Statistical Methods to Partition Effects of Quantity and Location During Comparison of Categorical Maps at Multiple Resolutions

R. Gil Pontius, Jr.

Abstract

New generalized statistical methods to measure agreement between two maps at multiple-resolutions, where each cell in each map has a multinomial distribution among any number of categories, are presented. This methodology quantifies agreement between any two categorical maps, where either map uses fuzzy or crisp classification. The method measures the agreement at various resolutions by aggregating neighboring cells into an increasingly coarse grid. At each resolution, the method partitions the overall agreement into correct due to chance, correct due to quantity, correct due to location, error due to location, and error due to quantity. In addition, the method computes six statistics that are useful to interpret the differences between maps, and shows how these statistics change with resolution. This technique is particularly useful for characterizing land-cover change and for validating land-cover change models. For illustration, this paper applies these theoretical concepts to the validation of a land-use change model for Costa Rica.

Introduction

The Need for Useful Indicators of Goodness-of-Fit

This journal's special issue concerning Characterizing and Modeling Landscape Dynamics is an indication of the tremendous growth in the general field of landscape modeling. Our field abounds with variations on Markov Chain models, Cellular Automata models, agent-based models, multi-nomial logistic regression models, etc. In fact, we are now producing models faster than we can validate them.

After a scientist runs a land-change model, usually the first question is "How well did the model perform?" To address this question, usually the first approach is to perform a visual examination between the output map that the model produces and a reference map that has been reserved for validation. After a visual comparison, the scientist may choose to compute statistical measures of goodness-of-fit. The most useful indication of goodness-of-fit would inform the scientist on how to improve the model. For example, if the model can control explicitly for patch size, then it would be useful to compare the average patch size between the model's output map and the validation map. If the model cannot control for patch size, then such a comparison would not be directly useful.

The Problem of Categorical Map Comparison

In the case where the model's output map and the reference map show a categorical variable, the most basic questions are "Does the model produce the correct quantity of area in each category?" and "Does the model place the specific categories in the correct locations?" This paper presents methods to answer those two questions in a manner that corresponds to the intuition of visual inspection and that is useful to scientists who need to improve models.

Figure 1 illustrates the distinction between agreement due to quantity versus agreement due to location. Figure 1 shows two raster maps, each of which has 16 cells. Assume that the first map is from a simulation model and the second map is the reference map used for validation. There are two categories, say deforested versus surviving forest. Each cell shows the proportion of membership in the deforested category. At the fine resolution, each of the cells belongs entirely to one of the two categories, hence the proportion membership in the deforested category is either 0 or 1. Figure 1 also shows this same pair of maps at a coarser resolution whereby four adjacent cells from the fine resolution map are aggregated, hence each coarse cell can have partial membership simultaneously in the deforested category and the surviving forest category. Figure 1 shows the proportion of membership in the deforested category for each coarse cell.

At the finest resolution, the proportion of cells classified correctly in Figure 1 is 12/16. The overall error is 4/16 and has two components: error due to quantity and error due to location. The proportion of the deforested category in the simulated map is 10/16 and in the reality map is 12/16, so there is an error in quantity of 2/16. The error of location is attributable to the fact that it is possible to swap the locations of a pair of cells in the simulated map in order to improve its agreement with the reality map. Specifically, in the simulated map, if we were to swap the location of the cell in row 3 column 1 with the cell in row 3 column 2, then the agreement between the simulated map and the reality map would improve from 12/16 to 14/16. After this swap, it would be impossible to perform additional swaps to improve agreement, because all of the remaining error would be due to quantity. Therefore, the error due to location is 2/16 at the finest resolution. At the coarse resolution, it is impossible to perform any swaps among the locations of four coarse cells in the simulated map in order to improve agreement with the reality map; therefore, at the coarse resolution, all of the error is due to quantity, which remains at 2/16.

Department of International Development, Community and Environment, Graduate School of Geography, Clark University, 950 Main Street, Worcester, MA 01610 (rpontius@clarku.edu)

Photogrammetric Engineering & Remote Sensing
Vol. 68, No. 10, October 2002, pp. 1041–1049.

0099-1112/02/6810–1041$3.00/0

Fine Resolution Maps

Simulated				Reality			
0	0	0	1	0	0	1	1
0	1	1	1	0	1	1	1
1	0	1	1	0	1	1	1
0	1	1	1	1	1	1	1

Coarse Resolution Maps

Simulated		Reality	
0.25	0.75	0.25	1.00
0.50	1.00	0.75	1.00

Figure 1. This shows an example of a map from a simulation model versus a reference map used for validation. The value in each grid cell is the proportion of membership in a category. When four adjacent cells of the fine resolution maps are averaged, the result is a coarse resolution cell that can show a partial membership in the category.

With this example in mind, let us examine the general case of comparison of two raster maps. Each cell in each map is classified as one of J categories. The most common (non-spatial) statistics are functions of a contingency table or confusion matrix, where the columns of the table have categories of one map and the rows are categories of the other map. Each entry in the table is the proportion (or number of cells) of the study area that falls into the combination of categories in each map. The contingency table yields familiar statistics, such as Chi-square, phi, tau, and kappa. GIS professionals use additional statistics, such as user's accuracy and producer's accuracy (Congalton, 1991; Congalton and Green, 1999). However, the basis for all these statistics is cell-by-cell agreement between the two maps, because the confusion matrix contains information about only cell-by-cell agreement, where each cell is crisp classified. The confusion matrix fails to distinguish between a near miss and a far miss. In other words, the confusion matrix records zero agreement when a cell is not classified correctly, even when the correct category is found in the neighboring cell, or even when the correct category is found nowhere near the cell. Also, the standard confusion matrix is not designed to account for partial success when the cell has partial membership in a category. Furthermore, the analysis of the confusion matrix usually treats the marginal totals as fixed; therefore, it confounds accuracy due to quantity and accuracy due to location (Pontius, 2000). For purposes of categorical map comparison, it would be better (1) to give some partial agreement for a near miss and less agreement for a far miss, (2) to be able to apply the comparison method to maps in which the cells are fuzzy classified, and (3) to separate agreement due to quantity from agreement due to location. This paper gives a method to accomplish all three of these goals with one approach, whereas other approaches accomplish only one or two of these goals. Therefore, this paper answers the numerous calls for research into this type of accuracy assessment (Wang, 1990; Gopal and Woodcock, 1994; Edwards and Lowell, 1996; Lambin *et al.*, 1999).

