A SPATIAL ECONOMIC AND ENVIRONMENTAL COMPUTER PREDICTIVE MODEL OF LAND USE CHANGE IN THAILAND USING A STATISTICAL TREE CLASSIFICATION APPROACH

John FELKNER
Harvard University
Graduate School of Design
gsd97jjf@gsd.harvard.edu

Working Group VI/4

KEY WORDS: Spatial, Land Use Change, Tree Classification, GIS, Remote Sensing, Thailand, Predictive.

ABSTRACT

This approach integrates the tools of Geographic Information Systems (GIS) and remotely sensed imagery to map land use change in two Provinces of Thailand from 1979 to 1999, and then uses a statistical tree classification model to predict future change.

1   INTRODUCTION

As population growth, conversion of land to agricultural uses, urbanization and deforestation have continued to grow at every increasing rates – especially in developing countries – interest in and the need for quantitative empirical models of deforestation and urbanization has grown in recent years (Kaimowitz and Angelsen 1998).

The two Provinces used in this study are Chachoengsao and Sisaket, which were chosen because they are representative of both economic and environmental variation within Thailand as a whole. Measured land use change from 1979 to 1989 – mapped in a raster GIS environment – is used as the dependent variable, and eight spatial GIS models representing socioeconomic and environmental factors are used as the independent variable predictors. The validity of the resulting statistical models to predict future change are evaluated by GIS comparison with existing 1999 land use, derived from Landsat satellite imagery.

This approach complements other statistical approaches in spatial deforestation and urbanization models by using a statistical tree classification approach (Kaimowitz and Angelsen 1998). The tree approach has a number of distinct advantages over more traditional statistical approaches, primarily in that it does not require normalization of input variables and provides an automated approach both for distinguishing existing homogeneous partitions and creating a hierarchy of predictive variable priorities in large disparate datasets (Venables and Ripley 1994; O'Connor, Jones et al. 1996).

2   DESCRIPTION OF RESEARCH

2.1   Economic and Environmental Models

This predictive land use model uses a tree classification analysis performed in Splus software using maps of land use change from 1979-1989 as the dependent variable and eight socioeconomic and environmental models as independent input factors. The basic hypothesis is that future land use change can be explained best by a combination of economic and environmental factors rather than by either economic or environmental factors alone.

The eight factors are all in the form of continuous raster grid maps of each province. All have grid cells of identical spatial resolution (60 x 60 meters).

The environmental models are the following:

i.   topography (based on a Digital Elevation Model (DEM))

ii.   soil moisture index – using a computer GIS model that has as its inputs precipitation, soil hydrologic group maps and slope to derive an index representing relative soil moisture of each pixel;

iii.   proximity to timber resources – a GIS raster grid in which each cell in the model is assigned a value reflecting its relative proximity to existing timber resources (derived from Landsat satellite imagery);

iv.   proximity to water resources – a GIS raster grid in which each cell in the model is assigned a value reflecting its relative proximity to existing water resources (from satellite imagery).

The socio-economic models are:

i.   population potential – a continuous raster grid attempting to model sub-provincial spatial population variability on average for the 1980s, based on population village values from the Community Development Department (CDD) Survey of the Thai Government data (conducted every two years). The model assumes that population distribution will occur along infrastructure networks as a function of travel time along those networks;

ii.   proximity to markets – a raster grid in which each cell has a value representing its relative proximity to markets (defined as villages, towns, cities, and infrastructure linkages to national and international markets external to the Province), assuming that access to markets will occur along infrastructure networks as a function of travel time along those networks;

iii.   proximity to infrastructure – a raster grid in which each cell has a value representing its relative proximity to infrastructure networks. The model takes into account steepness of slope as an inhibitor to infrastructure access. It also assumes that areas of land adjacent to major infrastructure segments (major highways) have a higher proximity value that land equally close to less important infrastructure elements (dirt roads);

iv.   poverty measure – a raster grid in which each cell has a value representing a model of sub-Provincial spatial variation in wealth. the model uses data pertaining to income or other indicators of wealth from the CDD data at the CDD village level, and then assumes that wealth will be spatially distributed along infrastructure networks as a function of access/travel time along those networks.

The population and proximity to markets models distribute data values derived from the CDD village values spatially along the road network by using a negative exponential gravity model. Once these values are distributed throughout the extent of the road network, a continuous surface is achieved by using an Inverse Distance Weighted (IDW) spatial interpolation.

2.2   Land Use Change Detection

Land use change detection was obtained by processing 1979 and 1989 Landsat Multi-Spectral Scanner (MSS) and Thematic Mapper satellite imagery for Sisaket and Chachoengsao Provinces. 1979 and 1989 images were first classified using and unsupervised classification to derive maps of landcover for each Province for both 1979 and 1989.

Land use change detection from 1979 to 1989 was then obtained by performing an unsupervised classification on a single dataset containing data from both the 1979 and 1989 images. The resulting clusters were classified with reference to the 1979 and 1989 landcover maps. The assumption was that certain clusters would represent areas of change, as described in (Jensen 1996).

