

# Modeling urban growth in data sparse environments: a new approach

Michail Fragkias

Center for Environmental Science and Policy, Stanford Institute for International Studies, Stanford University, Stanford, CA 94305-6055, USA; email: fragkias@stanford.edu

Karen C. Seto

Department of Geological and Environmental Sciences and Center for Environmental Science and Policy, Stanford Institute for International Studies, Stanford University, Stanford, CA 94305-6055, USA; email: kseto@stanford.edu

Manuscript word count: 10600

**Abstract.** Although there exist numerous urban growth models, most have significant data input requirements, limiting their utility in a developing world context. Yet, it is precisely in the developing world where there is an urgent need for urban growth models and scenarios since most expected urban growth in the next two decades will occur in those countries. This paper describes a physical urban growth model that requires few, but widely available, spatially explicit data. Utilizing binary urban/non-urban maps generated by satellite remote sensing, our model can inform urban planners and policy-makers about the most probable locations and periods of future urban land-use change. Using a discrete choice framework, the model employs a spatially explicit logistic regression analysis to evaluate probabilities of urban growth for a baseline period. It calibrates parameters, validates results, predicts urban land-use change and examines future growth scenarios. Future growth scenarios can be generated through the inclusion of land prohibited from development, transportation routes, or new planned urban developments. A novel and important element of the model is the incorporation of an explicit policymaking framework that captures and reduces model uncertainty (theory and specification uncertainties), effectively addressing problems of predictive bias; this framework also allows the user or policymaker to associate predictions with a loss function. The model is applied to three cities in southern China that have experienced dramatic urban land growth in the last two decades. From 1988 to 1999, urban land in the region increased by 451.6% or at an annual rate of approximately 16.5%. Results show that the model achieves 73%-77% accuracy for different cities at 30 m and 60 m resolutions. Aggregating the predictions to the county/administrative district, shows that prediction through thresholding underperforms compared to the technique of sample enumeration.

## 1 Introduction

Urban land-use change will be one of the biggest environmental challenges of the 21<sup>st</sup> century. Whereas land-use change research during the last two decades showcased the importance of deforestation, rangeland conversion, and agricultural land loss, urban expansion will be a necessary focus for the immediate future. The 20<sup>th</sup> century witnessed some of the most dramatic urban transformations of Earth's terrestrial environments in history. At the start of the 1900s, there were only 16 cities with populations over 1 million; by 2000, there were 417 (UNCHS 2002). In 1950, there were only two cities in the world with a population of over 10 million; in 2003 there were 20 (United Nations 2004). Currently, almost half of the world's population resides in urban areas, and it is estimated that this will increase to 61 percent by 2030 (United Nations 2004).

As urban areas expand, transform, and envelop the surrounding landscape, they impact the environment at multiple spatial and temporal scales through changing regional energy budgets, loss of wildlife habitat and biodiversity, and greater demand for natural resources. Interactions between and among urban land-use, policies, and Earth system function cannot be decoupled (Steffen et al. 2004). Compact growth can lead to efficient use of resources whereas expansive development can strain infrastructure and natural resource availability. Thus, reliable prediction of urban growth under different economic and political scenarios is critical.

Current urban growth models focus primarily on cities in the industrialized world where socioeconomic data are usually readily and widely available. Yet, it is in the developing world – Asia and Africa in particular - where most of the urban growth will occur in the next two decades (United Nations, 2004). Therefore, it is important to develop models that are applicable in these data-poor contexts. However, as many researchers have pointed out, problems regarding data

availability and accuracy in less developed and developing cities make their study difficult (Nelson and Geoghegan, 2002). In many cases, socioeconomic data can be nonexistent, incomplete, inaccurate, unreliable, or all the above. For example, forecasts of urbanization utilizing such data have been severely criticized (Cohen, 2004).

Recognizing that there is no foreseeable solution to acquiring more accurate data, we have developed an urban growth model that requires data that are widely available and routinely collected, namely satellite images. Urban growth models have been developed to: (i) understand and help develop land-use and development policy, (ii) understand the past and predict the future, and (iii) create scenarios for policy-makers and planners. Our model focuses on prediction. The model integrates policy-making decisions explicitly into its structure and the performance of the model is evaluated by its ability to accurately forecast urban growth, a major concern of policy-makers and urban planners. The paper also explores issues regarding the suitability and appropriate use of discrete choice statistical framework in this type of analysis.

## **2 Theoretical aspects**

Models are tools for organizing and describing the world. Used in many different disciplines, models are used to learn about interactions among parts of a system, to generate and test hypotheses about patterns and mechanisms, and to make testable predictions. All models help to explain a hypothetical or proven relationship between a response and one or more factors that are correlated with or cause the response (independent, explanatory, or predictor variables). Uncertainty enters into all aspects of model development, testing, and evaluation, but notably, at three phrases: 1) in the development of a conceptual model, the qualitative representation of relationships between parts of a system, the strength and direction of those relationships, and

how the relationships and system may be affected by different stressors; 2) in the development of a quantitative model, or identifying variables that represent the conceptual model; and 3) in parameterization of the empirical variables.

For example, the conceptual foundations of a land-use system (e.g., urban, agricultural, coastal marine) may be developed from a number of disciplines including geography, economics, engineering, or physics. In turn, each of these conceptual models of the system can be manifested into a number of empirical specifications. Thus, a model developer is faced with the dilemma of choosing both a conceptual framework and an empirical and functional form specification for the model. Yet, it is impossible for a single model—both conceptual and empirical—to capture all aspects of reality. This begs the question, “How should a model be selected?”

## **2.1 Land-use change policy evaluation and uncertainty**

A significant amount of research has been devoted to the task of model selection. How does one select the “best” or “true” model of reality, whether it be an urban economy or land-use system under study? Should one select a best model? Although a model can be identified as the “best” according to a statistical criterion, it represents only one of many potential data generating processes. That is, a model is only one representation of reality. Researchers have started to appreciate the idea of robustness across models in policy relevant studies and in recent years, model selection has been criticized as a weak basis for policy evaluation and derivation of future prescriptions. The main disadvantage of model selection is that it ignores the fundamental dimension of “model uncertainty”. As a solution, methodologies for robustness of the policy prescription across alternative model specifications have been proposed. Our urban land-use change model explicitly incorporates a methodology that addresses issues of theory and specification uncertainty.

Methodologies addressing the issues of theory and model uncertainty are gaining popularity and are systematically being incorporated in policy-relevant research (Brock et al., 2003). First, there exists uncertainty regarding competing theories; models may not be well informed by theory or there is more uncertainty about which theory of urban growth should be utilized due to institutional and cultural factors affecting land markets in developing countries or differing assumptions regarding decision making. Second, functional form specification for statistical models is also plagued by uncertainties due to subjective perceived relevance and endogeneity as well as the question of appropriate spatial and time lags, proxy variables, etc. Essentially, incomplete knowledge on the best model of an economy raises the issue of the sensitivity of any modeling approach to alternative specifications.

