snapshots of agricultural land use in this region using a statistical fusion of satellite data and agricultural census data, an underutilized ancillary data source available across Amazonia. The creation of these land-use maps, which have the spatial detail of a satellite image and the attribute information of an agricultural census, required the development of a new statistical technique for merging data sets at different scales and of fundamentally different data types. Here we describe and assess this nonlinear technique, which reinterprets existing land cover classifications by determining what categories are most highly related to the polygon land-use data across the study area. Although developed for this region, the technique

The figures in the printed version of this article appear in black and white. Color figures are available from the author upon request.

J. A. Cardille
Center for Sustainability and the Global Environment,
University of Wisconsin-Madison, 1710 University Ave.,
Madison, WI 53726, USA

M. K. Clayton
Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA

Present Address:
J. A. Cardille (✉)
Département de Géographie, Université de Montréal,
Montréal, QC, Canada H3C 3J7
e-mail: jeffrey.cardille@umontreal.ca

appears to hold broad promise for the systematic fusion of multiple data sets that are closely related but of different origins.

Keywords Polygon · Raster · Amazon · Land use · Land cover · Agriculture · Census · Deforestation · Crop · Pasture

1 Introduction

With land-cover conversion for cropland and pasture ongoing in Amazônia, concern has mounted over the current and future conditions within this rapidly changing area. In this data-poor region, most studies are limited to understanding land cover, which can be interpreted through a variety of remote sensing techniques and sensors (Global Land Cover Facility 1998; Instituto Nacional de Pesquisas Espaciais 2000; Saatchi et al. 2000; Skole and Tucker 1993). A growing body of work indicates that the future state of much of Amazônia depends on the land use to which converted areas are put (Moran et al. 2000), but land use in the tropics is difficult to assess from remotely sensed data alone (Hansen et al. 2000).

In an earlier paper (Cardille 2002), we presented new maps of cropland, natural pasture, and planted pasture for the entire Amazônian region for the mid-1990s. These new snapshots of land use were developed from a statistical fusion of satellite data and agricultural census data, an underutilized ancillary data source available across Amazônia. This creation of land-use maps, which have the spatial detail of a satellite image and the attribute information of an agricultural census, required the development of a new statistical technique for merging data sets at different scales and of fundamentally different data types.

The new technique fused census and classified satellite data using their jointly considered statistical properties and, by creating the first basin-wide satellite-based agricultural land-use maps, gave researchers a new tool for understanding this rapidly changing region. Unlike algorithms that use ancillary data to aid existing classification strategies, this new method uses polygon-based land-use totals as area-averaged training set data for a new classification algorithm. Its effect is to explore how land used for agriculture had likely been labeled during the land-cover data production process. By exploiting the statistical relationship between the labeled land-cover classes and agricultural totals, we created a new method for extracting useful information from polygon data.

Because polygon and raster data are of fundamentally different types (Worboys 1995), they contain substantially different information, and there are a limited number of strategies for simultaneously considering data of these contrasting types for the same area. The use of geographic information systems for simple spatial overlays of census-derived and satellite-derived variables can be informative (e.g., Fung and Siu 2000; Imhoff et al. 1997; Radeloff et al. 2000; Wright and Boag 1994). With the dramatic increase in satellite-based raster data and the need to produce pixel-by-pixel classifications using discrete categories, polygon data sets have been applied in each of three major stages of common image classification techniques (de Bruin and Molenaar 1999; Lillesand and Kiefer 2000). These include using polygon data to restrict the area under consideration as a part of image pre-processing (Lillesand and Kiefer 2000); to operate as an additional data channel during the operation of the classifier (e.g., Ricchetti 2000) (Lo and Faber 1997); and to aid in identifying spectral clusters during the post-classification phase (Ma et al. 2001, Vogelmann et al. 1998). Polygon information has also been used in a modified classification algorithm, as prior probability data in a maximum likelihood classifier (Bobo 1997; Gorte 1999; Maselli et al. 1992; Strahler 1980).

Confronted with existing census data and land cover classifications, Ramankutty and Foley (1998) developed a fusion technique that assumed a linear relationship between reported census cropland and crop categories from the IGBP global land-cover classification. Although the agricultural land-use map was successful over North America and Europe, poor results in developing countries suggested that relatively coarse census data, the difficulty of separating agriculture from background phenology, and highly generalized land-cover classification legends cannot be easily reconciled. To generalize this approach, we have developed a more general nonlinear technique that does not assume that spectral clusters have been labeled with the correct land use during the classification process. Instead, the method explores the land-cover classification's information content to determine dynamically what categories are most highly related to the polygon land-use data across the study area.

In order to develop this method we needed to answer the following questions: (1) What is an appropriate conceptual model for merging polygon-based land-use data with raster-based land-cover data, and which data sets should be fused? (2) What statistical tools can represent the relationship between land use and land cover? (3) What are the relevant representations of candidate models, and how are they evaluated? (4) How is an appropriate scale of prediction determined, and how can predictions be made at that scale? (5) How do we assess the quality of the resultant data sets? This paper addresses these questions for the Amazônia region, applying these techniques and tests to these example data sets and outputs.

2 Study area and available data

The study region chosen to illustrate this technique is the area within the hydrological borders of the Amazon and Tocantins river drainage basins (Fig. 1), as derived and presented in Costa et al. (2002). Covering large percentages of four countries (Table 1), this vast region of tropical forest, savanna, and river ecosystems spanning most of northern South America has become one of the most rapidly changing regions in the world. Beginning in the 1970s, rapidly escalating land conversion for agriculture—particularly for cattle use—began to transform the landscape into an area increasingly influenced by human activities (Dale et al. 1993; Fearnside and Guimaraes 1996; Guild et al. 1998; Lucas et al. 1996; Moran et al. 1996, 2000; Nepstad et al. 1991; Pichón 1997; Uhl et al. 1988; Vosti et al. 1998).

2.1 Agricultural census data

Agricultural censuses, an underutilized ground-based data source, gather information about many aspects of agricultural activity, including crop type and pasture extent, harvest and yield, animal counts, and farm size. Using data derived from the agricultural censuses from four countries and spatially explicit political borders (Fig. 1), we created a mid-1990s map of cropland and pasture density (Fig. 2). The polygons for which data were reported were either two or three political divisions below the national level, and are referred to here as an "administrative unit."

