Environment and Planning B: Planning and Design 2005, volume 32, pages 211 – 230

DOI:10.1068/b31152

Uncertainty in extrapolations of predictive land-change models

Robert Gilmore Pontius Jr, Joseph Spencer
Department of International Development, Community and Environment, Graduate School of Geography, Clark University, 950 Main Street, Worcester, MA 01610-1477, USA;
e-mail: rpontius@clarku.edu, JFSpiper@yahoo.com
Received 10 September 2004; in revised form 10 December 2004

Abstract. This paper gives a technique to extrapolate the anticipated accuracy of a prediction of land-use and land-cover change (LUCC) to any point in the future. The method calibrates a LUCC model with information from the past in order to simulate a map of the present, so that it can compute an objective measure of validation with empirical data. Then it uses that observed measurement of predictive accuracy to anticipate how accurately the model will predict a future landscape. The technique assumes that the accuracy of the model will decay to randomness as the model predicts farther into the future and estimates how fast the decay in accuracy will occur based on prior model performance. Results are presented graphically in terms of percentage of pixels classified correctly so that nonexperts can interpret the accuracy visually. The percentage correct is budgeted by three components: agreement due to chance, agreement due to the predicted quantity of each land category, and agreement due to the predicted location of each land category. The percentage error is budgeted by two components: disagreement due to the predicted location of each land category and disagreement due to the predicted quantity of each land category. Therefore, model users can see the sources of the accuracy and error of the model. The entire analysis is computable for multiple resolutions, so users can see how the results are sensitive to changes in scale. We illustrate the method with an application of the land-use change model Geomod to Central Massachusetts, where the predictive accuracy of the model decays to 90% over fourteen years and to near complete randomness over 200 years.

1 Introduction

1.1 Three phases of LUCC modelling

The science of land-use and land-cover change (LUCC) has been maturing in terms of its levels of sophistication. Lambin et al (1999) outline three foci, each of which has major activities and products that we hope will help policymakers to make better decisions. These products will either facilitate or hinder the decisionmaking process depending on how the scientists approach the work and how they present the products. If the method and presentation are clear, then the products can be helpful. Alternatively, if the method and/or presentation induce the audience to have an inappropriate amount of confidence in the products, then the scientific activity can actually harm the process of decisionmaking concerning the management of landscapes. For example, if the decisionmakers have too much confidence in a model that claims that new policies would be effective, then they may be likely to adopt a potentially expensive new policy that will ultimately fail. Conversely, if the decisionmakers have too little confidence in a model that claims that new policies would be effective, then they may be likely to overlook a potentially inexpensive new policy that would have succeeded. In both cases, better decisions are possible when the policymaker knows how to interpret the information and how to estimate an appropriate level of confidence to have in a model, whatever that level may be. This principle holds regardless of the level of sophistication of the analysis, for example, whether the analysis is based on a complex agent-based cellular automata dynamic simulation model or on a back-of-the-envelope calculation.

The principle also holds independent of both the accuracy of the model and the phase of the LUCC research.

The first phase of LUCC research has focused on the collection and documentation of data. If a research effort focuses on only this phase of analysis, the results can be extremely useful or not, depending on the approach and presentation. For example, the State of Massachusetts funded a major effort to have each of its towns create a digital 'buildout' map that shows which land is legally developable (http://www.state.ma.us/mgis/). The state's Secretary of Environmental Affairs toured the state to display the maps at regional meetings in order to allow stakeholders to see the implied consequences of a maximum buildout scenario in which all legally developable land would be developed. This approach was tremendously helpful in allowing the citizens of each town to understand the condition of the present zoning laws. The analysis was easy to grasp with a few sentences of explanation, and therefore could lead to fruitful discussion. The analysis attempted to estimate neither the locations that were relatively likely to be developed in the near future, nor the temporal rate at which such development is likely to occur. The presentation lacked any of the sophistication of a complex dynamic LUCC model. In fact, there was no estimate of the likelihood of such a maximum buildout scenario. Nevertheless, mere presentation of data with simple analysis was extremely helpful to inform decisionmakers and stakeholders.

The second phase of LUCC research has focused on analysis of data by using statistical models and predictive models. This second phase has the potential to be much more useful than the mere presentation of the data, but also has the potential to be extremely misleading, depending on how the analysis is performed and presented. For example, in order to make the Massachusetts buildout exercise more sophisticated, a LUCC modeler could perform statistical analysis to reveal the historical empirical relationship between the tendency for a specific location to be developed as a function of biophysical and social factors, similar to Schneider and Pontius (2001) and Pontius and Malanson (2005). This approach is potentially fraught with severe methodological problems, such as multicollinearity (Neter et al, 1983), spatial autocorrelation (Griffith, 1988), the modifiable areal unit problem (Openshaw, 1984), the ecological fallacy problem (King, 1997), and the category aggregation problem (Pontius and Malizia, 2004). If statistical analysis is performed, interpreted, and presented by a skilled scientist who is aware of how to address these daunting problems, then such an analysis might be useful to inform policymakers about the LUCC phenomenon. The obvious subsequent step in terms of sophistication of LUCC research would be to use those historical empirical relationships to calibrate a model to predict future LUCC (Clarke, 2004). This type of calibration is now ubiquitous in the practice of LUCC modeling (Veldkamp and Lambin, 2001). There exist elaborate techniques to calibrate models with empirical data, then to use the models to extrapolate future land change (Jantz et al, 2003; Silva and Clarke, 2002). It is routinely claimed both by the scientists who produce the extrapolations and by the decisionmakers who examine the output, that such models are useful. If it is possible for the clear presentation of good research products to be helpful, then it would also be possible for the opaque presentation of poor research products to be harmful. Scientists should have a method to tell whether their work is helpful or harmful, because high levels of effort, expenditure, and sophistication are no guarantee that the results will be helpful. Policymakers must have the help of scientists to judge the level of confidence that they should have in model output; therefore it is essential that scientists themselves have the necessary tools to measure the level of confidence that one should have in a model. Consequently, LUCC modelers are now pursuing a third phase of research.

This third phase of research focuses on the assessment and validation of predictive LUCC models. An appropriate approach to validation is to use empirical historical information to calibrate a model, then to simulate the change in the map from that historical point in time to a contemporary point in time where a truth map is available. Then the modeler can judge how well the model predicts a known point in time. Many researchers have taken such an approach (Brown et al, 2002; Geoghegan et al, 2001; Kok et al, 2001; Pontius and Malanson, 2005). Although this might be interesting from a scientific perspective, it is not particularly interesting from a decisionmaker's perspective because the model attempts to 'predict' something that is already known, that is, the contemporary point in time for which a truth map already exists. Therefore, the information of the validation is useless, unless it is used to inform the decisionmaker about something that is not already known. One important purpose of the validation exercise is to allow the modeler and decisionmaker to understand the appropriate level of confidence to have in the model as it extrapolates to points in time that are not known, for example, the future. Methods to tackle this challenge are beginning to appear in the literature (Pontius and Batchu, 2003; Pontius et al, 2003). This paper expands on that literature by taking into consideration multiple resolutions and the expected decay in accuracy of the predicted quantity of categories of land cover.

This paper offers a general method whereby a scientist can subject a model to validation, then use the information of the validation to compute the expected accuracy of the model as it extrapolates LUCC into the future. We offer this method as a technique for LUCC modelers to capitalize on the potentially valuable information that a validation exercise reveals. The basic idea of the technique is intellectually accessible. The technique assumes that the accuracy of the model decays to randomness as the model predicts farther into the future, and then estimates how fast the decay in accuracy will occur, based on how fast it occurred in the past. Therefore the method should work to the extent that the past is an indication of the future. The necessary mathematics to perform this analysis are nontrivial, but can be computed in a standard spreadsheet. Readers are invited to write to the authors to request a spreadsheet that performs the calculations. Although this paper gives the details of the mathematics necessary to perform the analysis, the technique is designed specifically to produce graphical figures that are as easy as possible for a nonexpert to interpret. Our intention is to offer a sophisticated analysis that is ultimately accessible to anyone who wants to grasp the important messages concerning the accuracy of a prediction. Accordingly, the results are presented as the anticipated percentage of the landscape classified correctly in the prediction of the future. Components of the accuracy are budgeted according to their sources.