We apply a statistical decision-theory framework that accounts for model uncertainty (Brock et al, 2003) to evaluate urban growth and assess urban policymaking processes. A policymaker (PM hereafter) examines a set of policies  $P$  for selection of a single policy  $p$  that can direct land-use change for administrative units (e.g. sub-city districts, cities, counties, provinces). The PM utilizes data about the urban land-uses  $d$  (a realization of a process) and the choice is conditional on a model  $m$  of the urban economy – ( $m$  can constitute alternative theories and statistical specifications). The PM minimizes the expected value of an objective (loss) function  $l(p, \theta)$  where  $\theta$  is a the exogenous state of nature (not controlled by the PM but affecting the influence of  $p$  on the loss,  $l$ ). For example, consider  $X$ , a vector of targeted growth rates per administrative unit that the PM has targeted for a particular time period. Different policies will drive different rates and patterns of urban growth. The deviation of these growth rates from the targeted growth rates can be expressed as a loss function, with each administrative unit weighted according to the policymaker's preferences. This weight is assigned according to the importance

of convergence to the target: the greater the importance of meeting the target, the higher the weight. Unknown exogenous factors that influence land-use decisions (the state of nature in a decision-theoretic framework) such as monetary or fiscal macroeconomic policies lead to the minimization of the expected value of the loss function by the PM.

In general, policy can be evaluated through the knowledge of the policy-maker's preferences and the conditional probability distribution of factors that affect the possible outcomes of the set of available policies (Brock et al., 2003). Typically, in such a framework, the probabilities of the states of nature is conditioned on existing data and selected models:  $\mu(\theta | d, m)$ . This framework changes in that  $\theta$  is not conditioned on  $m$  since the PM understands that there is probably no "best" or "true" model for the urban land use system. Thus, expected loss minimization utilizes the probability density function for  $\theta$ ,  $\mu(\theta | d)$  since it does not depend on a particular model:  $\mu$  is assumed conditional on the existing data  $d$  only (and not on  $m$ ). This is an artifact of accounting for model uncertainty; given the existence of a set of  $M$  possible models and accounting for their potentiality would require a probability distribution over the set of possible models. So, when incorporating model uncertainty,  $\theta$  is not conditioned on  $m$  since the PM takes into account all possible models (that is, in contrast to the simpler case of model selection). The optimal policy is the one that minimizes the expected loss

$$\min_{p \in P} E[l(p, \theta) | d] \text{ or } \min_{p \in P} \int_{\Theta} l(p, \theta) \mu(\theta | d) d\theta \quad (1)$$

where

$$\mu(\theta | d) = \sum_{m \in M} \mu(\theta | d, m) \mu(\theta | d) \quad (2)$$

If the loss function and the probability density function are specified, policies can be compared and the choice becomes a simple optimization problem. The expected loss does not depend on a

particular model. The methodology has its roots on what is defined as model averaging and econometric analysis attempts to derive estimates for  $\mu$ .

There are three advantages of this type of analysis. First, the policy-maker may desire and should be able to calculate a variety of characteristics (or, moments) of probability distributions of outcomes (such as the mean and variance) that may affect policy-making choices. Second, test statistics on model parameters do not inform us whether policy variables should be changed; they only reveal statistical significance of a single or a collection of estimated parameters. More focus should be placed on the magnitude (posterior distributions) of effects of policies. Third, as long as statistical methodology allows the convergence of the true value of underlying parameters and the estimates of those parameters, researchers can worry less about possible biases introduced by the nature and selection of particular statistical models.

The aforementioned simplified model of decision-making, although insightful, faces important limitations when applied to circumstances of multiple and simultaneous leverage level choices by PMs. The loss function defined suggests a single leverage: a policy tool that is examined and enters the statistical model through a single variable and associated estimable parameter. Unfortunately, models incorporating a multiplicity of policy instruments, although possible to derive, increase significantly the complexity of the model. Single and multiple policy leverage models have been already developed for macroeconomic economic growth and monetary rule models respectively.

In this paper we develop a model based on the idea that such frameworks should become increasingly more important in the study of urban growth, especially for rapidly growing cities in developing countries. Motivated by theoretical considerations, this framework is useful when models are not well informed by theory or when there is uncertainty about which theory of urban

growth should be utilized due to institutional and cultural factors affecting land markets.

Motivated by empirical considerations, this framework solves partially the problems of subjectivity and ad hoc specifications in uncertain environments.

## 2.2 Statistical modeling foundations

The profit (or utility) maximization problem of a land manager or a land user provides the conceptual framework for discrete choice model proposed for statistical analysis and the derivation of urban growth probabilities (Irwin and Geoghegan, 2001). We assume that a land manager chooses the land-use that maximizes the present value of the net expected returns. More formally, the decision of the land manager of parcel  $i$ , which exists in land-use  $u$  (for example, undeveloped land) chooses a land-use for a parcel among the alternatives  $j$  in period  $t$  that maximizes net expected returns. Thus, this one period decision is based on the following rule of conversion:

$$R_r(i, T | u) \geq R_j(i, T | u) \quad \forall j = 1, \dots, J \quad (3)$$

where  $R$  is the net expected returns from conversion of parcel  $i$  to land-use  $r$  at time  $t$  given the available land-use alternatives  $j$ . If equation (3) does not hold for at least one  $j$  other than  $u$ , then the parcel  $i$  remains unconverted in time  $t$ .

Assume that the net expected returns consist of two components:

$$R_r(i, T | u) = V_r(i, T) - \sum_{t=0}^{\infty} A(i, T + t) \delta^t \quad (4)$$

where  $V_r(i, T)$  is the one-time return from conversion to use  $r$  net of the costs of conversion in time  $T$  and  $A(i, t)$  is defined as the returns to undeveloped parcel  $i$  in time period  $t$ . The discount factor in this intertemporal maximization problem,  $\delta$ , is equal to  $1/(1+\rho)$  where  $\rho$  is the interest rate.

The fundamental assumptions of this framework are: i) that the land manager solves the intertemporal maximization problem of whether or not to develop a parcel of land by making a decision at one time based on the realization of a latent variable, the expected net returns. The land-use conversion (if it occurs) is thus the discrete outcome of a realization of a latent variable such as the net expected returns of a variety of possible conversions; ii) the likelihood that a parcel of land will be converted in any time period increases with characteristics of the parcel that increase  $V_r$  and decrease  $A$ . Such characteristics may include distances from major urban centers, transportation networks, the urban edge, and neighborhood and environmental amenities. Not all parcel characteristics are observable. Thus, a random component  $\varepsilon$  is introduced in the expression for the net expected returns. Denoting the characteristics of parcel  $i$  as  $X(i)$  and re-writing the optimal conversion rule as,

$$R_r(X(i), T | u) + \varepsilon_r(i, T | u) \geq R_j(X(i), T | u) + \varepsilon_j(i, T | u) \quad \forall j = 1, \dots, J \quad (5)$$

where the probability that a parcel  $k$ , which exists in undeveloped land-use  $u$  will be converted in use  $r$  in period  $t$  so that net expected returns are maximized is given by

$$\Pr(R_r(X(i), T | u) + \varepsilon_r(i, T | u) \geq R_j(X(i), T | u) + \varepsilon_j(i, T | u)) \quad \forall j = 1, \dots, J \quad (6)$$

The choice set for the  $j$  alternatives in the dataset used in this paper includes non-urban uses (forest, agriculture and other developable open spaces), residential, commercial and industrial uses. Assuming that the  $J$  disturbances  $\varepsilon$  are independent and identically distributed (iid), type I extreme value (EV) with a cumulative density function

$$F(\varepsilon_{ij}) = \exp(-e^{-\varepsilon_{ij}})$$

and letting  $y_i$  be a random variable for the choice made, it can be shown that the response probabilities are

$$\text{Prob}(y_i = j | x_i) = \frac{e^{\beta_j x_i}}{1 + \sum_{m=1}^J e^{\beta_m x_i}} \quad \forall j = 1, 2, \dots, J \quad (7)$$

Thus, the land manager's decision is econometrically estimated with the use of a dichotomous or polychotomous discrete choice model such as the multinomial logit model rooted in the random utility model (Greene, 2000).