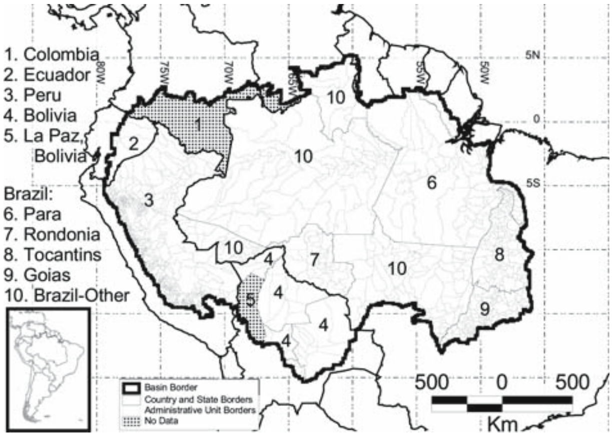

Fig. 1 Study area, political boundaries, and administrative unit framework for polygon-raster data fusion method. Within the study area, no ground-based data was available for this time period for Colombia or La Paz, Bolivia

Table 1 Agricultural census data sources and attributes

Country	Source	Median administrative area (ha)	Total area (10^6 ha)
Bolivia	INE 1990, FAO 2000	266.680	68.7
Brazil	IBGE 1997	63.195	454.2
Colombia	N / A	N / A	33.4
Ecuador	INEC 1995	56.563	12.9
Peru	INEI 1995	14.784	94.3

Administrative area amounts are from the referenced census source; total area amounts are for that part of the country within the study region

Abbreviations FAO: United Nations Food and Agriculture Organization; IBGE: Fundação Instituto Brasileiro de Geografia e Estatística; INE: Instituto Nacional de Estadística; INEC: Instituto Nacional de Estadística y Censos; and INEI: Instituto Nacional de Estadística e Informática

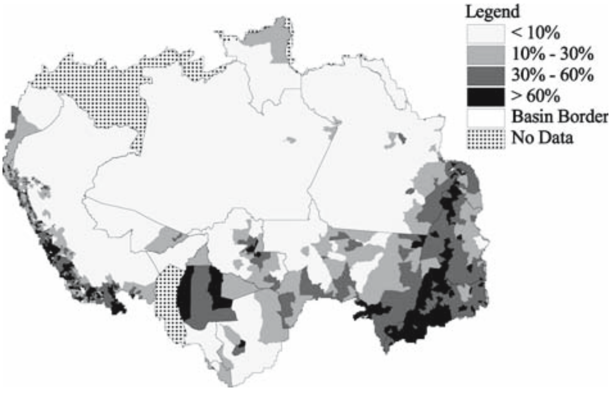

Fig. 2 Total cropland and pasture density in 1995 as reported in agricultural census data

2.2 Land-cover classifications

Two recent characterizations of land cover have expanded our knowledge of dominant land cover in Amazônia: the UMD (Hansen et al. 2000) and IGBP (Belward and Loveland 1996) global land-cover classifications (Fig. 3). These data sets, derived using monthly data from 1992 to 1993 from the Advanced Very High Resolution Radiometer (AVHRR), classified each 1-km pixel in Amazônia into one of either 13 (in the case of UMD) or 17 (in the case of IGBP) land-cover/land-use types.

Although these products classified nearly identical data into similar sets of classes the inclusion of band data for the UMD classification, ancillary data, and differences in production methods resulted in images that substantially disagree both globally (Hansen and Reed 2000) and over Amazônia (Table 1). This disagreement was especially acute outside of the Evergreen Broadleaf Forest core: in regions where census reports suggest high human land use such as in the eastern part of the study area (the Brazilian states of Pará, Tocantins, and Goiás), there is extreme disagreement between the land-cover classifications (Fig. 3).

In part because they were designed as land-cover products, land use is not well represented in either the IGBP or UMD classifications. Although both included Cropland as a category, the amount and location of this critically important land use differed substantially in the two classifications (Table 2). The Cropland/Natural Vegetation Mosaic category, which constitutes 17.3% of the IGBP classification in Amazônia, was not included in the UMD product. Perhaps more significantly, neither classification included Pasture, the dominant land use in the basin (Dale et al. 1993; Fearnside and Guimaraes 1996; Lucas et al. 1996; Moran et al. 1994, 2000).

In many areas, it appeared that many agricultural areas were categorized as a land cover consistent with that land use—for example, Grassland or Wooded Grassland. However, disagreement between the classifications made it difficult to simply adopt a single category—e.g., Grassland—as a direct proxy for cropland or pasture. This was consistent with a global comparison of the two data sets that found low per-pixel agreement among individual classes (Hansen and Reed 2000). Minimal ground truthing over Amazônia in both classifications (Hansen and Reed 2000) prevented the clear adoption of one of the products over the other.

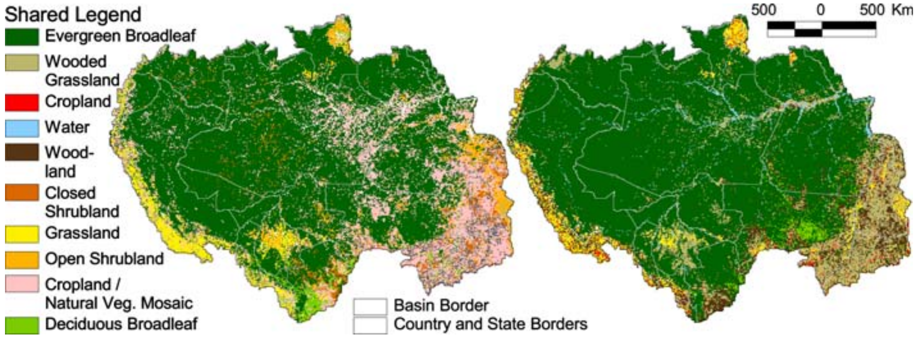

Fig. 3 IGBP DISCover (left) and Global Land Cover Facility (right) land cover classifications for the Amazon and Tocantins basins. Classifications are drawn with the same color legend

Table 2 Per-pixel analysis of comparable classes from the UMD and IGBP land cover classes in Amazonia

Classification	Evergreen broadleaf forest	Woody savanna (Woodland)	Savanna (Wooded grassland)	Cropland	Grassland	Cropland/Nat. Veg.	Other categories
IGBP proportion	64.3	2.7	4.5	2.7	3.6	17.3	4.9
UMD Proportion	69.9	8.0	13.0	1.0	3.4	–	4.7
Tau$_p$	high	low	low	low	low	–	-

Because categories' proportions could be similar but their locations quite different, the Tau$_p$ statistic (Ma and Redmond 1995) was used to quantify the agreement. The land cover classifications agrees well on the distribution and abundance of Evergreen Broadleaf Forest, but disagreed for all other categories. Category descriptions and names varied slightly; Hansen et al. (2000) was used to identify the comparable categories