Previous Approaches

Others have derived measures of multiple resolution goodness-of-fit in order to compare spatial patterns in landscapes. These methods generate windows at various resolutions, then plot the agreement within the window as a function of window size. Kok *et al.* (2001) compare maps of landscape change, in which there is an increase or decrease of each land type in each grid cell. They measure at various resolutions the extent to which the quantities of change in the cells of one map correlate with the quantities of change in the corresponding cells in another map. Turner *et al.* (1989) and Costanza (1989) offer additional methods of categorical map comparison at multiple resolutions, including methods to integrate measures at several resolutions into one overall measurement of agreement. Turner and Costanza apply their methods to crisp classification schemes.

Fuzzy classification gives the potential to be more descriptive than does crisp classification because fuzzy classification can show more information than crisp classification (Heuvelink and Burrough, 1993; Foody, 1999). Some researchers have developed methods to quantify agreement using fuzzy classification in order to improve estimates of quantity of land types (Woodcock and Gopal, 2000). Hay (1988) and Jupp (1989) use the confusion matrix to improve accuracy of quantity, but these methods do not incorporate proximity, nor do they analyze the accuracy of location separate from the accuracy of quantity.

A few researchers separate explicitly the accuracy of location from the accuracy of quantity. Monserud and Leemans (1992) use kappa to measure accuracy of location, and other methods to measure accuracy of quantity. Pontius (2000) shows how to separate the agreement due to quantity versus location for comparison of categorical maps; however, that method works for only a single resolution and only crisp classification.

As mentioned above, we could compare two categorical maps according to a variety of criteria, including average patch size, perimeter-to-area ratio, contagion, patch shape, etc. There are a variety of metrics for each of these characteristics of pattern structure. Ritters *et al.* (1995) describe and perform factor analysis on 26 such metrics. However, before a researcher examines these details of pattern structure, there are usually two more fundamental initial questions: How well do the maps agree in terms of the quantity in each category? How well do the maps agree in terms of the general location of each category?

This paper presents statistical methods that answer these fundamental questions with an approach that analyzes maps at multiple resolutions, works for both crisp and fuzzy classification, and partitions the agreement according to correct due to chance, correct due to quantity, correct due to location, error due to location, and error due to quantity. This paper describes, illustrates, and gives equations for all of these statistics and several derivative statistics that modelers will find helpful in model development.

Methods

Data Format

In order to illustrate the usefulness of the proposed methods to dynamic landscape modeling, this paper illustrates the methods with data from Costa Rica, for which there are two categories of interest, deforestation and surviving forest. Figure 2 shows a raster map of real deforestation from 1940 to 1983 (Sader and Joyce, 1988). For this example, we consider Figure 2 to be reality, which means it is our reference map of high accuracy. Of the cells that were forest in 1940, 70 percent became deforested between 1940 and 1983. Cells that are non-forest in 1940 are not part of this analysis because they are not candidates for new deforestation between 1940 and 1983.

Figure 3 shows a raster map of predicted deforestation simulated by a land-use change model similar to GEOMOD2 (Pontius *et al.*, 2001). GEOMOD2 is similar to many other land-use change models in which the user specifies the overall quantity of change and GEOMOD2 specifies the location of change based on a variety of biophysical and social characteristics.

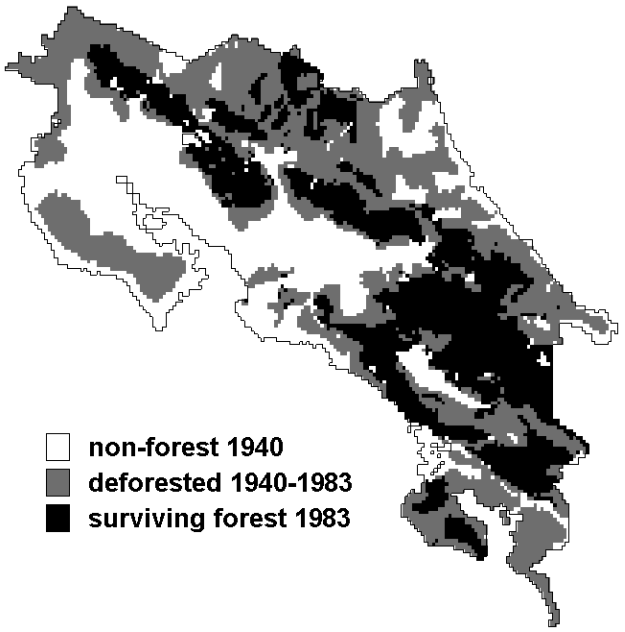

Figure 2. This map of real deforestation in Costa Rica shows that 70 percent of the 1940 forest area became deforested from 1940 to 1983. Each cell is two kilometers by two kilometers.

According to this particular run of the simulation model, there is 58 percent deforestation between 1940 and 1983. This specification of the quantity of deforestation is based on extrapolation of historical information. Therefore, we see that the model makes errors in terms of both the quantity of deforestation and the location of the deforestation. The focus here is on map comparison, regardless of the method of creation of the simulated landscape. It is not the purpose of this paper to discuss the method of simulation. Whatever the method of simulation, we will assume that the simulation model can make errors in terms of both the quantity of each predicted category and the location of each predicted category. Note that some models are calibrated such that they make no errors in the quantity of each category, in which case the methods of this paper still apply but are not as interesting.