1.2 Central Massachusetts

We illustrate the analysis with an application of a LUCC model in Central Massachusetts over four intervals of time: 1951 – 71, 1971 – 85, 1985 – 99, and 1999 – 2013. Figure 1 (over) shows the study area, which consists of the city of Worcester and the nine surrounding towns. Today Worcester is the third most populous city in New England. This has led to a substantial increase in built area, which tends to come at the expense of forest and agricultural land (Pontius et al, 2004). Policymakers are constantly looking for ways to manage such growth, because many newly formed citizen groups are alarmed about the rate of land change (Breunig, 2003).

Stationary changes are those for which the fundamental driving forces and processes are consistent across time intervals, and therefore are useful for prediction over time. Nonstationary changes are those for which the driving forces and processes are different in various time intervals, and therefore are not useful for accurate prediction

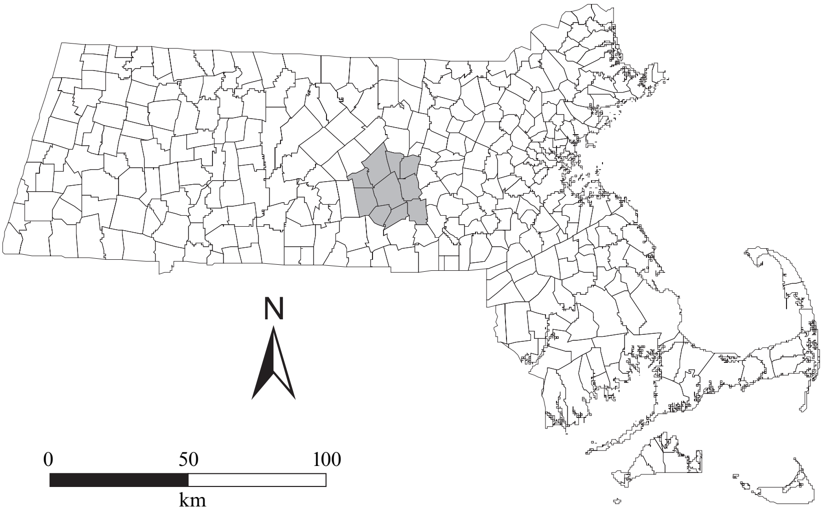

Figure 1. Ten towns in the study area of the Central Massachusetts Human Environmental Regional Observatory (CM-HERO).

across time intervals. Like any landscape, the Central Massachusetts study area has experienced stationary changes in some respects and nonstationary changes in other respects, where these respects are related to scale. Stationary trends that concern the overall quantity of land types include a monotonic increase in residential land and decrease in agricultural land. During each time interval, these changes have been driven by various forces, acting at various scales. At the beginning of the first time interval, Worcester was an industrial town, where many residents lived in the urban center and walked to work in factories. During this first period, racism and issues concerning schools drove suburbanization in much of Massachusetts. Worcester's manufacturing sector declined steadily during the second time interval in which two energy crises dramatically shocked the price structure of the entire economy (Hanson and Pratt, 1995). The increases in energy prices manifested on the landscape as people were confronted with the direct costs of heating for dwellings and commuting by automobiles. The third time interval experienced a near reversal of the energy crises, as the 1980s and 1990s witnessed a return to the days of cheap energy. In the fourth time interval, Worcester has become one of the nation's hottest housing markets, as interest rates have sustained their lowest levels since before 1951. Developers are creating a landscape of spacious single-family houses that require automobiles to access amenities and employment centers (Hanson and Giuliano, 2004). The construction of this landscape would make almost no economic sense if the high energy prices of the 1970s had persisted. The Worcester region is attractive to residents in part because it enables access by automobile to several New England cities, while it offers housing that is less expensive than in those other cities. Racism is less intensive than it used to be, but issues concerning schools continue to be fundamentally important in determining where new building occurs over short durations of time. Developers can obtain higher prices for houses built in higher quality school districts, because the location of the residence determines the public school that children attend, and there continues to be huge variation in the quality of public schools (Burgess, 1999). Alas, the quality of the schools is likely to be a nonstationary driving force because school quality and school policy change from decade to decade. Legal restrictions also guide land development in any single year; however, the zoning regulations change constantly, so knowledge of the historic restrictions does not necessarily offer predictive power over several decades (Pontius and Malanson, 2005). Whatever the variation in economic, social, and legal driving forces across time, developers have been faced consistently with the fact that it is easier to build on locations that are relatively flat and that possess desirable geologic characteristics. These characteristics are available in digital form and do not change substantially over time. This paper examines the predictive power of these stable course-scale drivers.

2 Methods

2.1 Data

The Central Massachusetts Human Environment Regional Observatory (CM-HERO) supplied the data for Worcester and the nine surrounding towns. CM-HERO has compiled a database that contains both tabular and map information from various publicly accessible sources. Tabular information gives the quantity of built land for each Massachusetts town in 1951 (MacConnell and Niedzwiedz, 1974). Map information includes surficial geology, slope, land use, and town boundaries (http://www.state.ma.us/mgis/). Each pixel of the surficial geology map gives one of three types: floodplain alluvium, till and bedrock, or sand and gravel. The slope map gives ordinal categories of slope in bins of one degree, plus an additional category for water, in order to separate flat water from flat land. The land-use maps for 1971, 1985, and 1999 show two types, built and nonbuilt, according to the definitions of Anderson et al (1976). Each map is on the same raster of 30 m × 30 m pixels.

The bold circles in figure 2 indicate the data, which show an increase in the percentage of built during the second half of the 20th century. An estimate of the percentage built for 1951 derives from tabular information, because a digital map of 1951 land cover is not yet complete (Holden et al, 2003). Maps give the percentages for 1971, 1985, and 1999. We examine the other features of figure 2 in subsequent sections.

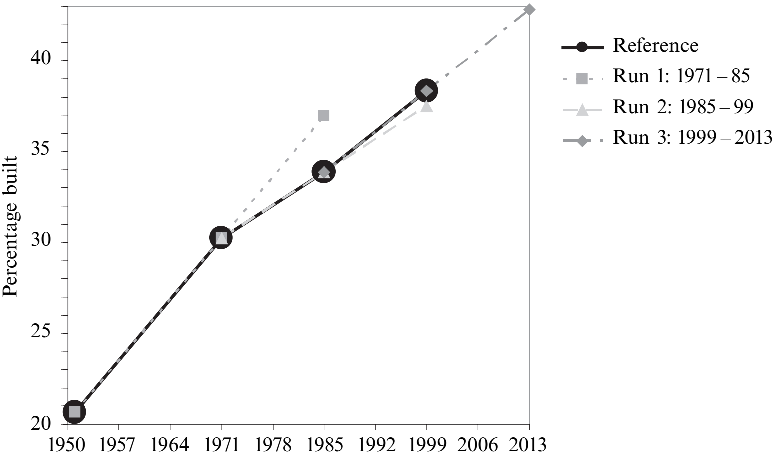

Figure 2. Reference data and extrapolation for three model runs of percentage of built land.