### **2.3 Comparisons to other urban land-use change models**

There exist numerous land-use change models (Brown et al., 2004; Irwin and Geoghegan, 2001; EPA, 2000, Briassoulis, 1999), and more models are regularly being developed (Brown et al., 2005). This proliferation exists because of differences in landscapes, types of land-use changes, the causal factors under consideration, and the hypotheses to be tested. Common classifications of urban land-use change models include, a 3-dimensional continuum of spatial scale, time scale and human decision-making (Agarwal et al. 2002), overlapping categories of equation-based, system, statistical technique, expert, evolutionary, cellular, hybrid and agent-based (Parker et al. 2003), and categorization as large-scale, rule-based, state-change and cellular automata (Klosterman and Pettit 2005).

Our model is a hybrid spatially explicit of urban land-use change with foundations in economic and statistical discrete choice models of land-use change (Geoghegan et al., 1998), adjusted for use in data-sparse environments. The model utilizes proxy variables of economic drivers of change and focuses on the tradeoff between predicting the location versus quantity of change (Veldkamp and Lambin, 2001).

It is a hybrid model of land-use change since it combines statistical estimation, calibration and validation and the explicit consideration of policies and decision-making in a

single framework. This paper offers a policymaker-centric approach since it introduces a technique that accounts for uncertainties regarding the model's theoretical basis and implemented specification. Following the planning support system classification proposed by Klosterman and Pettit (2005) our model falls within the state-change model category while it retains some similarities to rule-based and cellular automata models. The model employs a multinomial logit statistical model core found in models such as such as the land-use change component of the second generation of the California Urban futures (CUF II) and the California Urban and Biodiversity Analysis (CURBA) models (Landis and Zhang, 1998a; 1998b, Landis, 2001). For application in data sparse environments, it utilizes pixel level data and makes the assumption of decision making at the pixel level; this is due to the lack of parcel level data that would allow the estimation of a typical economic/econometric model of land-use change. The framework that this model employs could also operate with parcel level data when those are available. The model also validates its results, a topic that few models focus on and is importantly emphasized in the GEOMOD2 model (Schneider and Pontius 2001).

The problem of sparse data environments has been examined in deforestation studies (Nelson and Geoghegan, 2002) but its implications are not clearly identified in the study of cities in the developing world. The data requirements for our model are inspired by the parsimony of the data inputs of the urban growth part of SLEUTH CA model (Clarke et al., 1997); the main difference is that our model utilizes administrative unit boundaries but not slope. Other models with similar minimum data requirements include LUCAS (Berry et al., 1996) and LTM (Pijanowski et al., 1997) –with the exception of socioeconomic variables- and GEOMOD2 (Pontius et al., 2001; Schneider and Pontius, 2001). The model is scalable to data availability:

when more input data are available in georeferenced form, the model can be modified to accommodate them.

Our model differs significantly from rule-based urban CA models in that it does not emphasize transition choices to different land-use states, but rather the probabilities of these transitions. It can operationalize the functionality of a CA model and iterate growth steps as far in the future as the user considers reasonable. The transitions to different states though are driven by statistical analysis, namely the logit models, and not by CA rules.

Finally, the model utilizes a loose model/GIS coupling scheme: a GIS transforms the raw spatial input data into the file format required by the model (implemented in Matlab). Any GIS may be utilized for the processing as long as it can export information in image or ASCII raster file format. Once the data are in the correct format, the model reads them as input. The model then computes the results and outputs tables and graphics through Matlab's Image Processing Toolbox). These tables could easily be exported to a GIS and converted into spatially explicit layers, but in our study, we have retained them in table format.

### **3 Methodology**

#### **3.1 Input Requirements and Processing**

The model first reads the input data and creates binary raster difference maps of urban/non-urban land calculated from 2 satellite images; it also processes a transportation network, the central business district location and areas excluded from development. These inputs can be provided in the model in image or ASCII raster file format. The ASCII raster file format is an intermediate file format for transferring raster data. It consists of header data

followed by rows and columns of the cell values in ASCII characters. Furthermore, the model utilizes a raster ASCII format layer of distinct political administrative units such as districts or townships.

The model processes the input and produces images in the form of matrices. For each element of the matrix (basically, for each pixel) the output represents new urban growth between years (binary), the count of urban neighbors within a 3x3 and 5x5 pixel neighborhood, the Euclidean distance to the nearest transportation network branch, the distance to the nearest urban cell and the distance to the central business district (CBD).

### **3.2 Statistical Analysis and Calibration**

Next, the model employs statistical analysis for the calibration stage. It performs a random sampling on the initial urban/non-urban image (at time  $t_0$ ). It samples 1% of all developable pixels in the  $t_0$  image (all pixels other than excluded or urban in  $t_0$ ). An index for all developable pixels is created, the vector is randomly permuted, and the 1% of the elements of the vector is selected. It then creates a dataset (our build/calibration data,  $D^C$ ) with a single dependent variable and multiple sets of independent variables arranged in such a way that can be used automatically in regression analysis.

Using the initial ( $t_0$ ) and final ( $t_1$ ) images, the model runs multiple logistic regressions with binary dependent variable  $y$  as 'change to urban or no change' between time periods  $t_0$  and  $t_1$  by creating all possible sets of explanatory variables. For  $n$  explanatory variables that can be selected for the models, a total of  $2^n$  models are considered; for example, with 5 explanatory variables the total number of models generated is 32 while for 10 it is 1024. The upper limit in the choice of number of alternative explanatory variables entering the specifications is set by the

data availability of good proxies capturing the effects of exogenous variables and the computational capacity for estimation of thousands of models (especially at the stage of full image applications of results). This estimation process for all possible model specifications given a set of possible explanatory variables targets the consequent stage model averaging.

Thus the sets include the number of neighbors in the 3x3 and 5x5 neighborhoods, the distance to the nearest urban pixel, the nearest road, the central business district, and dummy variables representing whether a choice is made within a specific distance from a road and interactions among the independent variables. We also include district dummy variables, each representing one (or a collection) of the districts of each urban area in the study. A dummy variable representing a single district is always excluded in each model since it operates as the base dummy category. One of the main calibration features of this process is the estimation of the sets of regression coefficients and the generation of predicted probabilities of change. This is information that is eventually utilized for the validation stage and the prediction of future urban growth.

Using the calibration sample ( $D^C$ ) the model performs pseudo-Bayesian model averaging. This involves generating and utilizing a weighted average of the predicted probabilities of change (the fitted values of the dependent variable) for each sample (or population) point after collecting these estimates from the  $2^n$  logit model runs. The  $2^n$  sets of fitted values are weighted by their pseudo- $R^2$  statistic (the  $R^2$  statistics are first normalized so that they add up to one). The model also captures and reports the standard deviations of the predicted probability of change estimates (but only at the stage of the full image application).

One the advantage of model averaging is that it reduces bias introduced by uncertainties surrounding the specification of the model. It also potentially addresses the problem of predictive

bias when the predictions have policy relevance. Probabilistic models are inherently prone to the problem of predictive bias or “*the systematic tendency to predict on the low side or the high side*” (Hoeting 1999). Averaging models with alternative specifications increases the chances of averaging out the problem of predictive bias.