This paper gives general methods to quantify the agreement between two maps, such as Figure 2 versus Figure 3. However, this paper describes statistical methods that apply to cases that are much more general than Figure 2 versus Figure 3. For example, the method can compare any two maps where the categories are classified as any combination of fuzzy and crisp categories. Specifically, the methods can compare any two maps, denoted R for reality and S for simulated, that meet the following criteria. First, both map R and map S must have the same grid cell structure. Second, in each grid cell, there must be a multinomial distribution of categories such that Equations 1 through 4 hold, where j is a category, J is the number of categories, Rn,j is the proportion of category j in grid cell n of map R, Sn,j is the proportion of category j in grid cell n of map S, and Ng is the number of grid cells in the map at resolution g: i.e.,

$$0 \leq Rn,j \leq 1 \tag{1}$$

$$0 \leq Sn,j \leq 1 \tag{2}$$

$$\sum_{j=1}^{J} Rn,j = \sum_{j=1}^{J} Sn,j = 1 \tag{3}$$

$$\sum_{n=1}^{Ng} \sum_{j=1}^{J} Rn,j = \sum_{n=1}^{Ng} \sum_{j=1}^{J} Sn,j = Ng \tag{4}$$

Partial Agreement and Multiple Resolutions

The key to convert nearly any conventional statistic to a multiple resolution statistic is to allow for partial agreement between the cell of one map and the corresponding cell of the other map. In order to do this, one must allow for each cell to have partial membership in any of the categories. Let us define the agreement for category j in cell n to be the minimum of Rn,j and Sn,j (Prentice *et al.*, 1992). Equation 5 gives the total agreement over all J categories in any single cell: i.e.,

$$\text{proportion agreement in cell } n = \sum_{j=1}^{J} \text{MIN}(Rn,j,\ Sn,j) \tag{5}$$

A spatial aggregation technique generates a sequence of increasingly coarse resolutions. For our notation, the resolution is the length of a grid cell side. The finest resolution is the resolution in which the model creates its output map. The next resolution uses a 2 by 2 grid to aggregate four of the finest resolution neighboring cells, the subsequent resolution uses a 3 by 3 grid to aggregate nine of the finest resolution neighboring cells, and so on until the coarsest resolution aggregates all the finest resolution cells of the entire study area into one cell.

When one uses this aggregation technique, if there are many near misses in the comparison at the finest resolution, then the agreement between the maps will rise rapidly in the early stages of aggregation. If there are many far misses, then the agreement will not rise until later stages of aggregation. At

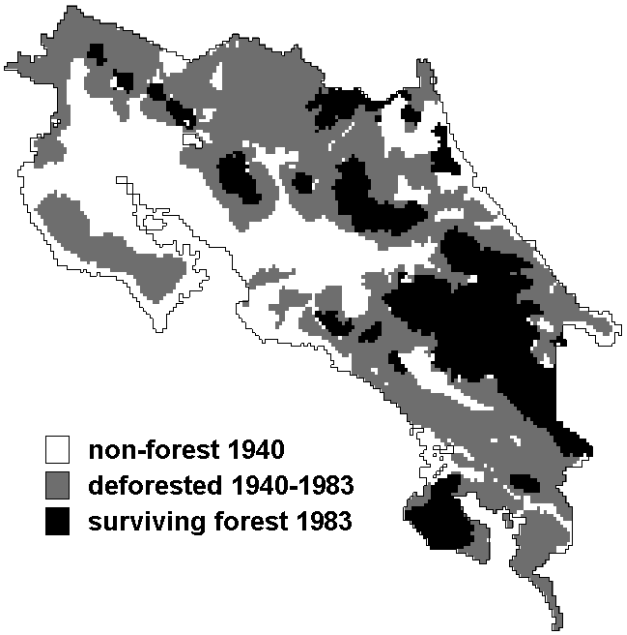

Figure 3. In this map of simulated deforestation in Costa Rica, 58 percent of the 1940 forest area is deforested between 1940 and 1983. Cells that are non-forest in 1940 are identical to those in Figure 2 and are not candidates for deforestation from 1940 to 1983.

the coarsest resolution, where there is just one cell, the agreement is a function of only the quantity of each category in each map; hence, location among the fine resolution cells plays no role in agreement.

If the study area is not perfectly square, as is the case in the Costa Rica example, then the aggregation technique will produce coarse resolution cells that are made up of different numbers of fine resolution cells. Therefore, it is important to weight each cell according to its importance in the analysis. A natural selection for each weight, *Wn*, is the number of fine resolution cells that constitute a coarse cell, *n*. Equation 6 gives the agreement between map *R* and map *S*, where each cell *n* has a weight *Wn*. Equation 6 applies to any specific resolution: i.e.,

$$\textit{total agreement at resolution } g = \frac{\sum_{n=1}^{Ng}\left[Wn \sum_{j=1}^{J} \text{MIN}(Rn,j,\ Sn,j)\right]}{\sum_{n=1}^{Ng} Wn} \tag{6}$$

Each resolution yields statistics of agreement, but it is also helpful to have one measure of overall agreement. Therefore, one can combine the level of agreement at each resolution into one measurement of agreement by taking a weighted average of agreement at each resolution. If it is important to measure agreement between the maps at precise locations, then the user should use large weights for fine resolutions. Equation 7 gives total agreement, where *Vg* is the weight for each map resolution *g*. The resolution weights are the most subjective component of this analysis. A natural selection for each *Vg* is the sum of the individual cell weights at resolution *g*, in which case total agreement simplifies to Equation 8, which is what the Costa Rica example uses.

$$\textit{weighted total agreement} = \left(\sum_{g=1}^{G} Vg \left\{ \frac{\sum_{n=1}^{Ng}\left[Wn \sum_{j=1}^{J} \text{MIN}(Rn,j,\ Sn,j)\right]}{\sum_{n=1}^{Ng} Wn} \right\}\right) \Bigg/ \sum_{g=1}^{G} Vg \tag{7}$$

$$\textit{simplified total agreement} = \left\{ \sum_{g=1}^{G} \sum_{n=1}^{Ng}\left[Wn \sum_{j=1}^{J} \text{MIN}(Rn,j,\ Sn,j)\right]\right\} \Bigg/ \sum_{g=1}^{G} \sum_{n=1}^{Ng} Wn \tag{8}$$

Location versus Quantity

Figure 4 shows nine mathematical expressions for the agreement between nine pairs of maps. The expression in middle column **m** and middle row M(**x**) gives the agreement M(**m**) between the reality map and the simulated map. The other eight expressions are idealized agreement between the reality map and a hypothetical simulated map. Each expression is based on the reality map and a combination of information that could be incorporated in the simulated map.

