Comparing Raster Map Comparison Algorithms for Spatial Modeling and Analysis

Matthias Kuhnert, Alexey Voinov, and Ralf Seppelt

Abstract

The comparison of spatial patterns is recognized as an important task in landscape ecology especially when spatially explicit simulation modeling or remote sensing is applied. Yet, there is no agreed procedure for doing that, probably because different problems require different algorithms. We explored a variety of existing algorithms and modified some of them to compare grid-based maps with categorical attributes. A new algorithm based on the "expanding window" approach was developed and compared to other known algorithms. The goal was to offer simple and flexible procedures for comparing spatial patterns in grid based maps that do not take into consideration object shapes and sizes of the maps. The difference between maps was characterized by three values: quantity, location, and distance between corresponding categories in the maps. Combinations of these indices work as good criteria to quantify differences between maps. A web-based survey was set up, in which participants were asked to grade the similarity of ten pairs of maps. These results were then used to compare how well the various algorithms can perform relative to the visual comparisons obtained; they were also used to calibrate existing algorithms.

Introduction

Overview

Analysis of spatial information and data frequently requires comparisons of spatial patterns. Several data structures are available for processing spatial information (raster, vector, or hybrid approaches). The grid-based approach seems to be the most frequent choice (Seppelt and Voinov, 2003), especially when spatially explicit simulation models are involved, and we need to compare, model output to spatially referenced data. Algorithms are essential that can measure the difference between two maps for assessing the similarity between the output and a data map (Borenstein, 1998). However, there do not seem to exist any agreed universal procedures for doing that, especially when categorical maps are to be compared.

Matthias Kuhnert is with GeoForschungsZentrum Potsdam, Enginerring Hydrology, Telegrafenberg, 14473 Potsdam, Germany (kuhnert@gfz-potsdam.de), formerly with the TU Braunschweig, Institut for Geoecology, Department for Environmental System Analysis, Langer Kamp 19c, 38106 Braunschweig, Germany.

Alexey Voinov is with the University of Vermont, Gund Institute for Ecological Economics, 590 Main Street, Burlington, VT 05405-0088.

Ralf Seppelt is with the UFZ Centre for Environmental Research Leipzig, Department for Applied Landscape Ecology, Permoserstrabe 15, 04318 Leipzig, Germany.

For categorical information (such as land-use of soil types or vegetation classification), the simplest way to do such a comparison is to run a cell-by-cell match to get the total number of matching cells. With this comparison, small spatial differences are treated in the same way as the large ones: even if there is a matching cell right near by, it will not be recognized by the algorithm, and will not be taken into account, no matter how important it would be for a quality fit estimate of the model. A good algorithm should access errors for small drifts less than for farther shifts (Pontius, 2002). An example that demonstrates this problem is the comparison of two chessboard patterns that are shifted by one cell. The cell-by-cell algorithm will give zero agreement, since there is not a single cell from the first map that will match a cell from the other one. Yet, if one looks at these maps, there will be an obvious similarity that we would not want to ignore.

Visual comparison of maps is quite fast and efficient. Human perception works very well in choosing the most appropriate scale for the comparison. The whole picture is quickly recognized and differences and similarities identified, both when cells are substituted or moved, or if only some values are changed. The problem is that a visual comparison does not establish a quantitative ranking. Also, it is hard to depend on visual comparisons if we need to compare dozens or hundreds of maps, as in some optimization tasks that are often necessary for model calibration. Quantitative algorithms are essential, but it is also important that they retain some of the features of qualitative comparison that we find so useful in quick analysis or *eyeballing*.

Recent State of the Art

Most of the existing comparison algorithms like the Chi Square Test, Cramer's V, or Kendall's Tau (Everitt, 1977) are based on a variety of statistical procedures applied to the total numbers of cells in different categories. These are quite straightforward procedures, but unfortunately they take into account no information about the spatial pattern (Pontius, 2001; 2002). For example, if two groups of cells in one map trade places in the other map, the total count of correctly classified cells may still be the same, and cell-by-cell algorithms will not recognize any difference in the map comparisons.

One of the most well-known measures for map comparisons is the Kappa (Cohen, 1960). It is an index for comparison based on statistics calculated for the so-called error matrix (Congalton, 1993; Stehmann, 1996): a quantification of matches and mismatches in different cells. Pontius (2000) suggested several modifications to the Kappa algorithms that

Photogrammetric Engineering & Remote Sensing
Vol. 71, No. 8, August 2005, pp. 975–984.