The model calibration stage performs the following steps: once the predicted probabilities for each  $D^C$  sample point are averaged, a series of binary sample sets of predicted urban/non-urban land are created utilizing an array of probability cut-off points (threshold values) that range from 0.5 to 1. The model compares the candidate sets of predicted urban/non-urban values with the actual realization of land-use during the time period under study and selects a threshold level that best matches the observed growth rate of urban land for the  $t_0/t_1$  period. To identify the best match, the model examines all possible probability cut-off points in  $[0.5, 1]$  at a two decimal point level and generates counts of differences between predicted urban land estimates and actual urban land. The cut-off point that generates the minimum difference between predicted urban land and actual urban land is selected as the optimal threshold for the next steps. This is a form of an external imposition of an urban growth rate scenario on the model. Note that a threshold can also be selected in such a way that an alternative urban growth scenario is portrayed. The threshold value is an important additional calibration parameter set by the model (in addition to variable coefficient estimates for the  $2^n$  models) and is utilized in the “percent correctly predicted” (PCP) validation stage together with the aforementioned coefficient estimates. Figure 1 describes the modeling steps in a methodology flow chart.

*[Figure 1 approximately here]*

### 3.3 Validation

Validation is an important step in every land-use change model (Pontius et al., 2004). Our model validates the results at two spatial scales: the individual pixel and an administrative unit (through the aggregation of the pixel level information). The model generates output that can be used for PCP validation and validation through sample enumeration. It selects a distinct second random spatial sample (our test/validation data,  $D^V$ ) for each city; thus the data used for validation are drawn from a different sample than those drawn for calibration – an out-of-sample validation technique. All  $2^n$  sets of independent variables are recalculated and take their values from the new validation sample. Together with the estimated sets of variable coefficients from the calibration stage they are applied to the fitted probability logit formula for each model. We then generate predicted probabilities of change for the sampled developable pixels for time period  $t_i$  utilizing the weighted average of the fitted/predicted probabilities of all the models. The normalized pseudo- $R^2$  score that the model achieves weights this average. Each pixel is assigned a number between zero and one (the weighted average of the predicted/fitted value from the logit models).

### 3.4 PCP Validation

A tool for output validation of land-use change models is the goodness-of-fit measure of “percent correctly predicted” (PCP). Such a measure of predictive ability can be generated in a statistical discrete choice model, such as a logit or a probit (of binomial or multinomial choice). The idea behind the measure is that an alternative that is predicted with the highest probability becomes the choice of an agent. Thus, in a binary framework, a predicted probability of change that surpasses 0.5 (or 50%) in a sample is classified as a choice of change (takes the value of one)

while the predicted probabilities of less than 0.5 are interpreted as no change and take the value of zero.

For validation using the PCP measure, the probability threshold value selected in the calibration stage needs to be applied in the new sample. Intuitively, the “golden” value for a PCP measure in binomial choice is the cut-off value of 0.5. In reality, the researcher or user can arbitrarily set any probability threshold value between 0 and 1 and generate binary predicted change values based on the estimated coefficients of the sets of variables and the full set of exogenous data. As seen in the previous section, we rely on an automated selection of this threshold based on the criterion of “best growth rate matching”.

This paper presents results of the PCP measure although places emphasis on the possibility of enriching models with alternative validation methods. Typically, research that conducts validation based on the PCP measure utilizes separation either by space or by time (Pontius et al., 2004). The former includes the application of calibrated coefficients in a different geographical area while the latter utilizes calibrated parameters for a subsequent time interval. In this study we use the form of separation by space within an out-of-sample modeling framework.

### **3.5 Critique of PCP Validation**

PCP validation has the capacity of validation at the pixel level and any other level that requires the aggregation of groups of pixels (such as administrative boundaries). Unfortunately, there are significant problems associated with this methodology and the use of thresholding in probability maps. Most importantly, the application of probability thresholds in statistical models is counter

to the notion of a predicted probability in such models. Moreover, the technique automatically discards information provided by the neighboring pixels' predicted probabilities.

Indeed, the PCP measure may not provide the best way to validate a statistical model, and some statistical discrete choice modeling researchers have advised against the use of the PCP measure. Train (2003), for example, suggests that a measure of goodness-of-fit such as the PCP should be avoided because the idea behind it is in direct conflict with the purpose of discrete choice models: the generation of choice probabilities. Limited information due to the unobservable component in the formulation of the choice process forbids the prediction of the choice of the economic agent; this is the reason that partially stochastic models are employed and the modeler can only state the probability that a certain alternative will be chosen by the agent.

The nature of the predicted probabilities in discrete choice models has a standard statistical "large sample repetition" interpretation. In a binomial choice example, if the agent makes a choice 100 times, one alternative may be chosen  $x$  times and the second alternative chosen  $100-x$  times; or out of 100 people,  $x$  will choose the first alternative and  $100-x$  the second. This interpretation differs greatly from the statement that, in this repeated choice interpretation, the alternative with the highest probability is the agent's choice each time. This does not mean that the PCP criterion is not useful in land-use modeling applications; it suggests that the researcher must be aware of probability interpretations in the application of statistical discrete choice models in land-use change modeling; the nature of such probabilities should be considered in the validation stages of the models.

Typically, urban land-use change models at the pixel level fail to predict changes correctly for a significant number of locations as can be verified by the application of thresholds and the PCP measure (Pontius et al., 2004). In a case of two discrete states of land-uses (urban

versus non-urban), there are two cases where predictions are accurate and two cases where predictions are wrong. These cases are important for the stage of validation through the PCP measure, or better, the reverse of the PCP measure: the cases of wrong predictions. Table 1 reports those intuitive instances.

*[Table 1 approximately here]*

Evidence of the aforementioned predictive bias can be found when the percentages of WPNU and WPU pixels are of disproportionate size. A disproportionately larger percent of WPNU pixels relative to the percent WPU signifies a systematic tendency of smaller probability values while the reverse signifies a systematic tendency of higher predicted probability values.

A significantly limiting characteristic of policy-relevant land-use change models is that they do not allow the user to define which type of error is more important. In statistics, when testing a hypothesis about a population, the researcher can control the probability of two distinct types of errors (namely the so called type I and. type II errors). In the parallel of a court trial where the innocence of the accused is the null hypothesis, a type I error is a situation where “an innocent man is convicted” and a type II error a situation where “a guilty man walks free”. The researcher can select a level of “significance” (an essentially arbitrary percentage level based on the beliefs and risk aversion of the researcher) that decreases the probability of one type of error at the cost of increasing the probability of the other type of error. It is easily understood that the dilemma of whether to adopt a high or low significance level is stronger when the stakes are higher (as in a case of an individual facing prison time or the death penalty).

In the case of a land-use change models, the severity of the two different types of errors (WPNU and WPU) should optimally be judged by the effects/consequences of the wrong predictions on environmental, monetary, fiscal, welfare and other outcomes. When a pixel is

forecasted as urban in a future date given a policy scenario and a PM bases his decision regarding a policy implementation (such as the building of new infrastructure) on this forecast, the cost of a wrong prediction (an under- or over- estimate) can be quantified. In the case of urban growth controls, for example, the cost of a policy that failed to restrict growth in an area was otherwise identified by the model as a “hotspot” of urban development can and should be explicitly associated with the model. Alternatively, if a PM makes a decision for adopting an area-specific growth-promoting policy basing the decision again on model predictions, the cost of a WPU use could also be associated quantitatively with the model. Such cost considerations are important for urban land-use change models that aim to be policy-relevant, and cost parameters are important when the relative costs of wrong predictions by models are significantly different. In land-use change models, the researcher (and more fundamentally, the user or PM) operates without a mechanism that would allow them to control the amount of different types of errors.