3 Methods

3.1 Conceptual model

Although agricultural census data was clearly an extremely valuable quantification of land use, its spatial resolution was too coarse to provide a clear view of agricultural activity across Amazônia. The Brazilian state of Rondonia, for example, had 40 *municipios* (roughly analogous to U.S. counties) in the 1995 census, with a statewide density of agricultural activity of 14% (Fig. 2). Satellite imagery, however, indicated that this agricultural activity was not uniformly spread throughout the region, but appeared to be limited to cleared areas within each *municipio* (Fig. 4b). The five-minute spatial scale (about 9 km × 9 km at the equator), intermediate between the coarse census data and fine satellite data, suitably captures most of this within-unit variation (Fig. 4c). This was selected for several reasons. First, the median size of a county unit was near 625 km^2, implying a characteristic length on a side of about 25 km, while the pixel size of the land cover data was 1 km. The 5-minute scale was nearly midway between these two scales in the length dimension. Second, the 5-min scale has the substantial practical benefit of matching the input requirements of many regional ecosystem models.

Our conceptual model (Fig. 5) was based on the observation that since both the agricultural census data and satellite data considered the same region during nearly the same time period, the two data sources should be related statistically. A visual inspection of the land-use (Fig. 2) and land-cover (Fig. 3) data indicated that areas with a low amount of Evergreen Broadleaf Forest, for example, tended to have more cropland and pasture, and that areas with substantial cropland or pasture density appeared to have been frequently labeled as Wooded Grassland or Grassland. We hypothesized that within an administrative unit, there was a relationship between the density of agricultural activity and the proportions of that unit classified as various

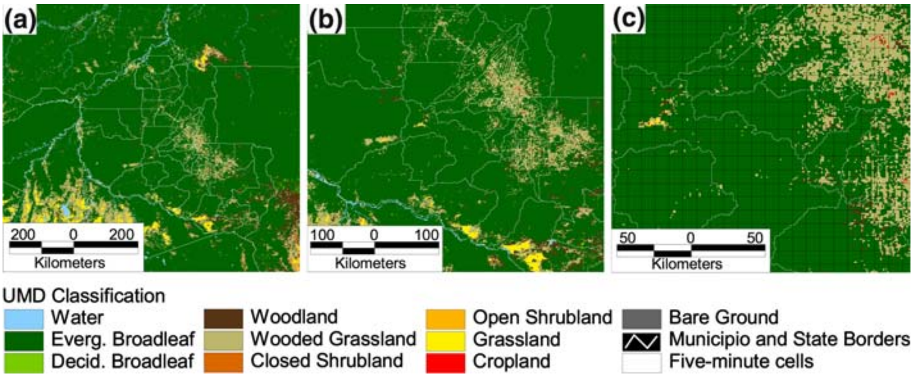

Fig. 4 Land cover assessment at three different scales in the Brazilian state of Rondonia. Areas not classified as Evergreen Broadleaf Forest are believed to be agricultural activity. Panel (a) shows the classification for the entire state at 1:8M scale and is clearly correlated with agricultural census data. Panel (b), at 1:4M scale, shows that agricultural activity is not uniformly spread across each municipio, but instead varies substantially both across and within municipios. Panel (c), at 1:75M scale, shows that five-minute (9 km) cells capture the spatial pattern of agricultural activity; this is the final resolution of the agricultural activity data set

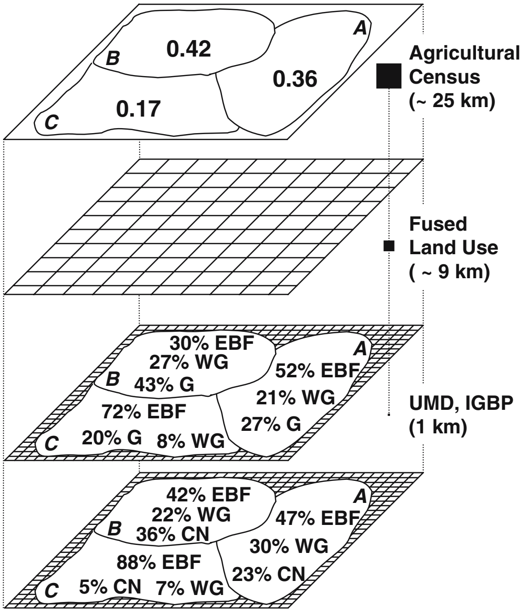

Fig. 5 Conceptual model of relationship between agricultural census and land cover from satellite classifications. For illustration purposes, UMD and IGBP classifications are simplified to consist of only three categories: EPF = Evergreen Broadleaf Forest; WG = Wooded Grassland; G = Grassland; CN = Cropland / Natural Vegetation Mosaic. Borders of hypothetical administrative units "A", "B", and "C" are shown in each layer. Spatial resolution of data sets are illustrated, including the resulting fused land use product

land covers, and that a fusion based on this statistical relationship could combine the attribute information of the land-use data with the spatial detail of the land-cover data.

3.2 Regression tree implementation

The relationship between land-use and land-cover data in Amazônia can be thought of as the following regression-like function:

$$\mathrm{y}_i = \mathrm{g}(\mathrm{igbp}_{i1}, \mathrm{igbp}_{i2}, \mathrm{igbp}_{i3} \ldots \mathrm{igbp}_{im}, \mathrm{umd}_{i1}, \mathrm{umd}_{i2}, \mathrm{umd}_{i3} \ldots \mathrm{umd}_{in}),$$

where y_i = proportion of unit i that is cropland or pasture; igbp_{ij} = proportion of unit i that is class j in the IGBP classification; umd_{ik} = proportion of unit i that is class k in the UMD classification; m = number of categories in the IGBP classification scheme; and n = number of categories in the UMD classification scheme.

To characterize g, the complex relationship between the cropland-pasture census proportion and each classification's category proportions, we used regression tree analysis (Breiman 1984; De'ath and Fabricius 2000). Regression trees have received wide attention in recent years as a tool for exploring relationships among ecological variables (Iverson and Prasad 1998; Lamon and Stow 1999; Michaelsen et al. 1994; Plant et al. 1999; Prince and Steininger 1999; Rathert et al. 1999) and have been used to develop remote sensing classifications (DeFries et al. 1997; Friedl et al. 1999; Hansen et al. 1996, 2003; Lawrence and Wright 2001).