2.3 Land Use Change Prediction: A Statistical Tree Classification Approach

For both the independent factor maps of land use change 1979-1989 and the input environmental and socioeconomic input factors, the value of each grid cell or pixel in each digital map is fed into the statistical software and a tree classification is performed (such an operation is referred to as a tree "classification" if the dependent variable is categorical, and a tree "regression" if the dependent variable is continuous). The tree classification process also allows for the prediction of the probabilities of each observation converting to a particular change class. For example, for pixel x,y, the tree classification will produce probabilities of that pixel converting to one of the four land use change categories: conversion of forest to agriculture; conversion of agriculture to urban; conversion of forest to urban; and no change.

The tree classification algorithm works by performing a binary split on the data that maximizes the overall reduction in deviance. The measurement of deviance at any particular "node" of the tree (where a split is performed) is measured by

$$D_i = -2 \sum_k n_{ik} \log p_{ik} \tag{1}$$

at each node i of the tree, with the probability distribution p over the classes as that node and the random sample of observations n at that node i. The deviance value for the entire tree is then given by

$$D = \sum_i D_i \tag{2}$$

for all nodes (Venables and Ripley 1994).

Since the tree classification has as the dependent variable observations that reflect land use change from 1979 to 1989, and since the independent variables are values that depict general conditions for selected economic and environmental variables during the 1980s, the tree classification predicted probabilities can be used to create a predicted map of land use change for 1999.

These probabilities of change for each pixel can be fed back into the GIS, and maps of probabilities can be created. For example, a map can be created in which each pixel in the map represents the probability of that pixel converting to, say, urban use in the future. Another map can be created showing the probability of each pixel converting from forest to agriculture, etc. These maps can be combined into one map through a series of overlay operations, assuming that priority is given for conversion of either forest or agriculture to urban. This single map can be said to represent probabilities of future land use conversion.

In order to finish the prediction process, a map of landcover from 1999 was derived from 1999 Landsat 7 satellite imagery. This map is compared with the 1989 landcover images. This comparison allows calculation of the number of pixels of new urban, new agriculture, etc., that actually occur from 1989 to 1999. These numbers of new pixels – the "demand" quantities for land use change – are then allocated on the predicted probability maps in rank order, prioritizing urban change first, then agriculture, etc.

2.4 Comparison Between Actual and Predicted Land Use Change

Comparison between the predicted and actual landcover maps is performed using systematic GIS comparison methods that allow for measures of prediction accuracy.

3 SPECIFICS OF DIGITAL ANALYSIS ENVIRONMENT

3.1 Raster Environment

The analysis completed for this research takes place in a digital context – i.e. it is a computer analysis and a computer predictive land use change model. All input economic and environmental factors and all land use and land use change maps derived from satellite imagery were converted to digital raster grids – that is, a "map" that consists of a regular grid of cells, or pixels, each of which is square in shape, has an x,y position, and a z-value assigned to it (reflecting whatever factor the map represents, such as topography, income, population, etc.).

The creation of the models in the GIS computer systems makes it easy to change the spatial resolution of the cells through mathematical resampling processes. However, in the case of this research, the minimum spatial analysis unit size is a pixel approximately **60 by 60 meters**, because data derived from Landsat MSS satellite imagery contains no details smaller than a 60 by 60 meter spatial resolution (1989 Landsat TM data has a 30 by 30 meter spatial resolution, but the 1979 Landsat MSS data has only 60 by 60 meters – thus, any resampling of pixel size below 60 x 60 meters is redundant).

3.2 Specific Digital Outputs

Digital map products from the research include the following:
    a. maps of land use change in Sisaket and Chachoengsao between 1979 and 1989 using processed Landsat satellite imagery, depicting the following land use changes:
        i. conversion of forest to agriculture;
        ii. conversion of agriculture to urban;
        iii. conversion of forest to urban;
        iv. no change.
    b. creation of continuous digital maps of each Province (Sisaket and Chachoengsao) representing sub-Provincial spatial variability of four socio-economic factors –
        i. population potential, 1980s;
        ii. income distribution, 1980s;
        iii. proximity/accessibility to markets, 1980s;
        iv. proximity to infrastructure networks, 1980s.
    c. creation of continuous digital maps of each Province representing sub-Provincial spatial variability of four environmental/physical factors –
        i. proximity to timber resources, late 1970s;
        ii. proximity to water resources, late 1970s;
        iii. soil moisture index, 1980s;
        iv. topography.
    d. creation of predicted land use change maps for 1999 for both Provinces using a statistical tree classification with maps of land use change 1979-1989 as the dependent categorical factor and the economic and environmental input factors as the independent variables.
        Three separate tree classifications were run:
        i. a tree classification predicting future land use change using the four economic input factors;
        ii. a tree classification predicting future land use change using the four environmental input factors;
        iii. a tree classification predicting future land use change using all eight economic and environmental input factors.
    (A major hypothesis of the research is that using both the economic and the environmental input factors will produce more accurate predicted maps of land use change than will either the economic or the environmental factors independently.)

e. comparison of predicted maps of land use change with an actual map of land use change derived from a classified 1999 Landsat 7 satellite image, through GIS map comparison methods. Subsequent evaluation will determine which combination of inputs into the tree classification (e.g. economic only, environmental only, or both) produces the most accurately predicted map.