0099-1112/05/7108–0975/$3.00/0

should distinguish between the error due to differences in the categories count (quantification error) and the error in spatial pattern misrepresentation (location error).

Equivalent to the Kappa statistics, Hagen (2003) created another indicator, K_{Fuzzy}, in which he used the fuzzy set theory to consider fuzziness of location and fuzziness of category for map comparison.

Another way to compare maps is based on the landscape pattern metrics (Riitters *et al.*, 1995). These are indices that describe different patterns for each category in a map. The indices for different maps are compared with each other to quantify the difference between two maps.

Turner and Costanza (1989) compared the spatial pattern analyses with the multiple resolution of "goodness of fit." The landscape pattern metrics is quite sensitive to the spatial pattern statistics, but it does not capture the location of the compared cells, therefore, potentially leading to confusing results, if we are more concerned with the location, rather than the pattern of change.

To account for the location of the mismatching cells, Costanza (1989) developed an algorithm that performs map comparisons over several resolutions. The algorithm scans the maps using a window, for which size is gradually increased. For each window size, a metric is calculated to compare the cells in the windows. These comparisons are then integrated across the whole map area and over multiple window sizes. Instead of using a moving window, Kok (2001) used two different resolutions of the project area to compare the simulation results with real data. This is equivalent to the moving window, but using only two fixed window sizes.

With the variety of these methods, it is not quite clear what are the benefits for particular applications, and how we should choose the right approach. In this paper, we compare a variety of existing algorithms to improve our understanding of how they perform and why they sometimes produce quite different results. We focus on comparing grid maps that contain different categories, with different category counts and arranged in various patterns. We also introduce yet another algorithm that we found computationally simpler and more flexible for modifications, than some of the other known algorithms. In attempt to generate a baseline comparison, we have performed a web-survey of nearly 100 participants who were asked to perform visual map fitting. The results of these visual comparisons were then compared to the results calculated by the quantitative methods.

Methods

Characteristic Values for Map Comparison

For a full characterization of a fit between two maps, one can distinguish between three different types of map mismatches characterized by:

1. The number of cells that changed from one category to another (quantity fit);
2. The number of cells that kept the category but changed location from one map to another (location fit);
3. The distance between the locations of these matching cells in the maps.

Pontius (2000) presented the first two characteristic values as agreement due to quantity and agreement due to location and in his later works (Pontius, 2002; Pontius *et al.*, 2004); he defined the distance characteristic. In the following, different indices, F_i, for comparing grid-based maps with categorical information are considered. In general, if the value of the function F is close to 1, the compared maps match well; if F is close to 0, the maps are totally different. Table 1 summarizes all notations in this paper.

TABLE 1. GENERAL NOTATIONS

a_{1i}, a_{2i}	numbers of cells with category i in Map 1, and Map 2
E_r	non-euclidean distance between corresponding cells in the compared maps
F_3	index for the moving window with a fixed window size of $w = 3$
F_{cbc}	index for the cell-by-cell comparison (fit for location)
F_d	index for the distance measure (different calculation in comparison to F_{dd})
F_{dd}	index for the distance measure (different calculation in comparison to F_d)
F_{ew}	index for the expanding window
F_{oall}	index for the overall comparison (fit for quantity)
F_t	index for an integrated value for the moving window
F_w	index for the moving window with window size w
i, j	the current category in Map 1, and Map 2
k	weighting factor for F_t
M_j	sum of the cells in category j in Map 2
N	the total number of cells in the map
N_i	sum of the cells in category i in Map 1
N_{id}	the number of direct matched cells in two maps
N_{ij}	number of cells that changed from category i in Map 1 to category j in Map 2
N_p	correct fit, if classification is perfect
n_s	number of cells in the calculated window s
q	total number of categories
Q_1	expected proportion correct due to chance
s	number of the window of one window size
t_w	number of windows with window size w
w	window size
W	weighting factor of the expanding window
x, y	coordinates for a cell in the map

Cell-by-cell Method

The most straightforward method of comparing two maps is to compute the cell-by-cell comparison that takes a cell in the first map (for example the model result) and matches it with the corresponding cell in the second map, (for example, the reference map with data). The count of a fit is 1, and a misfit counts as 0. The number of matched cells N_{id} divided by the total number of cells N gives a simple index:

$$F_{cbc} = N_{id}/N. \quad (1)$$

We may want to account for some more information about the types of mismatches observed and generate the so-called error matrix in Figure 1 (Congalton, 1993; Stehman, 1996). Here we assume categories of Map 1 in columns and