The horizontal axis of Figure 4 shows three levels of agreement in terms of the quantities of categories found in the simulated map. The left-most column **n** shows agreement of hypothetical simulations that have no information concerning quantity; thus, the proportion of every category *j* in the hypothetical simulated map is 1/*J* (Foody, 1992). The middle column **m** shows simulations that have some medium level of information concerning quantity. In general, **m** denotes the proportions of the categories in the simulated map. For our analysis of Figure 3, the deforestation category is 58 percent and the surviving category is 42 percent. The right column shows hypothetical simulations that have perfect information concerning quantity; thus, the proportion of each category *j* in the simulated map is equal to the proportion of category *j* in the reality map. Equation 9 gives the proportion of category *j* in the reality map, *R.,j* and Equation 10 gives the proportion of category *j* in the simulated map, *S.,j*: i.e.,

$$R.,j = \frac{\sum_{n=1}^{Ng} [Wn\ Rn,j]}{\sum_{n=1}^{Ng} Wn} \tag{9}$$

$$S.,j = \frac{\sum_{n=1}^{Ng} [Wn\ Sn,j]}{\sum_{n=1}^{Ng} Wn} \tag{10}$$

The vertical axis of Figure 4 shows three levels of agreement in terms of location between the reality map and simulated maps. The bottom row shows agreement for simulations that have no information concerning location; thus, an identical multinomial distribution of categories exists within all grid cells of the hypothetical simulated map. A simulated map that has no information of location is a map in which each category is distributed evenly across the landscape. The middle row shows agreement for simulations that have a medium level of information concerning location, as determined by the particular comparison we are making, which is Figure 2 versus Figure 3 in our example. In general, the level of the medium information of location depends on the simulation's ability to place categories at the proper locations, as indicated by the direct comparison between the simulated map and the reality map, denoted by M(**m**). Table 1 gives the formula for Klocation. The top row shows agreement of hypothetical simulations that have perfect information concerning location, so the given quantity of category *j* in the simulated map is placed as best as possible to match the locations in the reality map. If a simulated map has perfect location, then the only source of disagreement between it and the reality map is a difference in quantity.

For any single pair of maps, one can compute all nine of the points shown in Figure 4 for every resolution. Let each of the nine points be denoted by the notation (**x**,y), where **x** is a vector that gives the proportion of each category in the simulated map and y is the proportion agreement when the simulated map is compared to the reality map. Each of the three columns of Figure 4 refers to a combination of quantities in the simulated map. Let **x** = **n** for the left column; let **x** = **m** for the middle column; let **x** = **p** for right column. Let each of the rows of Figure 4 refer to a function of **x**: N(**x**) for the bottom row, M(**x**) for the middle row, and P(**x**) for the top row. The agreement between the reality map and the simulation map is given always by M(**m**). P(**p**) is always 100 percent. The other seven points are levels of expected agreement, which are derived statistically. The map of reality alone determines N(**n**), N(**p**), and P(**n**). The points in rows N(**x**) and M(**x**) change with resolution, and usually increase with increasing coarseness. The points in row P(**x**) do not change with resolution. All of the more advanced statistics given in this paper are functions of the nine crucial points of Figure 4. Pontius (2000) describes in depth the logic of Figure 4 for a single resolution where the categorization is crisp. The nine points partition an agreement space, which is described next.

Agreement Space

Figure 5 shows percent agreement versus quantity of deforestation in the simulated map. For any particular resolution, we

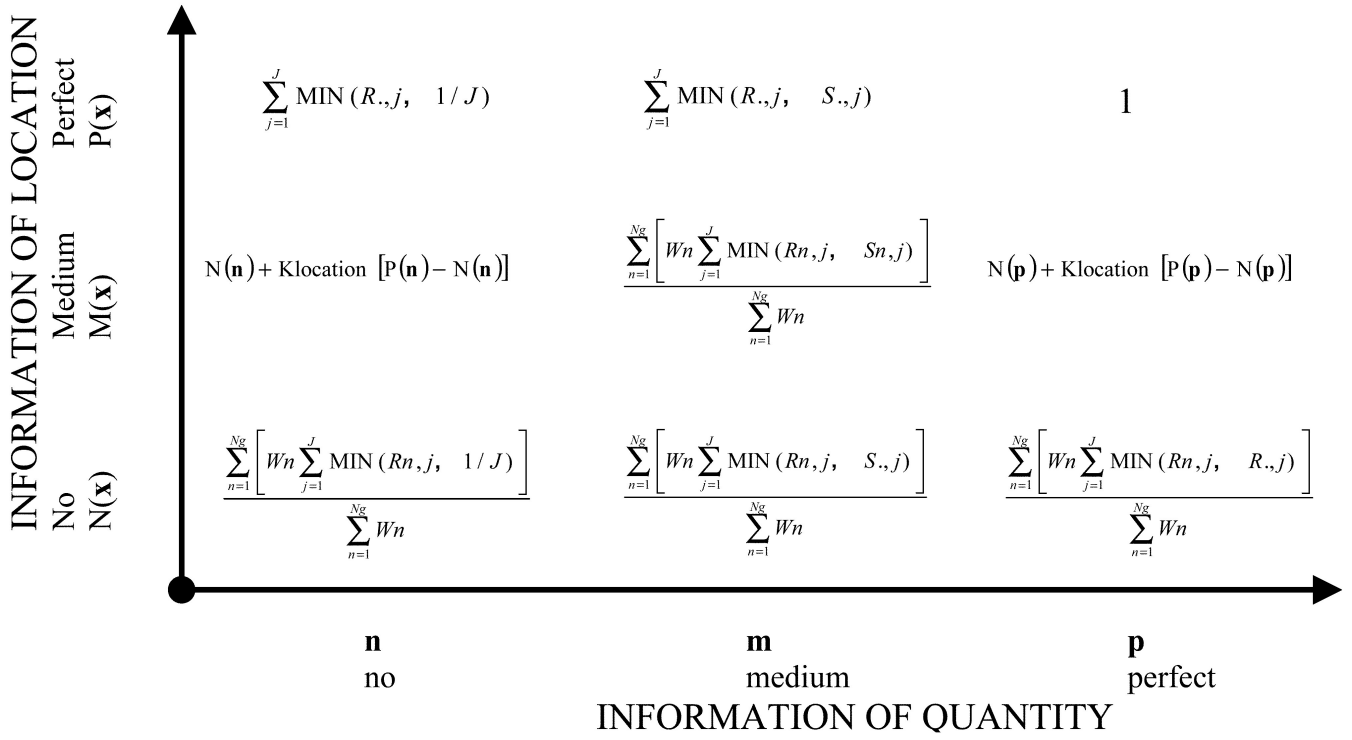

Figure 4. This figure gives mathematical expressions for nine points from which all other statistics of this paper derive. The expression in the middle column and middle row gives the agreement between the reality map and the simulated map at resolution *g*, where the variables are defined in the text. The other eight expressions are idealized agreement between the reality map and a simulated map, based on the combination of information available concerning quantity and location. All nine points are plotted in Figure 5.

could generate a figure similar to Figure 5. However, Figure 5 combines all resolutions into one figure, using the weighting method that places larger weights on fine resolutions as discussed above. Figure 5 shows the nine quantities in the same relative positions as given in Figure 4. The three points at 50 percent simulated deforestation show hypothetical simulations that have no information of quantity, that is, $\mathbf{x} = \mathbf{n}$; points at 58 percent simulated deforestation show simulations that have medium information of quantity, that is, $\mathbf{x} = \mathbf{m}$; points at 70 percent simulated deforestation show hypothetical simulations that have perfect information of quantity, that is, $\mathbf{x} = \mathbf{p}$.