PCP validation may be “vulnerable” to another type of error in predictive probabilistic models: lack of calibration. This is not to be confused with the calibration stage; in the statistics literature, it is defined as “*a systematic tendency to over- or understate predictive accuracy*” (Hoeting 1999, p.391). This potential problem is relatively unexplored in the literature of land-use modeling.

### **3.6 Validation through sample enumeration**

We propose the validation at a larger scale than the individual pixel through the help of sample enumeration. Sample enumeration is a technique of summing up predicted probabilities over a set of agents/observations for the consistent estimation of aggregate outcomes (Train 2003). We

utilize a spatial version of this popular aggregation and forecasting method for the purposes of validation. Following notation from Train (2003), a geographically explicit application is described. Consider a model that generates predicted probabilities  $P_{ni}$  that agent  $n$  will choose alternative  $i$ . Assume that a sample of  $N$  agents has been drawn from the population of decision makers (thus  $n=1, \dots, N$ ) that operate within distinct administrative boundaries or districts (so,  $d = 1, \dots, D$ ) and  $N_d$  is the number of agents in the sample within each district. A consistent estimate of the number of agents in the sample choosing alternative  $i$  in district  $d$  can be calculated as:

$$\hat{N}_{id} = \sum_{n=1}^{N_d} P_{ni}$$

A consistent estimate can be derived for the population also by applying the formula to the full population. This option is used at the prediction stage.

For the purposes of validation through sample enumeration it is desired that the sample is separated by space or time from the calibration sample. That is, we use a different sample to validate our results drawn from the same population of pixels. We employ the first formula for the validation sample of observations ( $D^V$ ) and compare the estimate of number of choices of urban change to the observed number of changes in the sample. Predictive accuracy is thus judged by the capacity of the averaged predicted probabilities to accurately capture change at the selected administrative unit.

### 3.7 Prediction

After establishing relevance through validation, the model can be used to inform policy. In this model we focus on prediction of changes not on the drivers of land-use change. In an effort to develop a model with minimal data requirements, we used spatial density and distance variables to check the predictive accuracy of a model.

The model can predict land-use using two methods. The first method thresholds predicted probabilities. In this stage of the model, prediction results are presented initially for the population of developable pixels in  $t_2$  and potentially for any discrete number of future iterations ( $t_3, t_4$ , etc.). In this paper we run a single iteration (for predictions in time period  $t_2$ ) that symbolizes the passing of a single - equivalent to the logit models - time period. The predictions utilize the estimated sets of regression coefficients and (potentially for each iteration) a newly created set of independent variables (since the urban/non-urban maps change each starting time period). As the landscape evolves, the variable values for each pixel change and are recalculated. Once a predicted probability map is produced, we apply the selected probability threshold from the calibration process (that represents a scenario for future urban growth rates), and a binary map of predicted urban/non-urban land is created.

The second method of prediction takes advantage of the potential of sample enumeration as a technique for forecasting at aggregate administrative unit levels. The only requirement is a dataset (scenario data,  $D^S$ ) that provide values for the independent variables that enter the model. In essence, the sample is adjusted so that it approximates a hypothetical future situation.

### **3.8 Scenario building and decision making**

A multiple scenario examination is an integral part of a policy decision-making process. Given the nature of the variables we have incorporated in the model, a simulation module can be used for the examination of policy-relevant alternative scenarios regarding simple policy leverages. The user/decision-maker is allowed to feed the model a collection of scenario images and define a loss function that connects predictions of urban growth with the objective the decision maker is trying to achieve (e.g. minimization of agricultural land loss). The model can provide predictions

of change and associated losses for each scenario, thus giving the decision maker the choice of the policy that minimizes the loss accounting for model uncertainty.

In the current setup, the collection of alternative scenario images falls within three distinct categories: first, the user can define areas of undeveloped land that are excluded from development; this way, the allocation of new urban pixels will be altered and a variation of hotspots of urban development will be discovered. Secondly, altered transportation routes can be designed according to existing plans of road or railway expansion. The user can incorporate this set of plans as road network images. Thirdly, the user can create patches of new developed land of high intentionality (e.g. a new airport); this type of input would be important for capturing spill-over effects of such developments that would otherwise be difficult to predict.

## **4 Application to three cities in the Pearl River delta, China**

### **4.1 The Study Areas: Shenzhen, Foshan and Guangzhou cities**

We apply the model to Shenzhen, Guangzhou and Foshan, three cities in the Pearl River Delta, southern China (Figures 2-5). The Delta generates more than 70 percent of the provincial GDP and is home to 21 million people, nearly one-third of the province's official population (Seto and Kaufman, 2003). Previous research shows that the region is undergoing rapid change (Seto et al., 2002). The region is fertile, has a strong agricultural history, and can support two to three crops per year. Although the cities are in close proximity, they differ in history, demographics, and economics (Seto and Fragkias, 2005; Seto, 2004).

Just a small fishing village until it was declared a Special Economic Zone in 1979, Shenzhen is located on the Hong Kong-China border and has experienced the most dramatic

economic growth and landscape changes of the cities in the study. Regionally, the city is the main attractor of foreign direct investment and has a large population of temporary workers; the total population of metropolitan Shenzhen is estimated at 5 million people. Accounting for unregistered workers or the floating population increases the population to 10 million. The Shenzhen metropolitan area consists of six districts. Four districts in the south (Nanshan, Futian, Luohu, and Yantian – moving from west to east) belong in the special Economic Zone (SEZ), while Baoan and Longgang districts in the north do not. The 4 SEZ districts border “the New Territories” of Hong Kong to the south and are high technology, city administration, financial, culture and information centers. Between 1988 and 1996, 45% of metropolitan urban growth occurred in Baoan, 28% in Longgang, 11% in Nanshan, 8% in Futian, 2% in Luohu and 1% in Yiantian (authors’ calculations).

The city of Foshan is located at the south-central part of Guangdong province, centrally within the PRD and just a few kilometers southwest of the city of Guangzhou (a proximity which has played an important role in the development of the city). In fact, Foshan city is positioned within 50 km of economically important cities such as Zhongshan, Jiangmen, Zhuhai and Dongguan, and within 100 km from Hong Kong. The government of Foshan is responsible for five county-level districts, namely Chancheng, Nanhai, Shunde, Sanshui and Gaoming. The official population of Foshan is estimated at approximately half a million people. The opening up of special economic zones affected the city of Foshan. In 1992, the Foshan National Hi-tech Industries Development Zone (IDZ) was approved. Urban development in Foshan was concentrated within the boundaries of Chancheng Qu (Foshan Shi) in 1988 with some scattered development extending towards several subdistricts/towns of the Foshan metropolitan area (mainly Luocun, Nanhai qu, Dongcong and Nanzhuang). By 2000, the development outside

Chancheng was extensive and followed a sprawled pattern. Between 1988 and 1996, 35% of metropolitan urban growth occurs in Chancheng, 23% in Nanzhuang, 12% in Dongcong, 9% in Luocun, 7% in Nanhai and 5% in Xiqiao (authors' calculations).