		Map2				Row
		1	2	⋯	q	Total
Map1	1	n_{11}	n_{12}	⋯	n_{1q}	$N_1 = \sum_{j=1}^{q} N_{1j}$
	2	n_{21}	n_{22}	⋯	n_{2q}	$N_2 = \sum_{j=1}^{q} N_{2j}$
	⋯	⋯	⋯	⋯	⋯	⋯
	q	n_{q1}	n_{q2}	⋯	n_{qq}	$N_q = \sum_{j=1}^{q} N_{qj}$
Column	Total	$M_1 = \sum_{i=1}^{q} N_{i1}$	$M_2 = \sum_{i=1}^{q} N_{i2}$	⋯	$M_q = \sum_{i=1}^{q} N_{iq}$	N

Figure 1. Error-Matrix for two compared maps: the categories of Map 1 in columns and categories of Map 2 in rows. Each element n_{ij} in the error matrix is the number of cells in category i in Map 1 and in category j in Map 2. N_i and M_j are the sums over the rows and columns, respectively.

categories of Map 2 in rows, then each element N_{ij} in the error matrix will present the number of cells that is in category i in Map 1 and in category j in Map 2; N_i and M_j are the sums of rows or columns, respectively.

Obviously the diagonal elements in this matrix will present the number of cells that are in the same category in both maps, and because of $N_{id} = \sum_{i=1}^{q} N_{ii}$ this approach is similar to Equation:

$$F_{cbc} = \frac{N_{id}}{N} = \frac{\sum_{i=1}^{q} N_{ii}}{N},$$

where q is the number of categories on both maps, and F_{cbc} is the index for cells with the same category placed at the same *location* in the two maps.

Overall Comparison

Another straightforward technique to compare maps is an overall comparison. This is the sum of the differences of the number of cells in each category. This is the disagreement due to differences in *quantity* of the categories (Pontius, 2002). The index gives the number of cells that do not change location but change the category:

$$F_{oall} = 1 - \frac{1}{N}\sum_{i=1}^{q}\left|a_{1i} - a_{2i}\right|, \quad (2)$$

where a_{1i}, a_{2i} are the numbers of cells with category i in Map 1 and Map 2, respectively.
The difference between F_{oall} and F_{cbc} estimates the number of cells that changed location.

Distance Index

The third characteristic value is the distance measure that tells us how far apart in space the disagreeing cells are. This index was not estimated previously; neither of the methods based on the error matrix can be used to calculate it. In the course of the paper, we will offer an algorithm that can calculate and take into account the *distance* E_r for each misplaced cell. There may be different ways to estimate the distance, for example, $E_r = \min(|x_r - x|, |y_r - y|)$. (x_r, y_r) are the coordinates of the r^{th} cell in Map 1, (x, y) are the coordinates of the cell with the same category in Map 2 that is the closest to the cell (x_r, y_r). In other words, E_r is the nearest distance (vertical, horizontal, or diagonal) to the cell in Map 1 where we find a matching category. An overall estimate for the distance measure could then be a metric based on:

$$F_d = \frac{1}{1 + \max_r(E_r)}, \quad (3)$$

or based on the average of all the distances:

$$F_{dd} = \frac{N}{N + \sum_r E_r}. \quad (4)$$

For a perfect fit $F_d = 1$, for the chessboard example $F_d = 0.5$ (as well as F_{dd}), since the matching cell is always found in the first next cell on the other map. Furthermore, we will use Equation 4 for the distance index.

Kappa

The error matrix (Figure 1) is used to define the well-known kappa index. This widely-used index has become almost standard in the remote sensing community (Cohen, 1960; Bishop *et al.*, 1975; Pontius, 1994). In calculating this index, we attempt to compensate for the "chance agreement" in the comparison and describe the index of agreement as the ratio of the observed accordance minus the probable accordance and the difference of maximum accordance minus probable accordance as proposed by Pontius (2000). F_k is defined by

$$F_k = \frac{F_{cbc} - Q_1}{N_c - Q_1}, \text{ with} \quad (5)$$

$$Q_1 = \sum_{i=1}^{q} N_i \cdot M_i / N^2, \text{ and} \quad (6)$$

$$N_c = 1. \quad (7)$$

In terms of the error matrix, the observed accordance is F_{cbc} (Equation 1). Manipulating the numbers in the error matrix, we can generate some other indices for map comparison. Also used in defining these indices is the so-called contingency table, which is similar to the error matrix, but with elements divided by the total number of cells N.