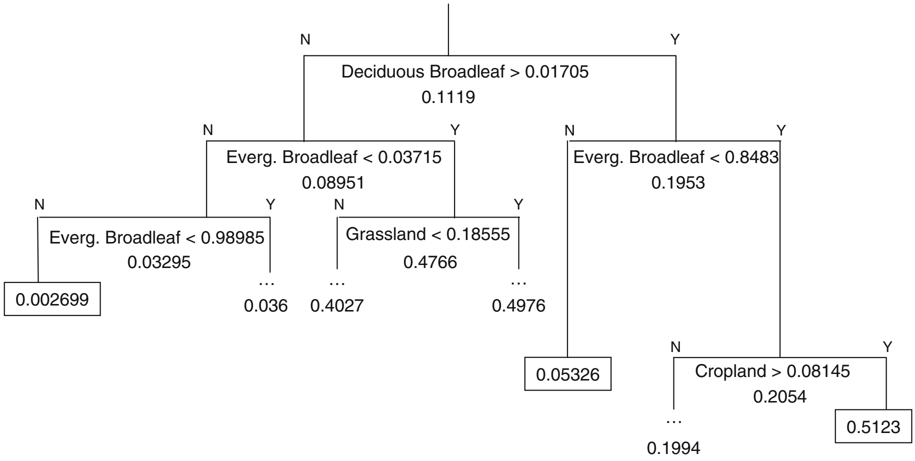

Fig. 6 Portion of a regression tree fitted from land use and land cover data in Amazonia. Decisions at each node split are based on the fraction of the area categorized in the land cover classification as the given type. Values under the node split are the mean agricultural density value of all cases in that subtree. For example, the mean agricultural density value of all cases is 0.1119; the mean value of all cases having Deciduous Broadleaf fraction $<$ 0.01705 is 0.08951, etc. Terminal nodes of the tree are shown within a box; other subtrees are truncated for display, with the mean value of the subtree shown. Trees were built using administrative units of Brazil and Peru as cases

Land-use and land-cover data within an administrative unit formed each case, and cases were weighted in the tree-building algorithm according to their land area. In the resulting trees, node splits were labeled with the proportion of a particular land-cover category, while each leaf of the tree was the mean cropland-pasture value of those administrative units having the specified land-cover conditions (Fig. 6). Administrative units from Peru and Brazil were used to create candidate trees; of the more than 1,300 cases, half were reserved for creating candidate tree models, and half reserved for model evaluation. Only census data from Brazil and Peru were used to form trees because they were reported in small administrative units and were nearly contemporary with the satellite data. Models and maps that included and excluded Bolivian data were indistinguishable from those presented here and are not shown. We developed three candidate tree models for evaluation: Model 1 used the 17 predictors from the IGBP classification, Model 2 used the 14 categories of the UMD classification, and Model 3 incorporated both the IGBP and UMD classifications for fusing the census and satellite data.

3.3 Prediction using candidate models

To produce a map of fused land use and land cover across the entire basin on a regular grid, we first reinterpreted the satellite classifications with respect to a five-minute (approximately 9 km × 9 km) grid. Each 5-min cell contained an average of 81 1-km land-cover cells (Fig. 4 and 5), and we used the land-cover cells to calculate the proportion of the cell's area belonging to each land-cover category from both the UMD and IGBP classifications.

We then used the land-cover proportions and the regression tree relationship between land use and land cover (derived at the administrative unit scale) to estimate

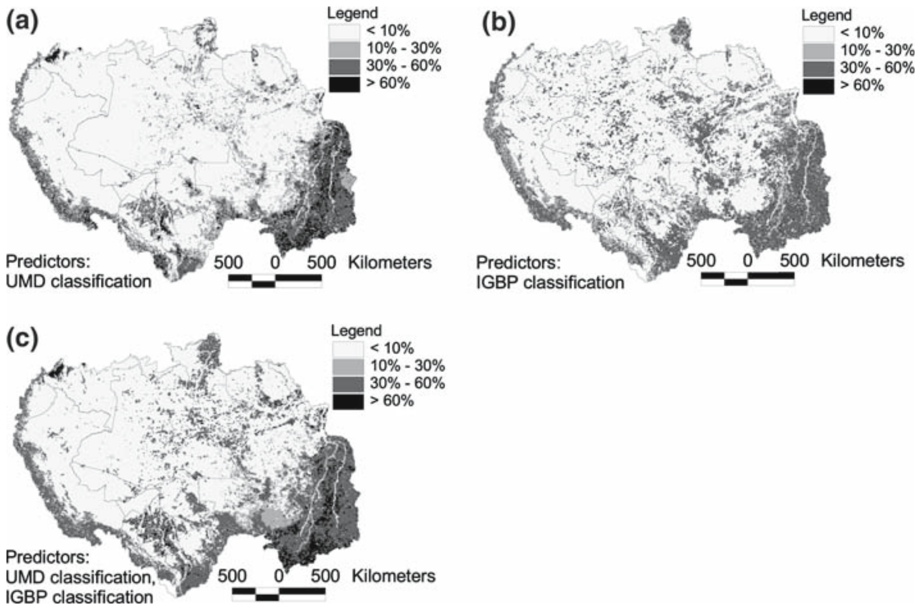

Fig. 7 Maps of mid-1990s density of agricultural land use in Amazonia using three models to fuse land use and land cover information. Panel (a): categories from UMD classification were used to relate census and satellite data. Panel (b): census data fused with IGBP categories. Panel (c): census data fused with both IGBP and UMD categories

the density of agricultural activity in each 5-min cell in Amazônia. We assumed that the statistical relationship between land use and land cover, although complex, was consistent across the entire study area. This allowed us to estimate agricultural activity in remote or dangerous areas such as Colombia, for which there was satellite information but no census data. We used the 5-minute maps of land-cover proportions and the three regression tree relationships to create three candidate surfaces of land use in Amazônia for evaluation (Fig. 7).

3.4 Evaluation methods

Because there was no existing basin-wide data set of land use in Amazônia against which we could compare the maps created from candidate trees, we were unable to use standard image classification assessment techniques, such as a confusion matrix (Lillesand and Kiefer 2000), to evaluate model quality. Instead, we developed tests at two different scales to evaluate the ability of each tree model to capture elements of both the land-cover and land-use data.

3.4.1 Evaluation at 5-min scale

At the 5-min scale, we explored the simple univariate relationship between the fitted agriculture maps (Fig. 7) and the individual categories of the land-cover classifications (Fig. 3). This simple test allowed us to evaluate, at the scale of prediction, our hypotheses that land use and land cover were related in a plausible and predictable way, that regression trees could capture that relationship, and that it was appropriate to project that relationship between spatial scales.