Guangzhou (Canton), capital of the Guangdong Province, is the oldest among the four cities in the study. Located at the mouth of the Pearl River, Guangzhou has been the cultural, economic, and industrial focal point of southern China. It is also a transportation hub; it has an international airport, one of the most active regional seaports, and railroad connections to all regions of the country. The traditional center of the city is formed mainly within Liwan, Yuexiu and Dongshan districts and well as within parts of Baiyun district to the north, Fangchun district to the southwest and Haizhu district to the south. Small scale scattered urban development was in place in the districts of Tianhe and Huangpu in 1988. By the year 2000, the previously peripheral districts of Baiyun, Tianhe, Huangpu, Haizhu, and Fangchun had already experienced an explosion of urban growth. The city's population is estimated at approximately 6 million people. For the 1988-1996 period, 22% of metropolitan urban growth occurs in Baiyun, 19% in Tianhe, 13% in Panyu and 12% in Huangpu (authors' calculations).

#### **4.2 Model application: calibration, validation and prediction**

We run the model for the three cities at two pixel resolutions, 30 m and 60 m. The latter is a product of the resampling of the first using a nearest neighbor algorithm. The model selects a 1% random sample from the developable pixels of the initial 1988 urban/non-urban image. For the 30 m resolution, this number is close to 14,000 pixels in the Shenzhen study area (Figure 3),

3,300 pixels in the Foshan study area (Figure 4) and 10,200 pixels in the Guangzhou study area (Figure 5).

*[Figure 3 approximately here]*

*[Figure 4 approximately here]*

*[Figure 5 approximately here]*

In its current implementation, the model explores all potential combinations of logit models created by allowing 5 variables to be included or excluded from the models – this results in 32 models. The logit models include district dummy variables (a set of an additional 32 models can potentially be generated without the inclusion of these dummies). The models for Shenzhen include dummy variables for all districts but the one containing the city’s downtown area; the models for Foshan aggregate the sub-districts into the four main districts of the metro area and create dummy variables based on this aggregation; the models for Guangzhou aggregate four small central districts into a single dummy category which becomes the base. After deriving all the sets of estimated coefficients and calculating the model-averaged predicted probabilities for  $D^C$ , the calibration stage identifies different threshold levels that best fit the observed urban growth rate for the period 1988-1996 for all cities and resolutions.

A second distinct random sample ( $D^V$ ) validates the results utilizing the model-averaged predicted probabilities for all cities and specified resolutions. In the case of PCP validation, the model applies the threshold value that generates the best match with the observed growth rate of urban land from the previous stage to  $D^V$  and creates a comparison vector between the observed and predicted values for the end year (1996) that generates and reports four classes of results (correctly predicted non-urban and urban; wrongly predicted non-urban and urban) for the cities at all resolutions (Table 2). Across all cities we observe a minor impact of resolution to the

percentages of correctly and wrongly predicted pixels; it is noteworthy that the percentages of total wrongly predicted pixels are not consistently higher or lower as one moves from the 30 m to the 60 m resolution.

*[Table 2 approximately here]*

The use of thresholds in the stages of validation and prediction compounds the problem of predictive bias in our models. Predictive bias in a PCP validation methodology artificially distorts the perceived success of the model. It is noteworthy that at high resolutions, as observed in Pontius et al. (2004), a prediction of no change (a so-called “null” model) achieves the lowest percent of wrongly predicted observations (result not reported in tables).

In the case of sample enumeration, the model sums the fitted probability values at the district level derived from the validation stage and then compares the results with the actual changes in the validation dataset.

*[Table 3 approximately here]*

*[Table 4 approximately here]*

*[Table 5 approximately here]*

We observe that the sample enumeration technique performs consistently better in the validation of the aggregate counts of change (Tables 3-5). Across all cities, we observe that the vector distances between the aggregate actual change (AAC) with aggregate sample enumeration predicted change (APC\_SE) and the aggregate thresholding predicted change (APC\_T) drop significantly as we lower the resolution from 30 m to 60 m. The performance of aggregate prediction through thresholding improves relative to aggregate prediction through sample enumeration but is still obviously far from satisfactory.

Following the validation stage, the process that generates the fitted probability values is repeated for the full population of observations for a full image probability map of development generation (Figures 6-8). Each pixel is assigned the predicted value from the logit formula using the sets of estimated coefficients.

*[Figure 6 approximately here]*

*[Figure 7 approximately here]*

*[Figure 8 approximately here]*

At the prediction stage, the model averaging process that generates the predicted probability values is repeated for the full population of observations; the  $t_1$  period (1996) land-use - and potentially transportation - data become the source from which we derive the values of independent variables and a full image of predicted probabilities of development is generated. Each pixel is assigned the averaged predicted value from the logit formula using the sets of estimated coefficients. Figures 9-11 map these predicted probabilities of development between 1996 and 2004 and the associated standard deviations of the predictions.

*[Figure 9 approximately here]*

*[Figure 10 approximately here]*

*[Figure 11 approximately here]*

We also use sample enumeration forecasting to predict amounts of urban land-use change for each district between years 1996 and 2004. For this aggregate estimate, we sum the predicted probabilities for all developable pixels to the district level and convert the number of pixels to the equivalent square kilometer area for the three cities (Table 6). For predictions based on thresholding, the model then generates images of urban/non-urban land-use for two time steps:

the end periods in 2004 and 2012. Figure 12 shows results for the latter year. The images are produced through thresholding, using the optimal thresholds from the calibration stage.

*[Table 6 approximately here]*

*[Figure 12 approximately here]*

## **5 Conclusions and future work**

The systematic incorporation of a variety of models (or specifications) accounts for model uncertainty in land-use change studies. This reduces uncertainties for a decision maker that are inherent in the use of models in the decision making process. Note that models in this framework can eventually incorporate the averaging of radically different models producing a final probability of change map.

Statistical models and PCP validation process may be incompatible; such pixel-level validation techniques may be misleading when used in statistical models. The problem with PCP validation may be due to the nature of the estimated probabilities. There exists a clear failure of spatially accurate prediction when using threshold probabilities for transition to choices.

Although the capacity of models to predict land use change at high-resolution pixel levels has been relatively disappointing (Pontius et al., 2004), the achieved levels might be adequate depending on the model use by policymakers. While the debate of prediction of location versus pattern is not resolved, we propose that sample enumeration can be successfully utilized for aggregating pixel level predictions to a larger administrative unit such as the neighborhood or the township. Thus, statistical models may be effectively limited to mainly the non-spatially-explicit sample enumeration validation and prediction. This limitation might not be severe. These

estimates could be coupled with a module that assumes the role of an allocation mechanism of predicted growth to individual pixels.

Our analysis of thresholding in the prediction stage shows that the user should primarily focus attention on the predicted probability maps as the output of discrete choice models. The modeling technique is more adequate for the identification of probable hotspots for development rather than for the generation of predicted maps of choices (change or not change in a binary setting). The model described in this paper allows the user to operate within both frameworks. Also, when utilizing a variety of models, the maps of standard deviations for the predicted probability maps convey important information regarding the spatially-explicit agreement of the various models.

In light of scarce information and the difficulty of dataset enrichment, predicted probability maps should be coupled with additional information about development trends that may be derived from stakeholders and policymakers. Generally, like in all models, the success of the model is constrained by the availability of quantifiable proxies. This model utilizes “calculatable” urban neighborhoods, distances to nearest urban pixel, nearest road and CBD and district dummies. Similarly, compared to other models of this type, scenario building requires the use of variables for which future values can be provided by the user.