TABLE 1. FORMULAS FOR INDICES OF MAP COMPARISON, INCLUDING RESULTS FOR THE EXAMPLES. ALL VALUES ARE IN PERCENT. FOR THE COSTA RICA EXAMPLE, THE VALUES ARE WEIGHTED OVER MULTIPLE RESOLUTIONS, WITH LARGER WEIGHTS AT FINER RESOLUTIONS

Parameter	Formula	Figure 1 Fine Resolution	Figure 1 Coarse Resolution	Costa Rica Reality Versus Simulation
Percent Correct	$\mathrm{M}(\mathbf{m})$	75	88	74
Kno	$\frac{\mathrm{M}(\mathbf{m}) - \mathrm{N}(\mathbf{n})}{\mathrm{P}(\mathbf{p}) - \mathrm{N}(\mathbf{n})}$	50	67	44
Klocation	$\frac{\mathrm{M}(\mathbf{m}) - \mathrm{N}(\mathbf{m})}{\mathrm{P}(\mathbf{m}) - \mathrm{N}(\mathbf{m})}$	60	100	54
Kquantity	$\frac{\mathrm{M}(\mathbf{m}) - \mathrm{M}(\mathbf{n})}{\mathrm{M}(\mathbf{p}) - \mathrm{M}(\mathbf{n})}$	50	50	42
Kstandard	$\frac{\mathrm{M}(\mathbf{m}) - \mathrm{N}(\mathbf{m})}{\mathrm{P}(\mathbf{p}) - \mathrm{N}(\mathbf{m})}$	43	60	40
VPIQ	$\mathrm{M}(\mathbf{p}) - \mathrm{M}(\mathbf{m})$	10	13	8
VPIL	$\mathrm{P}(\mathbf{m}) - \mathrm{M}(\mathbf{m})$	13	0	14

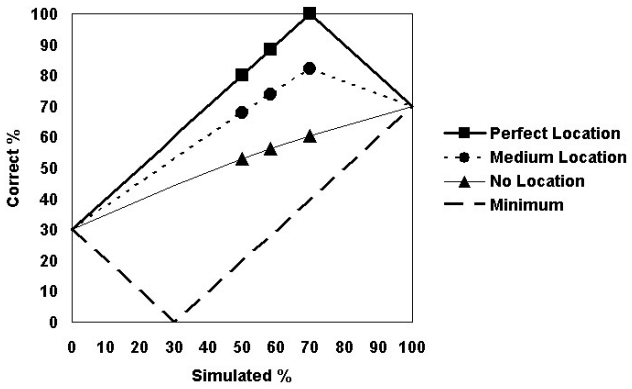

Figure 5. This agreement space shows percent correct classification versus percent of simulated deforestation. The relative positions among the nine points are in the same row and column arrangement as the mathematical expressions of Figure 4. We could make a figure such as this one for each particular resolution. However, Figure 5 is the one that results when all resolutions are combined by a weighted linear combination, with larger weights for finer resolutions.

The lines in Figure 5 show agreement as a function of the percent of deforestation in the simulation. The upper "Perfect Location" solid line shows the maximum agreement between the simulated map and the reality map. The points on the "Perfect Location" line correspond to hypothetical simulations that have perfect agreement in terms of location, which are in the P(**x**) row of Figure 4. The "No Location" line shows the agreement when the percent of simulated deforestation is assigned identically to every grid cell. The points on the "No Location" line correspond to scenarios that have chance agreement in terms of location, which are in the N(**x**) row of Figure 4. The points on the "Medium Location" line correspond to simulations that have a medium ability to specify location, which are in the M(**x**) row of Figure 4. The central circular point shows the agreement M(**m**) between the reality map (Figure 2) and the simulated map (Figure 3). The "Perfect Information" and "Medium Information" lines increase in percent correct as they approach perfect information of quantity. The "No Information" line shows that if the location of each category is distributed evenly across the landscape, then percent correct is maximized when the quantities are at extremes, in this case, 100 percent deforestation and 0 percent surviving forest.

Figure 5 portrays a two-dimensional slice through a three dimensional agreement space, where the three axes are (1) percent agreement, as a function of (2) resolution size, and (3) percent of a specific land type in the simulated map. For the Costa Rica example, the agreement space is three-dimensional because there are two categories, deforestation and surviving forest. For other applications, the agreement space would increase by one dimension for every additional land type. The axis of the additional dimension would be the percent of the additional land type in the simulated map.

Figure 6 shows a slice through the same three-dimensional agreement space as Figure 5. Figure 6 is orthogonal to Figure 5 and passes through the plane where percent deforested = 58 percent. The axes of Figure 6 are percent agreement versus resolution. Figure 6 shows the agreement between the simulated map and the reality map at various resolutions. For each resolution, the components of agreement are separated into proportion correct due to chance = N(**n**), proportion correct due to quantity = N(**m**) − N(**n**), proportion correct due to location M(**m**) − N(**m**), proportion error due to location = P(**m**) − M(**m**), and proportion error due to quantity = P(**p**) − P(**m**).