For each 5-min cell, the proportion taken up by each land-cover category was computed; we determined the simple correlation between these category proportions and the cropland-pasture proportions for each cell from the three candidate models. This allowed us to assess which land covers were most strongly associated with cropland and pasture; by our hypothesis, these land-cover categories were most often used to label areas used for cropland and pasture in the study region.

3.4.2 Evaluation at administrative unit scale

Since the models were created at the administrative unit scale but used at the 5-min scale, we developed a test that would assess how well each candidate model reproduced the agriculture proportions seen in each administrative unit. To compare the performance of the three trees at the administrative unit scale, we aggregated the 5-min cells within each administrative unit kept in reserve for validation, computing the mean fitted cropland-pasture density for each unit. These values of mean agricultural activity were then compared to the published agricultural census data (Fig. 2) through a linear regression using each unit's land area as the case weight. In addition to computing the regression parameters and correlation coefficient, we investigated the resulting studentized residuals for evidence of spatial pattern. This test allowed us to evaluate each tree's ability to capture variation in the agricultural census data, as well as well as to assess spatial variation across Amazônia in the quality of the fit.

4 Results and discussion

4.1 Model evaluation

4.1.1 Evaluation at 5-min scale

For each of the three models, this univariate comparison revealed a systematic relationship between land-use data and land-cover categories in Amazônia (Table 3). In all three models, there was a strong negative correlation between agricultural activity and Evergreen Broadleaf Forest (Table 3), consistent with the observation (e.g., Mittermeier et al. 1999) that much agriculture occurs in regions of former cerrado in the eastern part of the study area that have little or no forest.

The regression models also indicated a systematic relationship between certain land-cover classes and agricultural activity in Amazônia. Most strikingly, Models 2 and 3 indicated that cropland and pasture might have been systematically classified as Wooded Grassland in the UMD classification (Table 3), a plausible confusion. In models incorporating the IGBP classification, fitted agricultural activity was most strongly correlated with the Cropland/Natural Vegetation Mosaic category, as would be expected. Woodland and Wooded Savanna, two land-cover categories we might also expect to be used to classify cropland and pasture land uses, also showed a moderate positive relationship.

4.1.2 Evaluation at administrative unit scale

Regression results indicate that at the administrative unit scale, the fusion process generated tree models that captured much of the variability of the agricultural census

Table 3 Univariate relationship between fitted agricultural activity of each candidate model and per-category fractions from the UMD and IGBP classifications

UMD Class Name, IGBP Class Name	Model 1 (IGBP)	Model 2 (UMD)	Model 3 (UMD, IGBP)
Evergreen Broadleaf Forest	**−0.73**	**−0.81**	**−0.75**,**−0.66**
Deciduous Broadleaf Forest	0.11	0.01	0.05, 0.14
Mixed forest	0.01	0.00	0.00, 0.01
Closed shrub	0.06	0.05	0.10, 0.11
Open shrub	0.21	0.07	0.06, 0.14
Woodland, Wood savanna	**0.36**	**0.52**	**0.49, 0.30**
Wooded grassland, Savanna	0.28	**0.81**	**0.74**, 0.26
Grassland	0.24	0.26	0.23, 0.21
Permanent wetlands	0.02	–	–, 0.02
Cropland	**0.31**	**0.30**	0.26, **0.41**
Urban	0.03	–	–, 0.02
Cropland/Natural vegetation	**0.54**	–	–, **0.44**
Barren	0.04	0.05	0.03, 0.04
Water	−0.17	−0.12	−0.14,−0.14

Land cover fractions were computed for each five-minute cell within the study area and compared to the per-cell predicted agricultural activity of each regression tree model. Correlations above 0.30 are shown in bold

data. The map created from the tree built using only IGBP data (Model 1) captures much of the variability in the census ($r = 0.69$), while results from Model 2 were a much closer fit to census data ($r = 0.81$). The model built using both IGBP and UMD data (Model 3), although it was fitted using more predictor variables, was not as strong a fit ($r = 0.80$) to census data as the model built using only UMD data.

4.2 Model selection

Although the three tree models produced fitted maps of agricultural activity that did not appear dramatically different to the naked eye (Fig. 7), it is clear that of the UMD and IGBP classifications, the IGBP classification showed a substantially weaker relationship to census data. When only information from IGBP categories was considered (Model 1), the resultant fitted map was less strongly related to ground-derived agricultural activity (as measured against validation data at the administrative unit scale) than the fit from the other two models. When the model using only categories from the UMD classification (Model 2) was compared to the model using both UMD and IGBP categories (Model 3), the results were extremely similar, whether measured at the 5-min. scale (Table 3) or via regression against census data at the administrative unit scale. These results suggest that the UMD classification provided the most information when mapping land use in Amazônia, while the IGBP classification did not offer a substantial improvement.

Although the model using both classifications could be said to have provided an equally good fit to the census data, we judged that the benefit, if any, of including the 17 categories from the IGBP classification was outweighed by the cost of additional data preparation and model complexity. We chose the model using only UMD categories (Model 2, Fig. 7) as that which best fused land-cover data with agricultural census data for Amazônia.

4.3 Model quality: spatial variation in fit to census data

Studentized residuals from the comparison of agricultural census data and the aggregation of data from the fused product to the administrative unit scale indicated a very good fit. Few studentized residuals (3.8%) were outside the two-standard deviation range, and 1.9% were larger than three standard deviations. Maps of the residuals showed several interesting spatial trends (Fig. 8). Fitted values from the fusion were higher than the original census data throughout the Brazilian state of Amazonas and most of non-Andean Peru. Western Para was like neighboring Amazonas in that the fusion indicated more agricultural activity than reported the census, but Eastern Para, a known area of high agricultural activity, was under-predicted by the model. The same is true in Eastern Acre, an area of significant recent deforestation, where results indicate that the model under-predicted agricultural activity.

Spatial trends in the studentized residuals raise interesting questions about the source or sources of discrepancy. In particular, in remote areas the fused product indicates more agricultural activity than that reported in the census, but it is unclear whether this is due to an actual under-reporting of census data or whether the fusion method incorrectly interprets some land covers as agriculture. In these areas, it seems quite likely that the remoteness of the areas prevents census-takers from fully accessing farmers and ranchers. In the opposite scenario where the fitted data was lower than census values, it is also unclear whether farmers in Goias, for example, had an incentive to systematically over-report agricultural activity, or whether the fusion method simply failed to recognize actual agriculture. Further complicating error analysis is the use, in some areas, of multiple cropping seasons. Although the spatial trends are clear, they fortunately involve mostly small values: the low number of significantly large residuals indicates that the fusion of census data and satellite imagery was quite successful.