4   Data Sources

4.1   Vectorization of Contour Lines and Road Networks

Highly accurate contour maps obtained from the Thai government were first scanned into digital form. The contour lines and road networks were then vectorized – that is, converted to digital objects that are recognized by the GIS.

4.2   Creation of Digital Elevation Models (DEMs)

Once the contour lines were vectorized, a continuous digital surface was fitted to these contour lines through interpolation. This computationally intensive process was completed using ArcInfo software on a unix platform (using the ArcInfo TopoGrid system). The result is a surface which can be divided into a regular grid of cells of any specified spatial resolution, with each cell having a value that corresponds to its elevation in meters above sea level. Once this is completed, a simple mathematical process can be run in the GIS to calculate the slope of each pixel. These digital slope and elevation models make it possible to consider slope or elevation in any GIS modeling.

4.3   Creation of Digital Road Networks

Once vectorized, the road networks became GIS digital objects made up of a series of road line segments (called "arcs" in GIS terminology) that together form a continuous digital road network. Each road segment was assigned a distance/cost value corresponding to the approximate average travel speed along that road. For example, the average travel speed along a major highway may average around 65 mph, while on a dirt road the average may be around 30-40 mph. These distance/cost values were assigned to each road segment in the network depending on its road class, and distances along the network were calculated in terms of travel time (for example, radiating outward from a single point on the network in all directions along all roads for a distance of 2 hours travel time, taking into consideration the average speed of travel along the different road segments).

In addition to average travel times depending on road class, certain road segments were be assigned higher distance/cost values if they crossed certain barriers (such as an international boundary or are on very steep slopes) under the assumption that it will take more than the average time to cross these barriers using these roads.

These travel times along road networks were used in the creation of the continuous socio-economic models, described below, under the assumption that factors such as population and income are spatially distributed using road networks.

4.4   Soil and Rainfall Data Sources

In addition to data taken from the Thai government maps and the CDD data, soil sample data taken in Sisaket and Chachoengsao Provinces was used to interpolate soil quality maps for the Provinces using the kriging method (Myer 1991; Bonham-Carter 1994) (Laslett, McBratney et al. 1987) to generate a continuous spatial interpolation which was then used as an input to the soil moisture index model. The soil moisture model also used data collected by the Thai government from rainfall gauges as a precipitation input.

5   FUTURE DIRECTIONS FOR RESEARCH

Future possible directions for this research include the following:

A. using a method of logistic regression rather than tree classification for a predictive model – comparison with tree approach can be evaluated by the resulting accuracy of the predicted maps;
B. sensitivity testing of the parameters used to create the socioeconomic models and then considering the different predictions resulting from those models for accuracy;
C. evaluating predictions resulting from inputting different combinations of input models into the prediction regression model, to see which combinations result in the most accuracy;
D. extension of the modeling to other Thai Provinces (data has been collected for Lop Buri and Buriram Provinces) - this could allow analysis of relative influences of input models for different regions of a developing tropical country, perhaps showing that the same factors predict most accurately in each Province, or that different explanatory factors are different depending on regions with differing conditions;
E. exploration of various methods of spatial representation of the economic inputs under consideration using GIS and splus statistical software (e.g. kriging methods of spatial representation of inputs);
F. extension of the model to other tropical developing countries.

REFERENCES

Bonham-Carter, G. F. (1994). Geographic Information Systems for Geoscientists: Modeling with GIS. New York, Pergamon/Elsevier Science Inc.

Jensen, J. R. (1996). Introductory Digital Image Processing. Upper Saddle River, NJ, Prentice-Hall, Inc.
Kaimowitz, D. and A. Angelsen (1998). Economic Models of Tropical Deforestation - A Review. Jakarta, Center for International Forestry Research.

Laslett, G. M., A. B. McBratney, et al. (1987). "Comparison of Several Spatial Prediction Methods for Soil pH." Journal of Soil Science(38): 325-241.

Myer, D. E. (1991). "Interpolation and Estimation with Spatially Located Data." Chemometrics and Intelligent Laboratory Systems(11): 209-228.

O'Connor, R., M. T. Jones, et al. (1996). "Spatial Partitioning of Environmental Correlates of Avian Biodiversity in the Conterminous United States." Biodiversity Letters(3): 97-110.

Venables, W. N. and B. D. Ripley (1994). Modern Applied Statistics with S-PLUS. New York, Springer-Verlag.