2.2 The LUCC model Geomod

This paper uses the LUCC model called Geomod (Pontius and Malanson, 2005; Pontius et al, 2001). Geomod is designed to predict a one-way change over time from exactly one category to exactly one other category. In the case of Central Massachusetts, the transition is from nonbuilt to built. Geomod predicts the change from an initial time to a subsequent time by separating the extrapolation into two basic tasks. First, the model predicts the net quantity of additional built land. It performs this via linear extrapolation from the calibration information. Second, the model predicts the location of the additional built land by selecting the pixels that are most likely to transition to built according to a propensity map. Geomod creates the propensity map by computing the empirical relationship between the driver maps and the land-use maps that are used for calibration.

Geomod is an appropriate model for Central Massachusetts where the process of change tends to be dominated by a one-way transition from nonbuilt to built (Pontius et al, 2004). For this paper, the sophistication and accuracy of Geomod is not particularly important because the purpose of this paper is to present a general method to estimate the uncertainty in the prediction of a model, whatever it may be. The technique in this paper for assessing the uncertainty applies to any model that predicts the change over time among any number of land categories. In the next subsection we describe the strategy to compute the expected accuracy of the prediction.

2.3 Strategy and assumptions

Figure 2 indicates the strategy of this analysis, which consists of three runs of the LUCC model. Each run is designed to separate in time the calibration information from the validation information. The first run uses calibration information from 1951 and 1971 to predict the map of 1985, at which point the first run is subjected to validation using the reference map of 1985. The second run uses the reference maps of 1971 and 1985 to predict the map of 1999, at which point the second run is subjected to validation using the reference map of 1999. The third run uses the reference maps of 1985 and 1999 to predict the landscape of 2013, for which there cannot be validation until reference information for 2013 becomes available.

Each of the three model runs must perform two tasks. Each run must predict: (a) the quantity of built area, and (b) the location of the built area. Figure 2 shows that each model run predicts the quantity of built area by fitting a line through the two calibration points, and then be extrapolating that interpolated line to make the prediction.

Figure 3 shows a more detailed research design to show the information each run uses and produces. Each model run has three types of inputs: (1) the reference land-use map for the point in time where the extrapolation begins, (2) the percentage of built area for a previous point in time, and (3) the pair of driver maps.

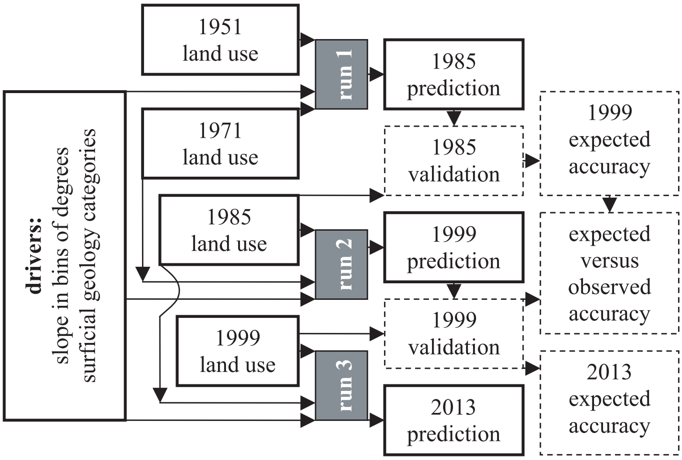

Figure 3. Steps of the methodology. The eight bold boxes represent maps. The three shaded boxes represent model runs. The five dashed boxes represent statistical analysis. The arrows show flow of information.

All three model runs use driver maps of surficial geology and slope to guide the model to specify the location of built area. Both of these driver maps show conditions that existed prior to 1971 and have been stable since 1971, therefore they are legitimate to use for calibration information for all three model runs. For each model run, Geomod computes an empirical relationship between the categories of the drivers and land use for the year that the extrapolation begins. This relationship is then used to predict the location of new built land. Geomod places new built land at the locations that have slopes and geologic characteristics similar to where built land existed in the year that the extrapolation begins.

The purpose of the first model run is to measure the accuracy of the model; therefore it must predict to a known point in time, that is, 1985. The inputs of the first run are: (1) the land-use map of 1971, (2) the tabular information of 1951, and (3) the driver maps. Geomod reads this information in order to produce a prediction of land use of 1985. A validation procedure compares the prediction of 1985 with the reference land-use map of 1985 to measure the accuracy of the prediction. This level of accuracy is then used to compute the expected accuracy of the second run.

The purpose of the second run is to see whether the observed accuracy of the second run is similar to the expected accuracy of the second run. The inputs of the second run are the two driver maps and the reference land-use maps of 1971 and 1985. Geomod reads these four maps in order to predict the land-use map of 1999. A validation step compares the prediction of 1999 to the reference land-use map of 1999 in order to compute the accuracy of the prediction. This observed accuracy is compared with the expected accuracy in order to assess the method that computes the expected accuracy.

The purpose of the third run is to extrapolate land use into the future with an expected level of accuracy. The inputs of the third run are the two driver maps and the reference land-use maps of 1985 and 1999. Geomod reads these four maps in order to predict the land-use map of 2013. The expected accuracy of the third run is computed from the observed accuracy of the second run.

There are three important assumptions that allow the model to extrapolate the expected future level of accuracy, based on available data. The first two assumptions concern stationarity. The third concerns the mathematical form of the extrapolation.

The first assumption is that some of the processes of landscape transformation over time are stationary. This assumption implies that past model performance indicates future model performance at some level. If some of the processes of land transformation are stationary, then the technique of extrapolation is appropriate at the scale of those processes. If none of the processes of land transformation is stationary, then our method of calibration, validation, and extrapolation with empirical data would indicate this by yielding: (a) low accuracy during validation, and (b) mismatch between expected model accuracy and observed model accuracy.

The second assumption is that some processes are nonstationary. The extrapolation technique assumes that the level of nonstationarity in the historic time intervals indicates the level of nonstationarity in the future. If a process of land transformation during the calibration interval is different than the process during the validation interval, then the model prediction will have low accuracy in the validation. The technique will then expect and extrapolate this low accuracy into the future. The amount of error is related directly to the level of nonstationarity, which can be a function of scale. Detailed factors (for example, school quality and legal restrictions) operate at finer scales and are less stationary than broad factors (for example, energy prices and geological characteristics); thus detailed nonstationary processes are likely to cause more errors at finer scales than at coarser scales.

The third important assumption is that the model accuracy experiences exponential decay over time and that the rate of decay is similar to the rate of observed accuracy in the validation measurement. We have selected the exponential decay function because it is sufficiently sophisticated to demonstrate the intuitive behavior of decay, but simple enough to produce results with a minimum of data. Other functional forms, such as logistic decay, are possible, but would require more data to fit uniquely. We use mathematics that require information from at most three points in time, because availability of historic maps is a major limitation in many land-change modeling projects. Three points in time can define one calibration interval and one validation interval, which are the minimum necessary to define a unique exponential decay curve for extrapolation. The validation measurement is used to extrapolate the accuracy into the unknown future, based on the assumptions that the accuracy of the prediction decays to randomness over time and that past accuracy indicates future accuracy. This paper uses information from a fourth point in time because it takes an additional step to assess the technique to extrapolate the model's expected accuracy. The next subsections give the mathematical details needed to compute the observed and the expected accuracies.

2.4 Map comparison

The mathematics of this paper are founded upon the principles of multiple-resolution categorical map comparison as described and derived by Pontius (2000; 2002) and Pontius and Suedmeyer (2004). The basic idea is that, when two maps of a common categorical variable are compared, it is possible to budget the components of agreement and disagreement in terms of the quantity of each category and the location of each category at multiple resolutions. The interpretation of agreement is the proportion of pixels classified correctly, and the interpretation of disagreement is the proportion of pixels classified incorrectly. Therefore the sum of agreement and disagreement always equals 100% of the study area. There are three components of agreement: agreement due to chance, agreement due to quantity, and agreement due to location. There are two components of disagreement: disagreement due to location and disagreement due to quantity. The next subsection gives the necessary mathematics to extrapolate the expected components of the agreement and disagreement between a reference map and a prediction map. At the beginning of the extrapolation, the components reflect the agreement between the reference map and itself; thus agreement is 100%. The agreement then decays to randomness as the extrapolation progresses in time. This paper uses the simplest possible mathematics to extrapolate each component at a rate consistent with observed measures of validation.