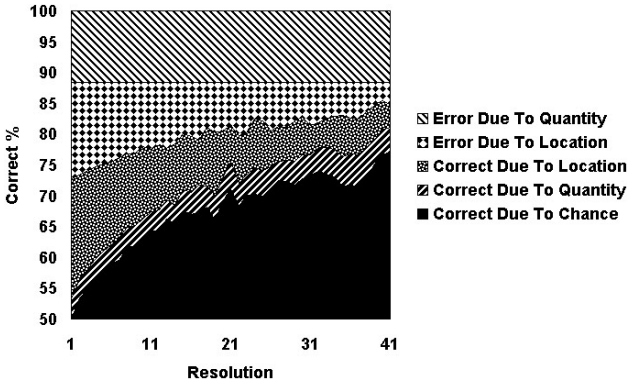

Figure 6. This agreement space shows percent correct classification versus resolution for the agreement between the map of reality (Figure 2) and the simulated map (Figure 3) of Costa Rica. The unit of the resolution axis is the length of the side of a coarse grid cell, expressed as a multiple of the length of the side of the fine resolution cell, so that the resolution of 1 is the finest and 41 is the coarsest. Notice how the proportion correct due to location and the proportion error due to location both shrink as the resolution becomes coarser.

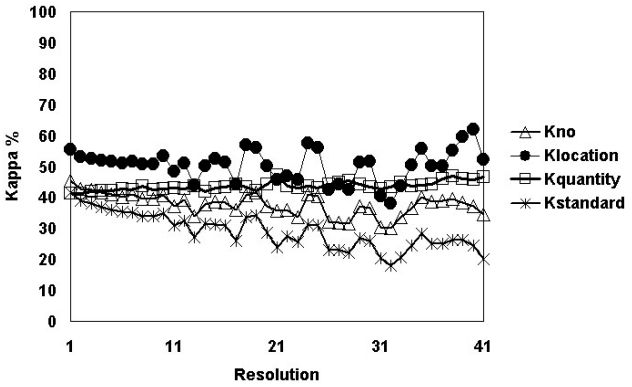

Figure 7. This agreement space shows variations in the kappa index of agreement versus resolution for the agreement between the map of reality (Figure 2) and the simulated map (Figure 3) for Costa Rica. The resolution of 1 is the finest and 41 is the coarsest, as in Figure 6.

Kappa

Figures 5 and 6 show that random chance can generate large agreement, especially when there are a small number of categories. Therefore, it is helpful to incorporate the expected proportion correct classification due to chance in an index of agreement (Cohen, 1960; Brennan and Prediger, 1981; Rosenfield, 1986; Hudson and Ramm, 1987). Equation 11 gives one of the most popular indices, kappa, where *Po* is the observed proportion correct which we denote as M(**m**), *Pc* is the expected proportion correct due to chance, and *Pp* is the proportion correct when classification is perfect. Thus, kappa is 1 when observed agreement is perfect; kappa is 0 when observed agreement is equal to the expected agreement due to chance. In the standard kappa, *Pc* is the quantity N(**m**) and *Pp* is the quantity P(**p**) given in Figure 4: i.e.,

$$kappa = \frac{(Po - Pc)}{(Pp - Pc)} \tag{11}$$

In addition to the standard kappa index of agreement, Pontius (2000) defines three variations: Kappa for no information (Kno), Kappa for location (Klocation), and Kappa for quantity (Kquantity). Kno is an overall index of agreement for which *Pc* = N(**n**) and *Pp* = P(**p**). Klocation is an index that measures the agreement in terms of location only, where *Pc* = N(**m**) and *Pp* = P(**m**). Kquantity measures the agreement in terms of quantity, where *Pc* = M(**n**) and *Pp* = M(**p**). Table 1 gives the formulas for these variations on kappa. Figure 7 shows how these indices of agreement change with resolution for the comparison between reality (Figure 2) and the simulated map (Figure 3) for Costa Rica.

Value of Perfect Information

Two additional indices of agreement are value of perfect information of quantity (VPIQ) and value of perfect information of location (VPIL), defined in Table 1. VPIQ is the additional increase in agreement that would be attained if the quantity in the simulated map were to match the quantity in the reality map, given the initial agreement M(**m**) and assuming a constant Klocation. VPIL is the additional increase in agreement

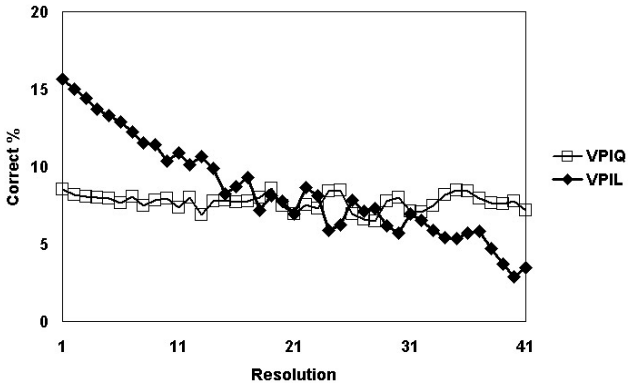

Figure 8. This agreement space shows the value of perfect information of quantity (VPIQ) and the value of perfect information of location (VPIL) versus resolution for the agreement between the map of reality (Figure 2) and the simulated map (Figure 3) of Costa Rica. The resolution of 1 is the finest and 41 is the coarsest. Notice that for finer resolutions, VPIQ is less than VPIL, and vice-versa.

that would be attained if the location in the simulated map were to match as closely as possible the reality map, given the initial agreement M(**m**) and assuming no change in the quantity in the simulated map. Figure 8 shows how these indices of agreement change with resolution for the comparison between reality (Figure 2) and the simulated map (Figure 3) for Costa Rica.

A Simplified Example

In order to guarantee that the reader has followed the logic of this analysis, I recommend that the reader work through the simplified example shown in Figure 1. For the fine resolution, Table 2 shows values of the nine crucial points in the same arrangement as their corresponding equations in Figure 4. Table 3 shows the same points for the coarse resolution maps in Figure 1. All the statistics of this paper derive directly from these nine points, so it is crucial to understand the logic of Figure 4 and Tables 2 and 3.

At the fine resolution, the overall agreement is M(**m**) = 12/16 and the error due to location is P(**m**) − M(**m**) = 2/16. This error of location is attributable to the fact that it is possible to swap the locations of a pair of cells in the simulated map in order to improve its fit with the reality map. At the coarse resolution, the overall agreement is M(**m**) = 14/16 and the error due to location is P(**m**) − M(**m**) = 0. At the coarse resolution, it is impossible to perform any swaps among locations of cells in the simulated map to improve agreement with the reality map. The near misses at the fine resolution are the reason for the perfect agreement in terms of location at the coarse resolution. At all resolutions, error due to quantity is P(**p**) − P(**m**) = 2/16. Table 1 gives all of the statistics that derive from the nine crucial points shown in Table 2.