Pontius (2000) presents an equation for K_{no} that is similar to F_0 in Equation 8:

$$F_0 = \frac{F_{cbc} - 1/q}{N_c - 1/q}. \quad (8)$$

The value for N_c is 1. In Equations 5 and 6, we reward a fit that is achieved for more categories, and penalize a fit that has the same value of matches but for a lower number of categories available. This F_0 is not what is actually called the standard Kappa index, which uses some sort of a distribution for the probable accordance.

Pontius (2000) modified Kappa trying to distinguish between the quantity error and the location error. He noted that the error in spatial location (pattern) can be independent of the error in the quantity of category matching. The calculation of Kappa is based on the maximum value for the proportion correct, $N_c = 1$, if the classification is perfect (Equations 5, 6, and 7). However, he argued that in certain cases, we might be concerned with matching only the quantity of the fits, or only the location of the fits. In those cases instead of using $N_c = 1$ for the maximal fit in (2), we may use values smaller than that, specifically

$$N_{c2} = \sum_{i=1}^{q} \min(M_i, N_i), \quad (9)$$

which Pontius (2000) used to define the so-called $F_{location}$ index that is defined as the "success due to the simulation's ability to specify location divided by the maximum possible success due to a simulation's ability to specify location perfectly":

$$F_{location} = \frac{F_{cbc} - Q_1}{N_{c2} - Q_1}. \quad (10)$$

It should be noted however, that this jargon may be somewhat misleading, since just like with all other Kappa measures it is still based only on the error matrix, which means that we still do not have the information about where exactly in the maps the differences are located.

For example, on the chessboard comparison, $F_0 = F_K = F_{location} = -1$, indicating that there is nothing similar about the maps, whereas visually we may still find them quite identical. Or similarly, in the example considered below in Figure 2, the error map is the same for both maps, and therefore all the indices based on the error matrix will be also the same, while one comparison should probably be graded higher.

Multiple Resolution Methods

Moving Window Approach

The problem with Kappa and its modifications is that it is entirely based on the cell-by-cell statistics. Maps that have a bias or have similar patterns, but slightly distorted or misregistered, may not agree well (Verbyla and Hammond, 1995). Costanza (1989) created an algorithm that goes

(a) (b)

Figure 2. The numbers of black cells in (a) Map A and (b) Map B are the same. If we compare these maps to a map entirely in grey using the sliding window algorithm, the results are different.

beyond the cell-by-cell comparison to include the cell neighborhood being considered. The maps are scanned using a window of increasing size. With window size 1 × 1, we prepared the regular cell-by-cell comparison. Then, we take a window of 2 × 2 cells and again scan the whole map. For each location, we calculate an index that is based on the difference between the total numbers of cells in each category in the window. If the two windows are identical, this difference is zero. The more mismatches we find, the larger this index. The window is then increased and the maps are scanned again. As the window size grows, the granularity of the maps gets blurred, and eventually we get a perfect fit assuming that the numbers of cells in the same category is the same for the whole area. The index for the moving window is calculated as:

$$F_w = \frac{1}{t_w}\sum_{s=1}^{t_w}\left[1 - \frac{\sum_{i=1}^{p}\left|a_{1i} - a_{2i}\right|}{2w^2}\right] \quad (11)$$

where t_w is the number of windows of window size w that is necessary to cover the whole map, and a_{1i} and a_{2i} are the number of cells in Map 1 and Map 2, respectively, for category i in the searched window. The graph of $F_w(w)$ is a steadily increasing function that gives some idea about the three characteristic values for the comparison (quantity, location, distance). The initial value (for $w = 1$) is the F_{cbc} index (location is similar on both maps), and the maximum value (for the maximal w, that covered the whole map) is the F_{oall} index (quantity is similar in both maps). The pattern of the slope is a proxy of the distance. The place where the curve starts to slope (with respect to the w value) describes the distance between cells that changed location, and the steepness of the slope represents the number of cells that changed the location. So, if the graph curves up quickly with w, we would reason that the spatial shifts between the maps are quite minor, and there is a good fit. If the graph goes up only for large w, we can say that even though there might be a good overall comparison, the spatial match is quite poor.