2.5 Mathematics

This paper uses the notation below. In this particular study, $G = 1116$, $N_1 = 651\,591$, $J = 2$, $T_1 = 1971$, $T_2 = 1985$, $T_3 = 1999$, and $\Delta_2 = \Delta_3 = 14$.

g is the grain size of resolution of the pixels as a multiple of 30 m, $g = 1, \ldots, G$;
G is the grain size of resolution at which the entire study area is in one coarse pixel;
n is the index for a specific pixel, $n = 1, \ldots, N_g$;
N_g is the number of pixels in the study area at resolution g;
j is an index for a specific category, $j = 1, \ldots, J$;
J is the number of categories in the study;
t is the time in years;
m is an index for prediction run, $m = 1, 2, 3$;
T_m is the year when prediction run m begins its extrapolation from a reference map;
W_{gn} is the weight at resolution g of pixel n;
R_{gnjt} is the membership at resolution g of pixel n to category j for time t in the reference map;

S_{gnjtm} is the membership at resolution g of pixel n to category j for time t in the prediction map from run m;
C_{gm} is the agreement between the reference map and a prediction map due to chance in both quantity and location at resolution g from run m;
D_{gtm} is the agreement between the reference map and a prediction map that distributes the quantities of the predicted categories uniformly in space at resolution g for time t from run m;
E_{gtm} is the agreement between the reference map and the prediction map at resolution g for time t from run m;
F_{tm} is the agreement between the reference map and a prediction map that has zero disagreement due to location, given the disagreement due to quantity in the simulation for time t from run m;
κ_{gtm} is the kappa for location statistic that indicates agreement in terms of location between the reference map and the prediction map at resolution g for time t from run m;
A is the horizontal asymptote as time approaches infinity for agreement between the reference map and a prediction map that has zero disagreement due to location;
Δm is the time interval in number of years from beginning of extrapolation for run m to year of the second interpolation point for run m, which is the same as the interval between the year of the beginning of extrapolation of the prediction for run $m - 1$and the validation year for run $m - 1$, for example $\Delta_2 = T_2 - T_1$, and $\Delta_3 = T_3 - T_2$;
Y_{gm} is the interpolation value for expected agreement between the reference map and a prediction map that distributes the categories uniformly in space at resolution g for run m at time T_m;
V_m is the interpolation value for expected agreement between the reference map and a prediction map that has zero disagreement due to location for run m at time $T_m + \Delta_m$;
H_{gm} is the interpolation value for expected kappa of location statistic at resolution g for run m at time $T_m + \Delta_m$;
B_{gm} is the decay coefficient that controls how fast F'_{tm} approaches A as time progresses at resolution g for run m.

Equation (1) gives the proportion of each category j for time t in the reference map, and equation (2) gives the proportion of each category j for time t in the prediction map for run m. The quantity of each category is independent of resolution, hence is denoted by a • symbol in the place of g and n. Equations (3)–(7) describe mathematically the terms defined above.

$$R_{\bullet jt} = \frac{\sum_{n=1}^{N_g} W_{gn} R_{gnjt}}{\sum_{n=1}^{N_g} W_{gn}} , \tag{1}$$

$$S_{\bullet jtm} = \frac{\sum_{n=1}^{N_g} W_{gn} S_{gnjtm}}{\sum_{n=1}^{N_g} W_{gn}} , \tag{2}$$

and

$$C_{gm} = \frac{\sum_{n=1}^{N_g} \left[W_{gn} \sum_{j=1}^{J} \min(R_{gnjt},\, 1/J) \right]}{\sum_{n=1}^{N_g} W_{gn}} , \tag{3}$$

$$D_{gtm} = \frac{\sum_{n=1}^{N_g} \left[W_{gn} \sum_{j=1}^{J} \min(R_{gnjt},\, S_{\bullet jtm}) \right]}{\sum_{n=1}^{N_g} W_{gn}} , \tag{4}$$

$$E_{gtm} = \frac{\sum_{n=1}^{N_g} \left[W_{gn} \sum_{j=1}^{J} \min(R_{gnjt},\, S_{gnjtm}) \right]}{\sum_{n=1}^{N_g} W_{gn}} , \tag{5}$$

$$F_{tm} = \sum_{j=1}^{J} \min(R_{\bullet jt},\, S_{\bullet jtm}) \ , \tag{6}$$

$$\kappa_{gtm} = \frac{E_{gtm} - D_{gtm}}{F_{tm} - D_{gtm}} \ . \tag{7}$$

For run 1, beginning year is 1971 and validation year is 1985. For run 2, beginning year is 1985 and validation year is 1999. For run 3, beginning year is 1999 and extrapolation is to 2013. Therefore, we can compute the observed values of equations (1) – (7) for run 1 at 1985 and for run 2 at 1999. We can compute the expected values of equations (3) – (7) for run 2 at any year greater than 1985 and for run 3 at any year greater than 1999. A prime symbol applied to any variable denotes the agreement that is expected during extrapolation, which is relevant for C'_{gm}, D'_{gtm}, E'_{gtm}, F'_{tm}, and κ'_{gtm}. The lack of a prime symbol denotes the agreement that is observed during validation.

2.6 Distant future

We presume that the accuracy of the prediction decays to randomness as the duration of the extrapolation increases. This means that the ability of the model to predict accurately decays towards the level of accuracy that we would expect for a random prediction; this does not mean that the landscape becomes random over time. Our procedure estimates how quickly the accuracy of the prediction decays to the level of accuracy that would be expected for a random prediction. Therefore, it is helpful to define the 'distant future' as a point in time at which the landscape is predicted with a level of accuracy equal to the level of accuracy expected through random chance. Let us denote that distance future point in time as ∞. We can describe statistically some of the characteristics of that distant future landscape based on its definition. Specifically, the information of today does not enable the modeler to make better than a random guess at predicting the distant future in terms of either the quantity of any category *j* or the location of any category *j*. Therefore, it is useful to portray $R_{\bullet j\infty}$ as a random variable that describes the proportion of the landscape for each category *j* at that distant point in time, because then the expected agreement between the distant future landscape and any prediction will be equal to the expected agreement due to chance. The expected value of $R_{\bullet j\infty}$ is $1/J$, but any particular realization of $R_{\bullet j\infty}$ will not be exactly $1/J$ because $R_{\bullet j\infty}$ is a random variable. Equation (8) models $R_{\bullet j\infty}$ as

$$R_{\bullet j\infty} = \frac{U_j}{\sum_{j=1}^{J} U_j} , \tag{8}$$

where each U_j is a random selection from the uniform distribution, that is,

$$U_j \sim \text{uniform } (0, 1) .$$

The modeler's most appropriate strategy to predict the map of this distant future is to predict $1/J$ as the proportion in each category and to distribute each category uniformly in space, so that the membership of every pixel to each category is $1/J$ for all J categories. Equation (9) gives the agreement between the reference map and a prediction map that has zero disagreement due to location for $t = \infty$.

$$A = \sum_{j=1}^{J} \min(R_{\bullet j\infty}, 1/J) . \tag{9}$$

$R_{\bullet j\infty}$ is a random variable; therefore A is also a random variable. Table 1 gives the expected value for A, which is as a function of only J and is independent of resolution g. Table 1 is based on a Monte Carlo technique that estimates the parameter A to within ±0.0001 at the 99.999% confidence level. Table 1 gives the values to the nearest percent for $J = 1, \ldots, 1000$. As J becomes larger than 1000, we think the expected value of A approaches 75%, but a formal proof eludes us.

Table 1. Asymptotic percentage correct with zero location error as the duration of the extrapolation approaches infinity.