Table 2. Values for the Nine Crucial Points to Compare the Simulated Map versus the Reality Map of Figure 1 at the Fine Resolution

Information of Location	Information of Quantity		
	No	Medium	Perfect
Perfect	12/16	14/16	16/16
Medium	10.4/16	12/16	13.6/16
No	8/16	9/16	10/16

Table 3. Values for the Nine Crucial Points to Compare the Simulated Map versus the Reality Map of Figure 1 at the Coarse Resolution

Information of Location	Information of Quantity		
	No	Medium	Perfect
Perfect	3/4 = 12/16	3.5/4 = 14/16	4/4 = 16/16
Medium	3/4 = 12/16	3.5/4 = 14/16	4/4 = 16/16
No	2.5/4 = 10/16	2.75/4 = 11/16	3/4 = 12/16

Results

Costa Rica Example

For the Costa Rica example, the overall agreement, denoted M(**m**), between the reality map (Figure 2) and the simulated map (Figure 3) is 74 percent, when the analysis is weighted over all resolutions, with higher weights for finer resolutions. Figures 5 and 6 show that the partitioning of agreement is 53 percent correct due to chance, 3 percent correct due to quantity, 18 percent correct due to location, 14 percent error due to location, and 12 percent error due to quantity. Table 1 shows that the values of kappa and its variations are near or slightly less than 0.5, which indicates the model is near or slightly less than half way between (a) the level of agreement expected by chance and (b) perfect agreement. The value of perfect information of location (VPIL) is 14 percent, and the value of perfect information of quantity (VPIQ) is 8 percent, which means that the modeler will gain more agreement from improving the model's ability to specify location than from improving the model's ability to specify quantity.

Figure 6 shows how percent agreement increases as resolution becomes coarser from 1 to 41 grid cells per side of each coarse grid cell. At the finest resolution, there are 9011 cells and, at a resolution of 41, there are 16 coarse grid cells. When the resolution reaches 199, there is exactly one coarse cell. Percent agreement increases from 73 percent to its maximum of 88 percent as one moves from the finest resolution to the coarsest resolution. At the finest resolution, correct due to chance is 50 percent, correct due to quantity is 3 percent, correct due to location is 19 percent, error due to location is 16 percent, and error due to quantity is 12 percent. As resolution becomes coarser, correct due to chance tends to increase, correct due to quantity varies unsystematically, correct due to location decreases, error due to location decreases, and error due to quantity remains constant.

It is important to examine how the various kappa indices of agreement change with resolution, because the percent agreement due to chance tends to increases with resolution. At the finest resolution, Kno is 45 percent, Klocation is 55 percent, Kquantity is 41 percent, and Kstandard is 42 percent. As resolution becomes coarser, Kno and Kstandard decrease while Klocation and Kquantity do not vary systematically (Figure 7).

Figure 8 shows that at the finest resolution VPIQ is 9 percent, and VPIL is 16 percent. As resolution becomes coarser, VPIQ remains steady while VPIL approaches zero. At resolutions more coarse than 29, VPIQ is larger than VPIL.

Discussion

Guidance for Model Improvement

The Results section gives statistics that are helpful to scientists who are trying to produce maps with high levels of agreement. Modelers and remote sensing specialists need to know how to adjust the simulation or classification methodology to increase accuracy.

In the Costa Rica example, VPIQ and VPIL show that an improvement in specification of location is likely to be more helpful in increasing accuracy than will an improvement in

specification of quantity at fine resolutions. However, at resolutions greater than 29 by 29 cells, improvement in specification of quantity becomes more important than improvement in specification of location. These types of indices can help modelers decide how to focus their modeling efforts. For example, if the output from a land-use change model will serve as input to other models, then the modeler should examine the land-use change model's VPIL and VPIQ at the resolution of the other models.

Furthermore, Figure 6 shows that, at the finest resolution, the error due to location accounts for 16 percent of the landscape, but at a resolution of 21 the error due to location accounts for about 8 percent of the landscape. This means that half of the errors of location happen over distances less than 42 kilometers, because the resolution of each cell is two kilometers.

Application to Landscape Characterization

With a trivial change in the definition of the categories in the maps, the techniques of this paper are extremely well suited to characterize general land-cover change over time. For example, in the Costa Rica case, suppose that in both the "simulated" map (Figure 2) and "reality" map (Figure 3), white meant out of study area, gray meant non-forest, and black meant forest. Also, assume that the "simulated" map is a map from some point in time, say 1960, and the "reality" map is of 1983. If this were the case, then the exact same calculations would compare the map of 1960 to the map of 1983. The resulting values of the statistics of agreement would be the same; however, the interpretation would characterize the change between 1960 and 1983. Total agreement is $M(\mathbf{m}) = 0.74$, which means that 74 percent of the landscape shows no change between 1960 and 1983. Of the remaining 26 percent that shows change, 14 percent of the change is attributable to the change in location between the two years and 12 percent of the change is attributable to changes in quantity.

In a land-cover change characterization, changes in location result from swapping locations between forest and non-forest cells. For example, suppose that between 1960 and 1983 there is forest regrowth on many non-forest cells and deforestation on an equal number of forest cells. In this case, all of the disagreement between the 1960 map and the 1983 map would be due to location, while none of the disagreement would be due to quantity. Alternatively, if between 1960 and 1983 there is no forest regrowth and some forest becomes deforested, then all of the disagreement between 1960 and 1983 would be due to quantity and none of the disagreement would be due to location.

This type of characterization is extremely important to scientists who want to understand and to model landscape dynamics. Even before running a simulation model, scientists can characterize the landscape change at any desired resolution, particularly at the resolution of the model. If the landscape characterization shows that change in location is larger than change in quantity, then the scientist should focus most of the modeling effort on specification of location and relatively less effort on specification of quantity. Conversely, if the landscape characterization shows that change in quantity is larger than change in location, then the scientist should focus most of the modeling effort on specification of quantity and relatively less effort on specification of location.