4.4 Scaling issues

The fusion method described here introduces several interesting scaling issues. Unlike in typical regression applications, model creation, model use, and model validation all

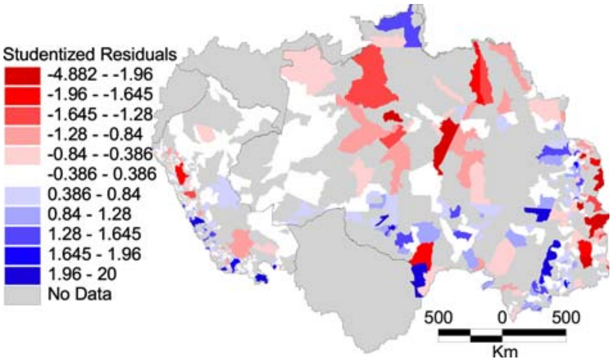

Fig. 8 Studentized residuals from Model 2. Maps of studentized residuals from each of the three models were similar and not shown here. Administrative units with negative values indicate the model overpredicted the recorded agricultural census data there; positive values indicate underprediction relative to census data. Since only Peru and Brazil units were used to build the regression tree models, studentized residuals were computed for only those countries

involve changes in spatial scale. It is important to remember that the model is created at the administrative unit scale; that is, we determined the relationship between census data and classified imagery using each administrative unit as a single case.

The statistical model is built at the administrative unit scale but used at a finer scale, and can thus be thought of as "scaling down" census data. That is, the algorithm distributes the coarse census data within the administrative unit, using what is known about the relationship between imagery and the census. It should be noted that within each administrative unit, the amount of agriculture in the fitted model is not constrained to be exactly the same as that in the census data; instead, it is allowed to vary from census results according to the underlying relationship. This scaling down preserves the basin-wide census totals, while effectively distributing the amounts among 5-min scale grid cells.

Questions of pruning regression trees are closely related to these issues of scale changes. In this method the unpruned trees were transferred between the administrative unit scale and the 5-min scale. Since the response set of regression trees is discrete, pruning tree nodes would result in a fused map having only a few possible levels of agricultural activity. By transporting the full regression trees without pruning, we take the largest amount of explanatory power in each model from the administrative unit scale to the 5-minute scale. The result is a map that, although having discrete levels, appears to be nearly continuous.

Since the scales of model creation and model testing differed we found that unlike in a typical regression, adding more predictor variables at the model building stage did not necessarily create a "better" model. In particular, when we added the predictors from the IGBP classification to the UMD predictors (forming Model 3), we did not get a model that was substantially better and was, in fact, arguably worse due to apparent noise (Fig. 7). That model quality is not simply a function of the amount of data suggests the need for tests at multiple scales to help choose among candidate models.

For the demonstration described here, the choice of the resulting scale was relatively easy since the scale halfway between the two sources of data was one commonly used for regional models. In situations where a scale coarser or finer than that near this midpoint were desirable, we expect that choosing a scale nearer to one data source than another would produce a resulting set with more characteristics of the nearer source. The development of objective measures of similarity at multiple scales are an area of active research, and tests of these measures may be helped with data sets of the form presented here.

4.5 Verification

In addition to technical questions of model quality measured at the scale of the administrative unit, there is the more fundamental question of verification of the resulting maps. The basin's size and the remoteness of many areas discourages employing the traditional concept of accuracy assessment of classified satellite imagery through visiting ground control points. If other maps of agricultural land use existed for this area, we could undertake a pixel-by-pixel comparison of the two data sets across the entire study region: since none exists, we are limited to alternate verification strategies.

A key directive of the Large-Scale Biosphere-Atmosphere Experiment in Amazônia (Nobre et al. 1997) is the development of new classified and unclassified imagery at multiple resolutions for the Brazilian part of the study area. Analyses using Landsat data, for example, over a relatively small part of the region could prove invaluable to the verification effort. For example, Alves et al. (1999) have used Landsat TM imagery to demonstrate the high levels of deforestation in a small part of Rondonia during the mid-1990s. Although agreement in this relatively small part of the study area could not completely verify the performance of the algorithm, a closer investigation could raise interesting questions about algorithm behavior, imagery dates, and land-cover/land-use comparability.

4.6 Caveats and algorithm improvements

Although the algorithm described here is quite effective in its ability to fuse census and satellite data, there are several caveats that should be considered when applying in other circumstances:

- If the land-cover classification categories are too general or error-prone, the quality of the fusion could be lowered. An example of this can be seen in Model 1, which used IGBP categories as its land-cover data set. The Cropland/Natural Vegetation category was correlated with cropland and pasture census data; however, the category appears to have been general enough that some non-agricultural pixels were included. This is reflected in the relatively low quality of the fit between IGBP categories and census data. A different interpretation of this low correlation, however, is also important to consider. It is possible that the IGBP classification is quite accurate, and that both census and UMD data were systematically biased toward only capturing high-density agriculture, while the IGBP product did not miss low-density activity. We would suggest that in the absence of ground truth information (as is the case with both the UMD and IGBP classifications), it seems more conservative to assume that the agreement between the UMD and census data is indicative of agreement between high-quality data sets, rather than indication of the same systematic error. For this particular group of data sets, it might be worthwhile to investigate the fit of the IGBP classification using the Seasonal Land Cover Regions classification scheme, which has more than 100 categories.
- It is important to realize that the scale of satellite observations may be too coarse to capture small-scale agriculture in this region. Because many pioneer areas involve deforestation finer than 1 km × 1 km, there could be significant agriculture missed by the satellite sensor. In this case, we could suggest that a finer-scale classification, if available, would allow a fusion with more accurate agriculture estimates. In this case, however, the only finer-scale classifications had too few categories to fuse appropriately with census data.
- It is difficult to assess the "noise" in the model. Because there is no ground-truth data for the classifications, it is difficult to objectively assess the balance between under- and over-prediction for a candidate model. If finer-scale classifications of agriculture were readily available over smaller areas, these could be used as ground truth data to discriminate among candidate models.
- This regression-tree-based technique produces a map with a discrete set of values. Depending on the amount of training data used to build the model, there may be too few tree nodes to create a useful map. For this example, the spatial resolution of the census data and the extent of the study area allowed a set of more than 50 nodes. A fitting technique with a continuous response (e.g., linear or logistic regression) could be considered, especially in situations where there are fewer training units available.