Number of categories	Percentage correct
2	81
3	78
4–6	77
7–17	76
18–1000	75

For the extrapolation into the future, equation (10) shows that A is the asymptotic limit for the agreement between the reference map and a prediction map that has zero disagreement due to location, given the disagreement due to quantity at time t. Equation (10) shows that when $t = T_m$ the agreement is 100%, which is the agreement at the very beginning of the extrapolation.

$$F'_{tm} = A + \{[1 - A] \exp[B_m (t - T_m)]\} , \tag{10}$$

$$B_m = \left[\ln \frac{V_m - A}{1 - A} \right] \Big/ \Delta_m . \tag{11}$$

Equations (11) defines B_m as a negative number such that F'_{tm} equals V_m at time $t = T_m + \Delta_m$. V_m is the interpolation value for time $T_m + \Delta_m$ for run m, where $m > 1$. V_m is set equal to the agreement observed in run $m - 1$ between the reference map and a prediction map that has zero disagreement due to location, given the disagreement due to quantity for run $m - 1$ at the validation time of run $m - 1$. V_m is in the interval $[0, 1]$, and equation (11) requires that $A < V_m$, therefore $B_m < 0$. Consequently, the factor in curvy braces of equation (10) approaches zero as time progresses from T_m

to infinity, thus F'_{tm} approaches A as time progresses. If $V_m \leqslant A$, F'_{tm} should be set to A for all t.

C'_{gm} is the expected agreement between the reference map for time T_m and a simulated map due to chance in terms of quantity and location. C'_{gm} does not change with time but grows larger as the resolution becomes coarser. D'_{gtm} decays to C'_{gm} according to equation (12). Y_{gm} is the interpolation value for time T_m at resolution g for run m, where $m > 1$. Y_{gm} is set equal to the agreement observed between the reference map and a simulation map that distributes the categories uniformly in space at resolution g for run m at time T_m.

$$D'_{gtm} = C'_{gm} + \{(Y_{gm} - C'_{gm})\exp[B_m(t - T_m)]\} \ . \tag{12}$$

Kappa for location of run m decays from one to zero as time progresses from T_m to infinity, according to equation (13). H_{gm} is the interpolation value for time $T_m + \Delta_m$ at resolution g of run m. H_{gm} is set equal to the observed kappa for location in the comparison between the reference map and the prediction map for the validation at resolution g of run $m - 1$. Equation (13) requires that $H_{gm} > 0$. If $H_{gm} \leqslant 0$, then κ'_{gtm} should be set for 0 for all time t. In most cases, $0 \leqslant H_{gm} \leqslant 1$, for which equation (13) shows exponential decay.

$$\kappa'_{gtm} = \exp\left(\frac{t - T_m}{\Delta_m} \ln H_{gm}\right) \ , \tag{13}$$

$$E'_{gtm} = D'_{gtm} + [(F'_{tm} - D'_{gtm})\kappa'_{gtm}] \ . \tag{14}$$

Consequently, as time progresses from T_m to infinity for run m, E'_{gtm} decays from F'_{tm} to D'_{gtm} according to equation (14). The expected percentage correct E'_{gtm} decays from 1 to C'_{tm}, because F'_{tm} decays from 1 to A, and D'_{gtm} decays from Y_{gm} to C'_{tm}. In other words, the expected percentage correct decays from perfect to random as the time interval grows from zero to infinity.

2.7 Coarser resolutions

All simulations are run at the finest resolution, $g = 1$, at which each pixel is 30 m × 30 m. We compare the predicted map and the reference map at multiple resolutions in the manner of Pontius (2002). Therefore, we can do the entire analysis at any resolution. At every resolution, the sum of the memberships over all categories equals unity for every pixel, such that equation (15) holds.

$$\sum_{j=1}^{J} R_{gnjt} = \sum_{j=1}^{J} S_{gnjtm} = 1 \ . \tag{15}$$

3 Results

Figures 4 and 5 show a visual comparison of the performance of the first run, which gives a prediction for 1985 that is 90% correct. Figure 4 is a reference map that shows the true change between 1971 and 1985. Figure 5 shows the predicted change over the same period. Figure 6 (over) shows the sources of agreement and disagreement between figures 4 and 5. Of the 10 percentage points of error, 4 percentage points are disagreement due to quantity, attributable to the model predicting too much additional built area as evidenced in figure 2. The remaining 6 percentage points of error are disagreement due to location, attributable to the model predicting change in the wrong place as evidence in the comparison of figures 4 and 5. The results from run 1 give information necessary to compute the expected accuracy of run 2.

Figure 7 (over) shows the expected accuracy of run 2 for any point in time between the beginning of the extrapolation and 2013. The expected agreement for run 2 at 1985 is 100%, because the time duration of the extrapolation is zero at 1985. The expected agreement shrinks as the time duration of the extrapolation projects farther into the future. At 1999 the expected agreement for run 2 is 90%, which is budgeted as 4% disagreement due to quantity and 6% disagreement due to location.

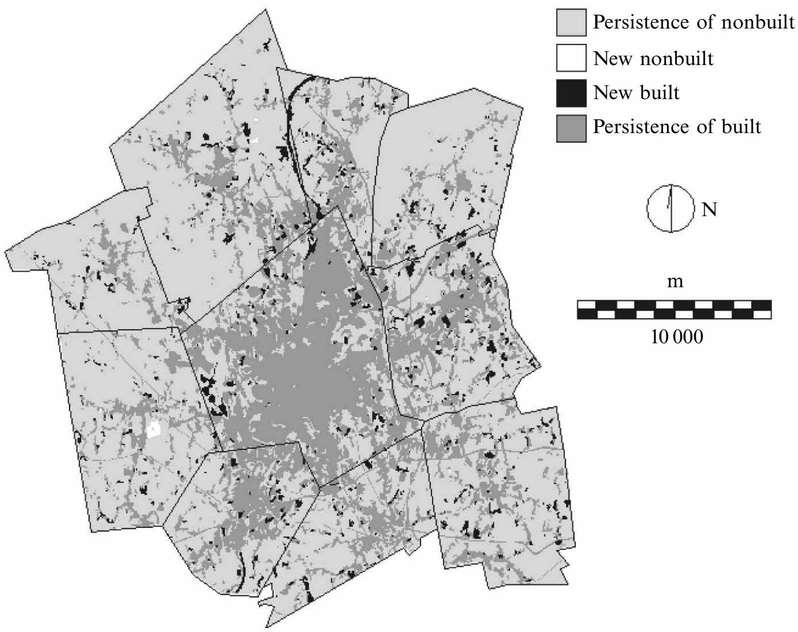

Figure 4. Reference map of change in built and nonbuilt from 1971 to 1985.

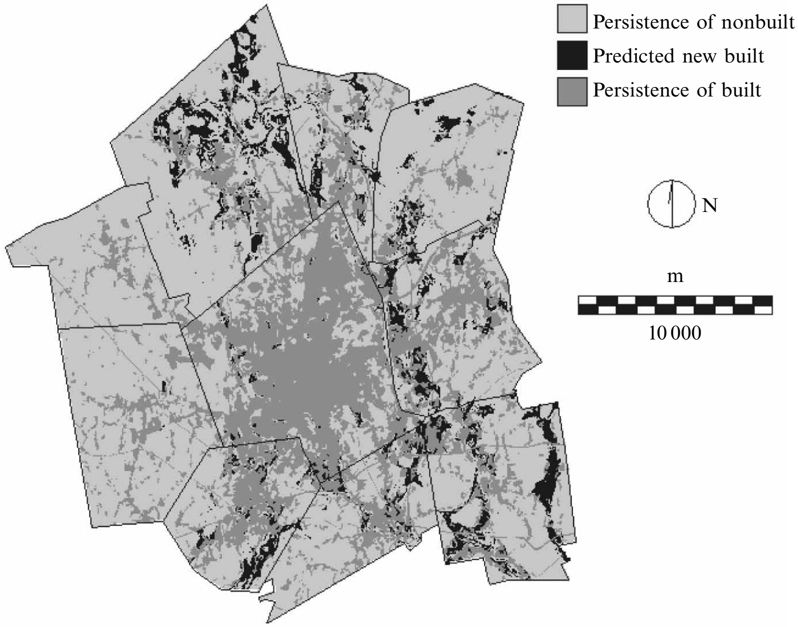

Figure 5. Predicted map of additional built between 1971 and 1985.