Alternative Indicators

This paper's techniques could compliment or replace conventional measures of goodness of fit. For example, R-squared is a common measure of agreement between empirical data and a regression surface. However, R-squared is almost always low and insensitive for logistic regression, when the dependent variable is binary (0 or 1) and the predicted variable is a probability. Furthermore, the standard R-squared fails to account for proximity and fails to partition the agreement and error according to quantity and location. The techniques in this paper are well suited to compare maps where the cells in one map are binary variables and the cells in the other map are modeled probabilities.

This paper presents methods to address two fundamental questions in any categorical map comparisons as mentioned in the introduction: Do the maps agree in quantity of each category? Do the maps agree in the location of each category? After the scientist addresses these questions, then more detailed questions arise, such as "Do the two maps look similar in terms of fractal dimension, patchiness, etc.?" In order to examine these issues, the scientist could compare the two maps in terms of landscape pattern indices such as those described by Ritters *et al.* (1999). However, it is recommended that the more basic methods of this paper be performed first, because the quantity and the general location of the mapped categories can constrain and influence substantially other indices of landscape pattern.

Conclusions

This paper gives mathematical expressions to enable scientists to separate overall classification agreement into (1) agreement associated with quantity versus (2) agreement associated with location. Moreover, this paper shows how to partition these components of agreement over multiple resolutions. It is important to examine multiple resolutions so that map comparison can take into consideration spatial proximity to agreement, and not be constricted to cell-by-cell agreement. These statistics can be used to compare any two maps of crisp or fuzzy categorical variables; hence, they should be useful for GIS modelers, remote sensors, and any other scientists who want to compare categorical maps.

Acknowledgments

The National Science Foundation (NSF) supported this research through the Water and Watersheds program grant DEB-9726862. Additional contributors include the Jesse B. Cox Charitable Trust, the Sweetwater Trust, and NSF's Long Term Ecological Research program OCE-9726921. Also, I thank Pablo Pacheco, Stephen Aldrich, anonymous reviewers, the Marsh Institute of Clark University, and the Center for Integrated Studies of Global Environmental Change at Carnegie Mellon University, with which this work is increasingly tied intellectually and programmatically. Hao Chen of Clark Labs is programming the techniques of this paper into the software *Idrisi*.

References

Brennan, R.L., and D.J. Prediger, 1981. Coefficient kappa: Some uses, misuses and alternatives, *Educational and Psychological Measurement*, 41:687–698.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1):37–46.

Congalton, R., 1991. A review of assessing the accuracy of classification of remotely sensed data, *Remote Sensing of the Environment*, 37:35–46.

Congalton, R., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, New York, N.Y., 137 p.

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

Edwards, G., and K.E. Lowell, 1996. Modeling uncertainty in photointerpreted boundaries, *Photogrammetric Engineering & Remote Sensing*, 62(4):337–391.

Foody, G., 1992. On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering & Remote Sensing*, 58(10):1459–1460.

———, 1999. The continuum of classification fuzziness in thematic mapping, *Photogrammetric Engineering & Remote Sensing*, 65(4):443–451.

Gopal, S., and C.E. Woodcock, 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets, *Photogrammetric Engineering & Remote Sensing*, 60(2):181–188.

Hay, A.M., 1988. The derivation of global estimates from a confusion matrix, *International Journal of Remote Sensing*, 9(8):1395–1398.

Heuvelink, G., and P. Burrough, 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification, *International Journal of Geographical Information Systems*, 7(3):231–246.

Hudson, W.D., and C.W. Ramm, 1987. Correct formulation of the kappa coefficient of agreement, *Photogrammetric Engineering & Remote Sensing*, 53(4):421–422.

Jupp, D.L., 1989. The stability of global estimates from confusion matrices, *International Journal of Remote Sensing*, 10(9):1563–1569.

Kok, K., A. Farrow, T. Veldkamp, and P.H. Verburg, 2001. A method and application of multi-scale validation in spatial land use models, *Agriculture, Ecosystems & Environment*, 85(1–3):223–238.

Lambin, E.F., X. Baulies, N. Bockstael, G. Fischer, T. Krug, R. Leemans, E.F. Moran, R.R. Rindfuss, Y. Sato, D. Skole, B.L. Turner II, and C. Vogel, 1999. *Land-Use and Land-Cover Change Implementation Strategy*, IGBP Report 48, IHDP Report 10, International Geosphere-Biosphere Program, The Royal Swedish Academy of Sciences, Stockholm, Sweden, 125 p.

Monserud, R., and R. Leemans, 1992. Comparing global vegetation maps with the kappa statistic, *Ecological Modelling*, (62):275–293.

Pontius, R.G., Jr., 2000. Quantification error versus location error in comparison of categorical maps, *Photogrammetric Engineering & Remote Sensing*, 66(8):1011–1016.

Pontius, R.G., Jr., J.D. Cornell, and C.A.S. Hall, 2001. Modeling the spatial pattern of land-use change with GEOMOD2: Application and validation for Costa Rica, *Agriculture, Ecosystems & Environment*, 85(1–3):191–203.

Prentice, I.C., W. Cramer, S.P. Harrison, R. Leemans, R.A. Monserud, and A.M. Solomon, 1992. A global biome model based on plant physiology and dominance, soil properties and climate, *Journal of Biogeography*, 19:117–134.

Ritters, K.H., R.V. O'Neill, C.T. Hunsaker, J.D. Wickham, D.H. Yankee, S.P. Timmins, K.B. Jones, and B.L. Jackson, 1995. A factor analysis of landscape pattern and structure metrics, *Landscape Ecology*, 10(1):23–39.

Rosenfield, G.H., 1986. A coefficient of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering & Remote Sensing*, 52(2):223–227.

Sader, S.A., and A.T. Joyce, 1988. Deforestation rates and trends in Costa Rica, 1940 to 1983, *Biotropica*, 20(1):11–19.

Turner, M.G., T. Costanza, and F.H. Sklar, 1989. Methods to evaluate the performance of spatial simulation models, *Ecological Modelling*, 48:1–18.

Wang, F., 1990. Improving remote sensing image analysis through fuzzy information representation, *Photogrammetric Engineering & Remote Sensing*, 56(8):1163–1169.

Woodcock, C.E., and S. Gopal, 2000. Fuzzy set theory and thematic maps: Accuracy assessment and area estimation, *International Journal of Geographical Information Science*, 14(2):153–172.