- The conceptual model is easily adaptable to include additional factors that may influence agricultural density, such as elevation, soil characteristics, or population. The conceptual model described here was designed to explore what land-cover categories were used to label areas of agriculture. Users who believe this land-cover labeling may have varied spatially according to other mapped factors, or who are interested in exploring what factors influence the density of agriculture, can easily extend this conceptual model.

5 Conclusion

This technique has allowed us to merge the best elements of two high-quality data sets. Although they are not highly spatially detailed, agricultural censuses provide the best ground-based information available about agricultural activity across this vast, data-poor area. Satellite data products can distinguish among land covers, yet lack detail about the use of human-controlled lands. The fused product has elements of both; the attributes of census data are distributed within administrative units, scaled down using spatial attributes of satellite imagery.

In this study, the land-cover classes most strongly associated with land-use data were quite plausible and reflected known difficulties of the land-cover classification process. An additional goal of this study was to examine whether, in this region, one of the two existing whole-basin land-cover classifications was a superior fit to census data. We developed models and tests that explore the ability of these three combinations to be fused with land-use data.

Both census data and satellite imagery contains errors, and analyzing the regression at the administrative unit scale allows us to hypothesize about possible biases in the data. This method may lead to a clarification of those regions where better census data are needed, as well as identifying areas where the satellite classifications are particularly prone to error.

The validation test that we devised compares census data kept in reserve to fused data re-aggregated within administrative units. That the fit is strong even after transporting the statistical relationship across scales and then re-summarizing is an indication of the clear, previously unrevealed relationship between land cover, census data, and land use in Amazônia.

This method appears to have general applicability. It can be used, for example, to better locate the spatial distribution of cities from the fusion of population census data and classified imagery, or to relate historical land-use records to a land-cover legacy in a region. We hope this method will be adopted and refined by others seeking to fuse attributes from any polygon-based data set with a finer-scale classified raster image.

Acknowledgements Jon Foley provided perspective on the modeling technique, and Navin Ramankutty encouraged the idea further. This work was funded by EPA STAR Fellowship Grant U-915332 and LBA-ECO, a component of the Large-Scale Biosphere-Atmosphere Experiment in Amazônia (LBA). We thank Jon Foley, Monica Turner, Frank Scarpace, Tom Gower, and Kevin McSweeney for their helpful comments on an earlier version of this manuscript.

References

Alves DS, Pereira JLG, De Sousa CL, Soares JV, Yamaguchi F (1999) Characterizing landscape changes in central Rondonia using Landsat TM imagery. Int. J. Remote Sens 20:2877–2882

Belward AS, Loveland TR (1996) The DIS 1 km land cover data set. In Global change newsletter, vol 27. The International Geosphere-Biosphere Programme: A study of global change (IGBP) of the International Council of Scientific Unions, pp 7–8

Bobo MR (1997) Incorporation of categorical ancillary data into multispectral image classifications using a modified Bayesian decision rule. M. S. Thesis University of Wisconsin-Madison

Breiman L (1984) Classification and regression trees. Wadsworth International Group Belmont Calif.

Cardille JA (2002) Characterizing patterns of agricultural land use in Amazonia by merging satellite imagery and census data. Ph.D Thesis, University of Wisconsin-Madison

Costa MH, Henrique C, Oliveira C, Andrade RG, Bustamante TR, Silva FA, Coe MT (2002) A macrohydrological dataset for the Amazon basin. J Geophys Res 10.1029/2000JD000309

Dale V, O'Neill R, Pedlowski M, Southworth F (1993) Causes and effects of land-use change in central Rondonia Brazil. Photogramm Eng Remote Sens 59:997–1005

de Bruin S, Molenaar M (1999) Remote sensing and geographical information systems. In: Stein, A et al. (eds) Spatial statistics for remote sensing Kluwer Academic Publishers, Dordrecht, pp 41–54

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192

DeFries R, Hansen M, Steininger M, Dubayah R, Sohlberg R, Townshend J (1997) Subpixel forest cover in central Africa from multisensor, multitemporal data. Remote Sens Environ 60:228–246

Fearnside P, Guimaraes W (1996) Carbon uptake by secondary forests in Brazilian Amazonia. For Ecol Manage 80:35–46

Friedl MA, Brodley CE, Strahler AH (1999) Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Trans Geosci Remote Sens 37:969–977

Fundação Instituto Brasileiro de Geografia e Estatística. (1997) Censo agropecuário: 1995–1996 IBGE, Rio de Janeiro

Fung T, Siu W (2000) Environmental quality and its changes, an analysis using NDVI. Int J Remote Sens 21:1011–1024

Global Land Cover Facility (1998) Results of the NASA landsat pathfinder humid tropical deforestation project. Geography Department, University of Maryland, College Park, MD USA

Gorte B (1999) Supervised image classification. In: Stein A et al (eds), Spatial statistics for remote sensing Kluwer Academic Publishers, Dordrecht, pp. 153–163

Guild LS, Kauffman JB, Ellingson LJ, Cummings DL, Castro EA (1998) Dynamics associated with total aboveground biomass C, nutrient pools, and biomass burning of primary forest and pasture in Rondonia Brazil during SCAR-B. J Geophys Res-Atmos 103:32091–32100

Hansen MC, Reed B (2000) A comparison of the IGBP DISCover and University of Maryland 1km global land cover products. Int J Remote Sens 21:1365–1373

Hansen M, Dubayah R, DeFries R (1996) Classification trees: an alternative to traditional land cover classifiers. Int J Remote Sens 17:1075–1081

Hansen MC, DeFries RS, Townshend JRG, Sohlberg R (2000) Global land cover classification at 1 km spatial resolution using a classification tree approach. Int J Remote Sens 21:1331–1364

Hansen MC, DeFries RS, Townshend JRG, Carroll M, Dimiceli C, Sohlberg RA (2003) Global percent tree cover at a spatial resolution of 500 meters: first results of the MODIS vegetation continuous fields algorithm (available online). Earth Interact 7:1–15

Imhoff ML, Lawrence WT, Elvidge CD, Paul T, Levine E, Privalsky MV (1997) Using nighttime DMSP/OLS images of city lights to estimate the impact of urban land use on soil resources in the United States. Remote Sens Environ 59:105–117

Instituto Nacional de Pesquisas Espaciais (2000) Monitoramento do desflorestamento bruto da Amazônia Brasileira (Monitoring the Brazilian Amazon gross deforestation). Minsterio da Ciencia e Tecnologia, São José dos Campos, São Paulo, Brazil

Iverson LR, Prasad AM (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. Ecol Monogr 68:465–485