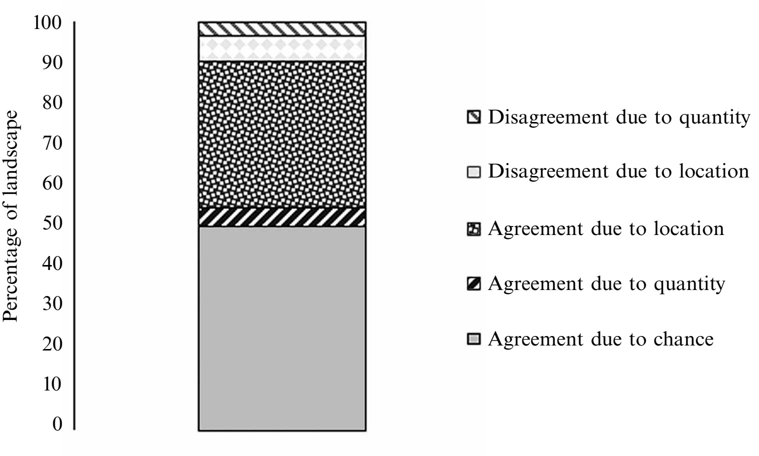

Figure 6. Budget of sources of agreement and disagreement for the comparison of the prediction map for 1985 versus the reference map for 1985.

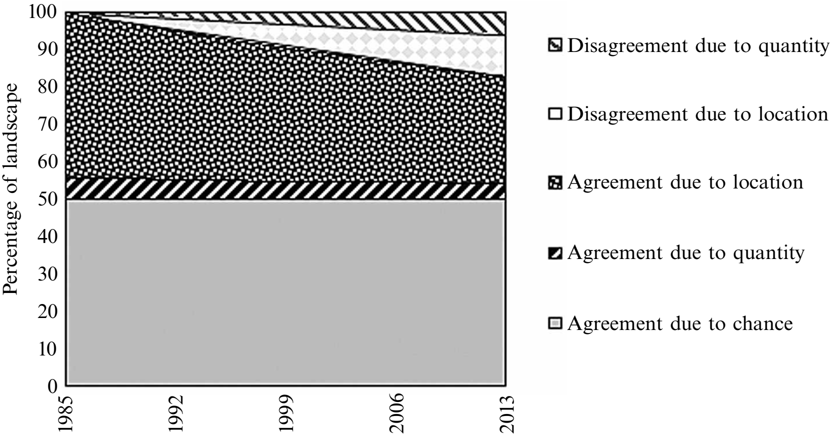

Figure 7. Expected budget of sources of agreement and disagreement for prediction from 1985 to 2013.

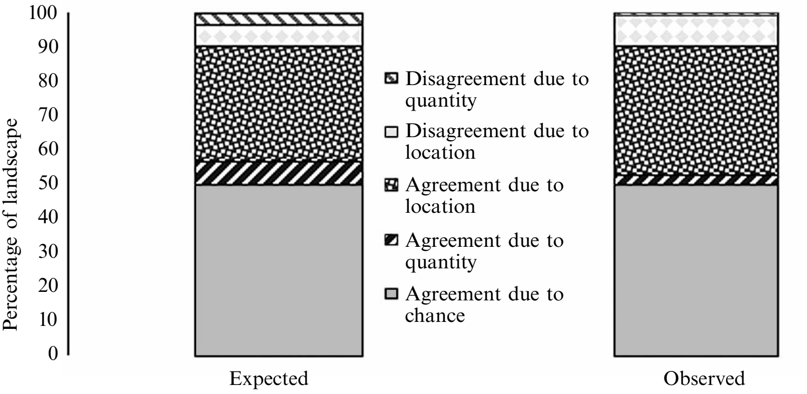

Figure 8. Expected accuracy versus observed accuracy of run 2 in terms of the budget of sources of agreement and disagreement for the comparison of the prediction map for 1999 versus the reference map for 1999.

The observed agreement between the 1999 reference map and 1999 predicted map from run 2 is 90%, which is budgeted at 1% disagreement due to quantity and 9% disagreement due to location. Figure 8 compares the components of the expected accuracy versus the observed accuracy of run 2 for 1999. The expected total accuracy is within 1 percentage point of the observed total accuracy for 1999, but the two components of the disagreement do not match perfectly. Almost half of the expected disagreement is attributable to quantity; whereas almost none of the observed disagreement is attributable to quantity.

Figure 9 shows the prediction for 2013 from run 3. Figure 10 shows that the expected accuracy of this prediction at the 30 m resolution is 90%, where the disagreement is budgeted as 1% due to quantity and 9% due to location. Figure 11 (over) shows

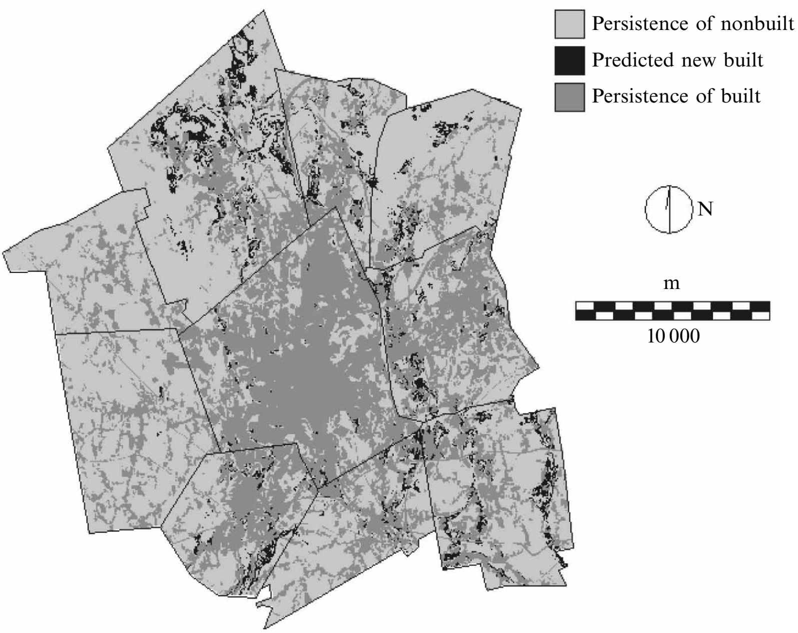

Figure 9. Predicted map of additional built between 1999 and 2013.

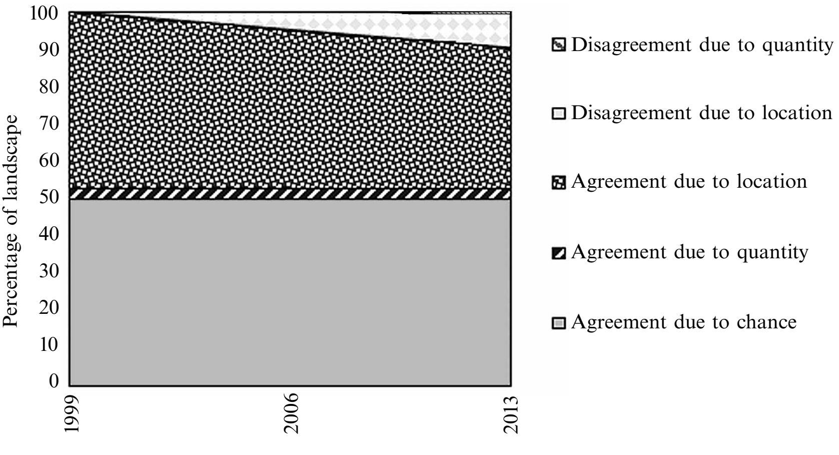

Figure 10. Expected budget of sources of agreement and disagreement for prediction from 1999 to 2013 at the 30 m resolution of the raw data.