For future implementation, we are considering (i) validation through separation by time; (ii) the more immediate incorporation of calibration schemes to different localities/districts that calibrate the model and capture in more detail the urban growth rates (the quantity of change) at the district level; (iii) the coupling of this model with other models operating at the county level that would help shed more light into which hotspot of urban development identified in this model

are more likely to be observed in reality; and (iv) the collection of scenarios (e.g., transportation routes).

### **Acknowledgements**

This research was supported by the US NASA New Investigator Program, Grant NAG5-10534 and the US NSF CAREER Program, Grant BCS-348986. The authors would like to thank Nathaniel Aden and Anthony DeLisi for background research and GIS development, Alexandre Boucher for helpful discussions and two anonymous reviewers for helpful comments. We also thank participants at the CUPUM 05 conference for their comments.

### **References**

- Agarwal C, Green G M, Grove J M, Evans T, Schweik C, 2002, "A review and assessment of land-use change models: Dynamics of space, time, and human choice." *Joint publication by the Center for the Study of Institutions, Population, and Environmental Change at Indiana University-Bloomington and the USDA Forest Service*. USDA Forest Service Northeastern Forest Research Station: Burlington, VT
- Berry M W, Flamm R O, Hazen B C, MacIntyre R L, 1996, "The Land-Use Change and Analysis System (LUCAS) for evaluating landscape management decisions", *IEEE Computational Science & Engineering* **3** 24-35
- Bockstael N E, 1996, "Modeling economics and ecology: the importance of a spatial perspective", *American Journal of Agricultural Economics* **78** 1168 - 1180
- Briassoulis H, 1999 *Analysis of Land Use Change: Theoretical and Modeling Approaches*, (Regional Research Institute, West Virginia University)

- Brown D G, Walker R, Manson S M, Seto K C, 2004, Modeling land use and land cover change. In *Land Change Science: Observing, Monitoring, and Understanding Trajectories of Change on the Earth's Surface*. Gutman G, Janetos A, Justice C, Moran E, Mustard J, Rindfuss R, Skole D, Turner, B L II. (eds). Dordrecht, Netherlands: Kluwer Academic Publishers, 395-409.
- Brock W A, Durlauf S N, West K D, 2003 “Policy evaluation in uncertain economic environments”, *Brookings Papers on Economic Activity* **1** 235-322
- Clarke K C, Hoppen S, Gaydos L, 1997, “A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area”, *Environment and Planning B* **24** 247-261
- Cohen B, 2004. “Urban Growth in Developing Countries: A Review of Current Trends and a Caution Regarding Existing Forecasts.” *World Development* **32** 23-51
- EPA, 2000, “Projecting Land-Use Change: A Summary of Models for Assessing the Effects of Community Growth and Change on Land-Use Patterns”, EPA/600/R-00/098, U.S. Environmental Protection Agency, Office of Research and Development, Cincinnati, OH.
- Geoghegan J, Pritchard L J, Ogneva-Himmelberger Y, Roy Chowdury R, Sanderson S, Turner B L II, 1998, “ “Socializing the pixel” and “pixelizing the social” in land-use/cover change”, in *People and pixels* Eds D Liverman, E F Moran, R R Rindfuss, P C Stern (National Research Council, Washington DC) pp 51-69
- Greene W H, 2000, *Econometric Analysis*, Fourth Edition, (Prentice Hall: New Jersey)
- Hoeting J A, Madigan D, Raftery A E, Volinsky C T, 1999, “Bayesian model averaging: a tutorial” *Statistical Analysis* **14** 382-417

- Irwin E G, Geoghegan J, 2001, "Theory, data, methods: developing spatially explicit economic models of land use change" *Agriculture Ecosystems & Environment* **85** 7-23
- Klosterman R E, Pettit C J, 2005, "Guest editorial" *Environment and Planning B* **32** 477-484
- Landis J, Zhang M, 1998a, "The second generation of the California Urban Futures Model: Part 1, Model logic and theory" *Environment and Planning B: Planning and Design* **25** 657-666
- Landis J, Zhang M, 1998b, "The second generation of the California Urban Futures Model. Part 2, Specification and calibration results of the land-use change submodel" *Environment and Planning B: Planning and Design* **25** 795-824
- Landis J, 2001, "CUF, CUF II, and CURBA: a family of spatially explicit urban growth and land-use policy simulation models", in *Planning Support Systems: Integrating Geographic Information Systems, Models and Visualization Tools*, R K Brail, R E Klosterman (eds), (ESRI Press, Redlands, CA) 157 - 200
- Nelson G, Geoghegan J, 2002, "Deforestation and land use change: sparse data environments" *Agricultural Economics* **27** 201-216
- Parker D C, Manson S M, Janssen M A, Hoffmann M, Deadman P, 2003, "Multi-agent systems for the simulation of land-use and land-cover change: a review." *Annals of the Association of American Geographers* **93** 314-37
- Pijanowski B C, Long D T, Gage S H, Cooper W E, 1997, "A Land Transformation Model: Conceptual Elements, Spatial Object Class Hierarchies, GIS Command Syntax and an Application to Michigan's Saginaw Bay Watershed" Land Use Modeling Workshop. Sioux Falls, South Dakota, June 3-5, 1997, Sponsored by NCGIA and USGS

- Pontius R G, Huffaker D, Denman K, 2004, "Useful techniques of validation for spatially explicit land-change models" *Ecological Modelling* **179** 445-461
- Pontius R G, Cornell J D, Hall C A S, 2001, "Modeling the spatial pattern of land-use change with GEOMOD2: application and validation for Costa Rica" *Agriculture, Ecosystems and Environment* **85** 191-203
- Schneider A, Seto K C, Webster D R, 2005, "Urban growth in Chengdu, Western China: application of remote sensing to assess planning and policy outcomes" *Environment and Planning B: Planning and Design* **32** 323 - 345
- Seto K C, Woodcock C E, Song C, Huang X, Lu J, Kaufmann R K, 2002, "Monitoring land-use change in the Pearl River Delta using Landsat TM" *International Journal of Remote Sensing* **23** 1985-2004
- Seto K C, Kaufmann R K, 2003, "Modeling the drivers of urban land use change in the Pearl River Delta, China: Integrating remote sensing with socioeconomic data" *Land Economics* **79** 106-121
- Seto K C, 2004, "Urban Growth in South China: Winners and Losers of China's Policy Reforms", *Petermanns Geographische Mitteilungen* **148** 50-57
- Seto, K C, Fragkias M, 2005, "Quantifying spatiotemporal patterns of urban land-use change in four cities of China with time series landscape metrics" *Landscape Ecology* **20** 871-88.
- Steffen W L, Sanderson A, Tyson P D, Jäger J, Matson P A, Moore B III, Oldfield F, Richardson K, Schellnhuber H J, Turner B L II, Wasson R J (Eds.), 2004, *Global Change and the Earth System: A Planet Under Pressure* (Springer Verlag, Heidelberg)
- Schneider L, Pontius R G, 2001, "Modeling land-use change in the Ipswich watershed, Massachusetts, USA" *Agriculture, Ecosystems and Environment* **85** 83-94

Train K, 2003 *Discrete choice methods with simulation* (Cambridge University Press, New York)

UNCHS, 2002 *The state of the world's cities report 2001* (United Nations Centre for Human Settlements Habitat).