Lamon EC, Stow CA (1999) Sources of variability in microcontaminant data for Lake Michigan salmonids: statistical models and implications for trend detection. Can J Fish Aquat Sci 56:71–85

Lawrence RL, Wright A (2001) Rule-based classification systems using classification and regression tree (CART) analysis. Photogramm Eng Remote Sens 67:1137–1142

Lillesand TM, Kiefer RW (2000) Remote sensing and image interpretation, 4th edn. John Wiley & Sons, New York

Lo CP, Faber BJ (1997) Integration of landsat thematic mapper and census data for quality of life assessment. Remote Sens Environ 62:143–157

Lucas RM, Curran PJ, Honzak M, Foody GM, do Amaral I, Amaral S (1996) Disturbance and recovery of tropical forests: balancing the carbon account. In J H. C. Gash, et al. Amazonian deforestation and climate (eds) Wiley, pp. 383–398

Ma ZK, Redmond RL (1995) Tau coefficients for accuracy assessment of classification of remote-sensing data. Photogramm Eng Remote Sens 61:435–439

Ma ZK, Hart MM, Redmond RL (2001) Mapping vegetation across large geographic areas: integration of remote sensing and GIS to classify multisource data. Photogramm Engi Remote Sens 67:295–307

Maselli F, Conese C, Petkov L, Resti R (1992) Inclusion of prior probabilities derived from a nonparametric process into the maximum-likelihood classifier. Photogramm Eng Remote Sens 58:201–207

Michaelsen J, Schimel DS, Friedl MA, Davis FW, Dubayah RC (1994) Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J Veget Sci 5:673–686

Mittermeier RA, Mittermeier CG, Myers N, Robles Gil P (1999) Hotspots : earth's biologically richest and most endangered terrestrial ecoregions, 1st English edn. Cemex, Mexico City

Moran EF, Brondizio E, Mausel P, Wu Y (1994) Integrating Amazonian vegetation, land-use, and satellite data. Bioscience 44:329–338

Moran EF, Packer A, Brondizio E, Tucker J (1996) Restoration of vegetation cover in the eastern Amazon. Ecolo Econ 18:41–54

Moran EF, Brondizio ES, Tucker JM, da Silva-Forsberg MC, McCracken S, Falesi I (2000) Effects of soil fertility and land-use on forest succession in Amazonia. Forest Ecology and Management 139:93–108

Nepstad DC, Uhl C, Serrao EAS (1991) Recuperation of a degraded Amazonian landscape–forest recovery and agricultural restoration. Ambio 20:248–255

Nobre CA, Artaxo P, Becker A, Brown IF, Dolman H, Dunne, T, Gash JHC, Grace J, Janetos AC, Kabat P, Keller M, Krug T, Marengo JA, McNeal RJ, Prince SD, Silva Dias, PL Tomasella J, Victoria RL, Vörösmarty CJ, Wickland DE (1997) Large scale biosphere-atmosphere experiment in Amazonia (LBA) extended science plan. LBA Project Office

Pichón FJ (1997) Settler households and land-use patterns in the Amazon frontier: farm-level evidence from Ecuador. World Dev 25:67–91

Plant RE, Mermer A, Pettygrove GS, Vayssieres MP, Young JA, Miller RO, Jackson LF, Denison RF, Phelps K (1999) Factors underlying grain yield spatial variability in three irrigated wheat fields. Transactions of the ASAE 42:1187–1202

Prince SD, Steininger MK (1999) Biophysical stratification of the Amazon basin. Global Change Biol 5:1–22

Radeloff VC, Hagen AE, Voss PR, Field DR, Mladenoff DJ (2000) Exploring the spatial relationship between census and land- cover data. Soc Nat Resour 13:599–609

Ramankutty N, Foley JA (1998) Characterizing patterns of global land use: An analysis of global croplands data. Global Biogeochem Cycles 12:667–685

Rathert D, White D, Sifneos JC, Hughes RM (1999) Environmental correlates of species richness for native freshwater fish in Oregon USA. J Biogeogr 26:257–273

Republica de Bolivia: Instituto Nacional de Estadística (1990) II censo nacional agropecuario (1984): Resultados departamentales INE La Paz, Bolivia

Ricchetti E (2000) Multispectral satellite image and ancillary data integration for geological classification. Photogramm Eng Remote Sens 66:429–435

Saatchi SS, Nelson B, Podest E, Holt J (2000) Mapping land cover types in the Amazon Basin using 1 km JERS-1 mosaic. Int. J Remote Sens 21:1201–1234

Sistema Estadístico Agropecuario Nacional: Instituto Nacional de Estadística y Censos (1995) Encuesta nacional de superficie y producción agropecuarias de areas. INEC Quito

Skole D, Tucker C (1993) Tropical deforestation and habitat fragmentation in the Amazon - satellite data from 1978 to 1988. Science 260:1905–1910

Strahler AH (1980) The use of prior probabilities in maximum likelihood classification of remotely sensed data. Remote Sens of Environ 10:135–163

Uhl C, Buschbacher R, Serrao EA (1988) Abandoned pastures in eastern Amazonia. 1. Patterns of plant succession. J Ecol 76:663–681

United Nations Food and Agriculture Organization (2000) FAOSTAT agricultural data. FAO Computerized Information Series Statistics

Vogelmann JE, Sohl T, Howard SM (1998) Regional characterization of land cover using multiple sources of data. Photogramm Eng Remote Sens 64:45–57

Vosti SA, Witcover J, Carpentier CL (1998) Arresting deforestation and resource degradation in the forest margins of the humid tropics: policy, technology, and institutional options for western Brazil. International Food Policy Research Institute, Washington, DC

Worboys M (1995) GIS, a computing perspective. Taylor & Francis, Bristol, PA, USA

Wright GG, Boag B (1994) The application of satellite remote-sensing and spatial proximity analysis techniques to observations on the grazing of oilseed rape by roe deer. Int J Remote Sens 15:2087–2097

Biographical sketches

Jeffrey A. Cardille is an associate professor in the Département de Géographie at the Université de Montréal. He received his Ph.D. from the Institute of Environmental Studies at the University of Wisconsin-Madison in 2002 and completed his postdoctoral research there in September 2005. His research analyzes the connections between spatial patterns and ecological processes, often merging varied forms, sources, and qualities of spatial data.

Murray K. Clayton is Professor and Chair of Plant Pathology and Professor of Statistics at the University of Wisconsin-Madison. He received his Ph.D. in Statistics from the University of Minnesota in 1983. He conducts research on the application of statistical methods to problems in the agricultural, biological, and environmental sciences, with a particular focus on spatial statistics.