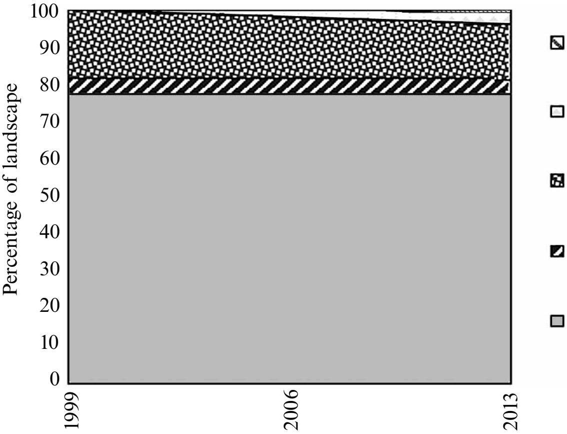

Figure 11. Expected budget of sources of agreement and disagreement for prediction from 1999 to 2013 at 4 km resolution.

that the expected accuracy of this prediction at the 4 km resolution is 96%, where the disagreement is budgeted as 1% due to quantity and 3% due to location.

The expected accuracy of the prediction of 2013 is sensitive to a variety of factors, especially the selection of the time intervals for calibration, validation, and extrapolation. If the calibration interval were 1951 – 71, the validation interval were 1971 – 85, and the extrapolation interval were 1985 – 2013, then figure 7 shows that we would expect the prediction for 2013 to be 82% correct, with 7% error due to quantity and 11% error due to location. This allocation of expected components of error is different than the allocation of expected components in figure 10.

4 Discussion

4.1 Interpretation with respect to stationarity

The match between the expected and observed overall accuracy is an indication that the process of land transformation is somewhat stationary at a coarse scale, as figure 2 shows that growth in built area increases monotonically. The calibrated model parameters for the first run are nearly identical to the independently calibrated parameters for the second run, which indicates that the relationship between the dependent variable and the independent variables are stationary over these particular time intervals. In both runs, pixels that have the largest propensity for built land are on relatively flat slopes and on sand and gravel. If the processes were even more stationary, then we would see even more of a match between the expected error budget and the observed error budget.

The lack of perfect match between the expected components of error and observed components of error is an indication that the process of land transformation is somewhat nonstationary at a fine scale. Figure 2 shows that the rate of increase in the quantity of built land was smaller in the interval 1971 – 85 than in any other interval. There may be a variety of reasons, including the previously mentioned energy crises of the 1970s. Consequently, run 1 predicts too much newly built land, and run 2 predicts too little newly built land (figure 2). If the processes were less stationary, then we would see less of a match between the expected error budget and the observed error budget.

4.2 Is the model good or poor?

Some readers might consider the observed accuracy of 90% to be ‘good’. We recommend that scientists avoid using such simple adjectives because they can be extremely misleading. Different people tend to interpret such words quite differently. 'Good' usually implies that the model is doing what it is supposed to do, that is, predicting landscape change accurately. This is not necessarily the case.

To evaluate the performance of the model, we should compare the model prediction to what we could have predicted without the LUCC model. A null model would predict pure persistence, that is, no change, over the duration of the simulation. Specifically, a null model of the change between 1985 and 1999 would simply use the 1985 map as the prediction for 1999. In this case, the accuracy of the null model would be 94%, because there is only 6% change on the landscape between 1985 and 1999, when measured at the resolution of the data, that is, the 30 m resolution. This 6% change is the sum of a 5% gain in built and a 1% loss in built. It is typical that a null model predicts better than a LUCC model at the fine resolution of the raw data. Pontius et al (2004) explore this issue in depth.

When confronted with the comparison to a null model, many modelers are tempted to ignore the pixels that persist, in order to focus on the pixels that change, because the model tries to predict change. Pontius et al (2004) advise against elimination of any pixels that are legitimate candidates for change because potentially serious problems can occur when the pixels that persist are ignored. The purpose of the model is to discriminate between the pixels that change from those that persist, so the measure of validation should consider all pixels that are candidates for change, whether or not they change in truth and/or in the model. This situation underscores the importance of proper interpretation of the measure of validation. The percentage correct used in this paper indicates the ability of the model to predict the entire landscape, including change and persistence. Proper interpretation for model assessment requires that one must be cognizant of the amount of true change and the amount of predicted change, as can be seen in figures 2, 4, and 5. Specifically, if the purpose of the model is to predict change that occurs on a small percentage of the landscape, then the overall accuracy for the entire landscape can be as high as 90%, even when the model predicts the location of change inaccurately, which is the case in the application of Geomod in this paper.

However, this does not necessarily mean that the model prediction is 'poor'. In this case, the word 'poor' is as useless and misleading as the word 'good'. The null model performs with a higher level of accuracy than Geomod at the 30 m resolution of the raw data. This is because Geomod cannot predict land-cover change accurately to within 30 m, but 30 m spatial precision might be irrelevant to the policymaker. The policymaker might want to know whether Geomod predicts the correct location within the general neighborhood, for example, within a few kilometers of the true location. If this were the case, then Geomod should be compared with a null model at some spatial resolution coarser than 30 m. In fact, there are many resolutions at which Geomod is more accurate than the null model, therefore it is important to perform the validation at multiple resolutions.

4.3 Importance of multiple resolutions

Figure 2 shows that Geomod's prediction is better than the null model at the very coarsest spatial resolution. At this coarsest resolution, the entire study area is in one very coarse pixel; therefore the entire prediction concerns only the quantity of built land, as shown in figure 2. The quantity of built land is 34% in 1985, 38% in 1999, and 37% as predicted by the linear extrapolation to 1999. A null model of the change between 1985 and 1999 would 'predict' that the percentage built in 1999 is the same as the percentage built in 1985, thus the null model would have an error due to quantity of 4 percentage points. The linear extrapolation predicts the percentage built in 1999 with an error of 1 percentage point, which is more accurate than the null model.

The null model predicts the map more accurately than Geomod at the fine resolution, and Geomod predicts more accurately than the null model at a coarse resolution. Therefore, there is a resolution at which a null model predicts just as accurately as Geomod. That resolution is called the null resolution (Pontius et al, 2004). For the Geomod example, the null resolution is 4 km, which means that Geomod predicts better than a null model at resolutions coarser than 4 km. Equivalently, the null model is more accurate than Geomod at a spatial precision finer than 4 km. This type of analysis illustrates the importance of examining the model performance at multiple resolutions (Costanza, 1989).

Figure 11 gives the anticipated model performance at a resolution of 4 km. At that resolution, most of the anticipated error in 2013 is attributable to the inability of the model to predict location correctly. As resolution becomes coarser, disagreement due to location shrinks while disagreement due to quantity is unaffected, because change in spatial resolution influences the measurement of location but is independent of the measurement of quantity.