United Nations, 2004 *World Urbanization Prospects: The 2003 Revision* (UN Press, New York)

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

**Table 1.** Prediction Outcomes

Prediction	State of the world	
	Urban	Non-urban
Non-urban prediction	Wrongly predicted non-urban (WPNU)	Correctly predicted non-urban (CPNU)
Urban prediction	Correctly predicted urban (CPU)	Wrongly predicted urban (WPU)

**Table 2.** PCP validation results: CPNU, CPU, WPNU and WPU pixels (in %) for the three cities and different resolutions\*

	Shenzhen		Foshan		Guangzhou	
	CPNU	CPU	CPNU	CPU	CPNU	CPU
30 m resolution (900 sq m per pixel)	62.42%	12.06%	71.75%	4.40%	61.73%	12.19%
	WPNU	WPU	WPNU	WPU	WPNU	WPU
	13.91%	11.61%	14.26%	9.59%	14.21%	11.87%
Total wrongly predicted	25.52%		23.85%		26.08%	
60 m resolution (3600 sq m per pixel)	58.91%	13.54%	69.94%	6.79%	62.84%	13.04%
	WPNU	WPU	WPNU	WPU	WPNU	WPU
	12.16%	15.38%	9.94%	13.33%	13.35%	10.76%
Total wrongly predicted	27.54%		23.27%		24.11%	

\* CPNU: correctly predicted non-urban; CPU: correctly predicted urban; WPNU: wrongly predicted non-urban; WPU: wrongly predicted urban.

Optimal threshold criterion: cut-off that best matches urban growth levels; applied for validation: FS-30m-0.30 | FS-60m-0.25 | SZ-30m-0.40 | SZ-60m-0.35 | GZ-30m-0.40 | GZ-60m-0.35 | GZ column reports results after dropping dummies (1-4), essentially collapsing four small central districts into one (central districts are too small to capture variation effectively).

**Table 3.** Validation through sample enumeration at different spatial resolutions for the city of Shenzhen; comparison with aggregation from thresholding

District Code (Name)	30 m			60 m		
	AAC	APC_SE	APC_T	AAC	APC_SE	APC_T
1 (Baoan)	809	826.20	753	408	417.08	502
2 (Futian)	140	129.05	213	70	67.95	117
3 (Longgang)	495	499.94	499	251	252.53	296
4 (Luohu)	34	35.52	0	19	23.2	4
5 (Nanshan)	198	192.63	221	87	83.89	80
6 (Yiantian)	30	26.57	7	7	7.1	1
7 (Other)	107	93.99	10	52	47.44	6
Vector distances between:						
AAC and APC_SE	25.54			11.71		
AAC and APC_T	149.01			124.48		

AAC: Aggregate actual change; APC\_SE: Aggregate predicted change through sample enumeration; APC\_T: Aggregate predicted change through thresholding

**Table 4.** Validation through sample enumeration at different spatial resolutions for the city of Foshan; comparison with aggregation from thresholding

District Code (Name)	30 m			60 m		
	AAC	APC_SE	APC_T	AAC	APC_SE	APC_T
11 (Beijiao)	5	11.59	0	0	4.43	0
12 (Chen cun)	6	8.52	0	2	3.19	0
13 (Dongcong)	80	50.62	0	10	15.48	0
15 (Shatou)	1	0.69	0	0	0.27	0
29 (Pingzhou)	14	16.83	0	6	4.88	0
30 (Nanhai)	49	46.02	0	10	9.94	1
31 (Chancheng)	181	206.02	440	52	51.26	126
32 (Luocun)	72	57.9	0	17	12.93	3
33 (Nanzhuang)	161	149.09	2	29	33.22	19
34 (Xiqiao)	28	26.37	14	11	8.12	11
38 (Jinsha)	8	5.71	1	0	1.39	1
40 (Xiaokang)	8	14.82	4	1	4.56	5
41 (Shishan)	0	0.11	0	0	0.08	0
42 (Dali)	2	2.68	0	0	0.67	0
43 (Yanbu)	0	0	0	0	0	0
Vector distances between:						
AAC and APC_SE	44.18			10.52		
AAC and APC_T	326.91			77.52		

AAC: Aggregate actual change; APC\_SE: Aggregate predicted change through sample enumeration; APC\_T: Aggregate predicted change through thresholding

**Table 5.** Sample runs for validation through sample enumeration at different spatial resolutions for the city of Guangzhou; comparison with aggregation from thresholding

District Code (Name)	30 m			60 m		
	AAC	APC_SE	APC_T	AAC	APC_SE	APC_T
1 (Liwan)	0	1.47	2	0	0.60	1
2 (Yuexiu)	9	10.42	17	3	1.67	3
3 (Dongshan)	5	6.81	12	6	5.2	10
4 (Tianhe)	245	246.29	271	58	50.76	67
5 (Huangpu)	166	173.1	165	28	26.20	19
6 (Haizhu)	113	123.75	112	35	31.65	47
7 (Fangchun)	66	77.54	81	19	13.95	7
8 (Baiyun)	297	300.3	248	68	79.38	86
11 (Zhencheng)	78	87.39	27	25	18.04	3
12 (Panyu)	171	174.13	111	53	34.35	20
-9999 (Other)	193	196.2	178	41	38.49	40
Vector distances between:						
AAC and APC_SE		20.67			25.04	
AAC and APC_T		99.23			48.63	

AAC: Aggregate actual change; APC\_SE: Aggregate predicted change through sample enumeration; APC\_T: Aggregate predicted change through thresholding

**Table 6.** Sample enumeration forecasting of urban growth for districts in the three cities

District Code (Name)	Summation of predicted probabilities	Predicted area of urban change in km <sup>2</sup> *
<i>Shenzhen</i>		
1 (Baoan)	45840	165.024
2 (Futian)	3594.1	12.94
3 (Longgang)	27454	98.83
4 (Luohu)	2543	9.1548
5 (Nanshan)	8806	31.7016
6 (Yiantian)	1374.5	4.9482
7 (Other)	7152.6	25.7494
<i>Foshan</i>		
11 (Beijiao)	1350.6	1.2255
12 (Chen cun)	1298.2	1.1684
13 (Dongcong)	6084.9	5.4764
15 (Shatou)	38.362	0.0345
29 (Pingzhou)	1679.1	1.5112
30 (Nanhai)	3378.9	3.0410
31 (Chancheng)	12164	10.9474
32 (Luocun)	4247	3.8223
33 (Nanzhuang)	13568	12.2114
34 (Xiqiao)	2291.8	2.0626
38 (Jinsha)	406.24	0.3656
40 (Xiaokang)	1531.7	1.3786
41 (Shishan)	6.1644	0.0055
42 (Dali)	247.58	0.2228
43 (Yanbu)	3.2617	0.0029
<i>Guangzhou</i>		
1 (Liwan)	118.64	0.4271
2 (Yuexiu)	157.09	0.5655
3 (Dongshan)	284.1	1.0228
4 (Tianhe)	10340	37.2238
5 (Huangpu)	7579.7	27.2869
6 (Haizhu)	4851.9	17.4667
7 (Fangchun)	2851	10.2637
8 (Baiyun)	16049	57.7747
11 (Zhencheng)	6657.5	23.9670
12 (Panyu)	12483	44.9391
-9999 (Other)	8249.3	29.6973

\* At 30 m resolution for Foshan and 60 m resolution for Shenzhen and Guangzhou; note that the predictions for the peripheral districts in the study cover only the portion of the districts that are included in the study area.