4.4 Extrapolation over long intervals

The technique extrapolates the certainty of the model prediction to any time in the future. The farther in time, the closer the anticipated level of accuracy will be to randomness. The importance of the technique is to offer a visual method to show how fast the accuracy decays to randomness. For our example, if the certainty is extrapolated to 200 years in the future, then the anticipated accuracy of the model decays to 54%, which is within 4 percentage points of complete randomness at the 30 m resolution. Most of the agreement during such an extrapolation can be attributable to persistence on the landscape. Therefore, these results suggest that transitions over the next couple of centuries are not predictable with a level of accuracy that is much better than random. These results seem to be consistent with our knowledge of how humans transform landscapes in New England (Cronon, 1983; Foster and Aber, 2004). Two hundred years ago, the primary landscape transition was from forest to agriculture, while the human population was increasing. A simple extrapolation of that trend would have had essentially no predictive power, because the basic mechanisms of land transformation have changed in the last two centuries. Even the broadest macroprocesses have been nonstationary over long time intervals in New England. The human population has simultaneously grown and moved from the rural locations to urban centers mainly because of changes in the way humans use energy. Consequently, forests have regrown on abandoned agricultural land. If we experience a substantial change in the driving forces of land transformation in the future, then the model prediction will probably be no better than random at that future point in time. The methods of this paper offer a technique to quantify how fast we will approach that point, based on empirical data.

5 Conclusions

This paper presents a general method to compute the expected accuracy of a prediction from a GIS-based model that simulates land-use and land-cover change. The premise is that the expected accuracy of the extrapolation decays to randomness as the duration of the extrapolation grows. The techniques in this paper estimate the rate at which the decay occurs, on the basis of the observed accuracy of the model at a point in time where reference data are available for validation. The expected percentage correct is budgeted according to: agreement due to chance, agreement due to quantity, and agreement due to location. The expected percentage error is budgeted according to: disagreement due to location and disagreement due to quantity. The method predicts the total error to within 1 percentage point, and the two components of disagreement to within 3 percentage points when applied to a landscape in Central Massachusetts over fourteen years at a 30 m resolution. The concepts and equations can be applied at multiple resolutions to any LUCC model that predicts dynamic transitions among any number of categories. The purpose is to determine the level of trust that we should have in predictive LUCC models. We hope other modelers will use this technique to convey the certainty that one should have in extrapolations of land change into the future. The visual display of results will help scientists to communicate the appropriate level of confidence to have in model predictions, which is what is needed in order for modeling exercises to be truly useful.

Acknowledgements. Data for this analysis were supplied by the Central Massachusetts HERO project, which is supported by the National Science Foundation (NSF) via the grant "Infrastructure to Develop a Human – Environment Regional Observatory (HERO) Network" (Award ID 9978052). In addition, this research was supported by the Center for Integrated Study of the Human Dimensions of Global Change, through a cooperative agreement between the NSF (SBR-9521914) and Carnegie Mellon University. Clarklabs has made the building blocks of this analysis available in the GIS software, Idrisi®.

References

Anderson J R, Hardy E E, Roach J T, Witmer R E, 1976, "A land use and land cover classification system for use with remote sensor data", professional paper 964, US Geological Survey, Reston, VA

Breunig K, 2003 *Losing Ground: At What Cost?* (Massachusetts Audubon Society, Lincoln, MA)

Brown D G, Goovaerts P, Burnicki A, Li M, 2002, "Stochastic simulation of land-cover change using geostatistics and generalized additive models" *Photogrammetric Engineering and Remote Sensing* **68** 1052 – 1061

Burgess P, 1999, "Conservation and development in conflict", Master of Arts thesis, International Development Program, Clark University, Worcester, MA

Clarke K, 2004, "The limits of simplicity: toward geocomputational honesty in urban modeling", in *GeoDynamics* Eds P Atkinson, G Foody, S Darby, F Wu (CRC Press, Boca Raton, FL) pp 215 – 232

Costanza R, 1989, "Model goodness of fit: a multiple resolution procedure" *Ecological Modelling* **47** 199 – 215

Cronon W, 1983 *Changes in the Land* (Harper Collins, Toronto)

Foster D R, Aber J D (Eds), 2004 *Forests In Time: The Environmental Consequences of 1000 Years of Change in New England* (Yale University Press, New Haven, CT)

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

Griffith D A, 1988 *Advanced Spatial Statistics* (Kluwer Academic, Dordrecht)

Hanson S, Guiliano G (Eds), 2004 *The Geography of Urban Transportation* 3rd edition (Guilford, New York)

Hanson S, Pratt G, 1995 *Gender, Work and Space* (Routledge, London)

Holden M, Lippitt C, Pontius R G Jr, Williams C, 2003, "Building a database of historic land cover to detect landscape change" *Biological Bulletin* **205** 257 – 258

Jantz C A, Goetz S J, Shelley M K, 2003, "Using the SLEUTH urban growth model to simulate the impacts of future policy scenarios on urban land use in the Baltimore – Washington metropolitan area" *Environment and Planning B: Planning and Design* **30** 251 – 271

King G, 1997 *A Solution to the Ecological Inference Problem* (Princeton University Press, Princeton, NJ)

Kok K, Farrow A, Veldkamp A, Verburg P H, 2001, "A method and application of multi-scale validation in spatial land use models" *Agriculture, Ecosystems and Environment* **85** 233 – 238

Lambin E F, Baulies X, Bockstael N, Fischer G, Krug T, Leemans R, Moran E F, Rindfuss R R, Sato Y, Skole D, Turner II B L, Vogel C, 1999, "Land-use and land-cover change implementation strategy", IGBP Report 48 IHDP Report 10, Royal Swedish Academy of Sciences, Stockholm

MacConnell W P, Niedzwiedz W, 1974, "Remote sensing 20 years of change in Worcester County Massachusetts 1951 – 1971", Report 10, Massachusetts Agricultural Experiment Station, University of Massachusetts, Amherst, MA

Neter J, Wasserman W, Kutner M H, 1983 *Applied Linear Regression Models* (Richard D Irwin, Homewood, IL)

Openshaw S, 1984 *The Modifiable Areal Unit Problem* (GeoBooks, Norwich)

Pontius R G Jr, 2000, "Quantification error versus location error in comparison of categorical maps" *Photogrammetric Engineering and Remote Sensing* **66** 1011 – 1016

Pontius R G Jr, 2002, "Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions" *Photogrammetric Engineering and Remote Sensing* **68** 1041 – 1049

Pontius R G Jr, Batchu K, 2003, "Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India" *Transactions in GIS* **7** 467 – 484

Pontius R G Jr, Malanson J, 2005, "Comparison of the accuracy of land change models: cellular automata Markov versus Geomod" *International Journal of Geographical Information Science* **19** 243 – 265

Pontius R G Jr, Malizia N, 2004, "Effect of category aggregation on map comparison", in *Lecture Notes in Computer Science 3234* Eds M J Egenhofer, C Freksa, H J Miller (Springer, Berlin) pp 251 – 268

Pontius R G Jr, Suedmeyer B, 2004, "Components of agreement in categorical maps at multiple resolutions", in *Remote Sensing and GIS Accuracy Assessment* Eds R S Lunetta, J G Lyon (CRC Press, Boca Raton, FL) pp 233 – 251

Pontius R G Jr, Cornell J, Hall C, 2001, "Modelling the spatial pattern of land-use change with GEOMOD2: application and validation for Costa Rica" *Agriculture, Ecosystems and Environment* **85**(1 – 3) 191 – 203

Pontius R G Jr, Agrawal A, Huffaker D, 2003, "Estimating the uncertainty of land-cover extrapolations while constructing a raster map from tabular data" *Journal of Geographical Systems* **5** 253 – 273

Pontius R G Jr, Huffaker D, Denman K, 2004, "Useful techniques of validation for spatially-explicit land-change models" *Ecological Modelling* **179** 449 – 461

Schneider L C, Pontius R G Jr, 2001, "Modeling land-use change in the Ipswich watershed Massachusetts USA" *Agriculture, Ecosystems and Environment* **85** 83 – 94

Silva E, Clarke K, 2002, "Calibration of the SLEUTH urban growth model for Lisbon and Porto Portugal" *Computers, Environment and Urban Systems* **26** 525 – 552

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