For a comparison with the previously introduced indices, this function needs to be aggregated to a single value. Costanza (1989) proposes

$$F_t = \frac{\sum_{w=1}^{n} F_w e^{-k(w-1)}}{\sum_{w=1}^{n} e^{-k(w-1)}}, \quad (12)$$

which assigns an exponentially decreasing weight to the comparisons performed at lower resolutions; n is the total number of cells in the map, and k is a penalty coefficient. This algorithm makes use of a multi-scale comparison approach, which makes them distinct from all other map comparison indices introduced above. However, there are a couple of related problems.

Increasing the window size makes the number of "no-data" values a sensitive parameter in map comparison. If we consider "no-data" as another category, we may considerably distort the result, since there may be a large number of cells in this category that we will be matching. If "no-data" are ignored, we get windows along the boundaries of the study area that will have different numbers of active cells. To avoid this problem, it makes sense to replace Equation 11 by another formulae that ignores the "no-data" cells:

$$F_w = \frac{1}{t_w}\sum_{s=1}^{t_w}\left[1 - \frac{\sum_{i=1}^{p}\left|a_{1i} - a_{2i}\right|}{2n_s}\right]. \quad (13)$$

Here, n_s is the number of cells in the window with data. The boundary effects still show in the comparison results. Consider, for example, a comparison with the window size of 3 × 3. Suppose in one case, two corresponding windows will contain nine cells of different categories, whereas a different pair of windows at a different location, but with the same size of this comparison, will contain only one cell with a category value and eight no data values. In this latter case if this one cell is changed, the fit for the window changes by 100 percent. In contrast, in the former case, the change in one cell influences the fit for the window by only 11 percent. In the equation, the weighting of these windows is the same, but in reality they should be quite different.

In addition, the shape of the window causes unintended behavior for the distance measure. While scanning the map with the window, it would be best if we could look at each cell the same numbers of times. However, this does not happen, and the cells on the edge of the map are considered less often than the cells in the center (Figure 3). As a

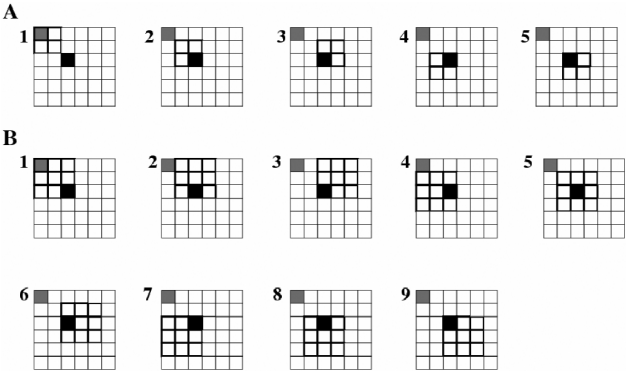

Figure 3. Cells in the middle and the cells at the edge get different consideration in the moving window algorithm. Independent of the window size w the cell in the corner gets noticed only once (A 1 and B 1). The cell in the middle will be considered once for window size $w = 1$, four times for window size $w = 2$ (A 2–5) and nine times for window size $w = 3$ (B 1–9), and so on. The number of hits will decrease for larger window sizes.

result a changed cell will be considered nine times with a 3×3 window, because it is a middle cell, and another not-changed cell will get considered only once, because it is a corner cell. This asymmetry may have a substantial effect on the result of the comparison.

Let us illustrate this problem with the following example. Suppose we have two grid maps and each of the maps is square (20×20 cells). The comparison is made for two categories (black and gray). One map has black cells on the edge (Figure 2a), and one map has black cells in the middle (Figure 2b). The number of black cells is equal in both maps. We compare each of these two maps with a map that contains only gray cells. It is interesting to watch how the F_w value changes as we increase the size of the window (Figure 4). Contrary to our expectations that we should be getting a higher value, the larger the window we use to scan the maps, we see that the index follows a different pattern. In some cases it first grows, then drops, in other cases, the opposite. Just as in the Figure 3 example, the algorithm is more influenced by the cells in the middle of the maps than on the periphery. If there is a similarity of cells in the middle and differences of the cells at the edge, the result will be overestimated, and vice versa (Figure 4). Still, the resulting index F_t seems to capture the difference pretty well (unlike the Kappa test, which is actually zero for both maps): for Map A, $F_t = 0.81$, for Map B, $F_t = 0.48$.

Expanding Window Approach

Based on the results of this analysis for the moving window comparison, we developed another algorithm that attempts to merge the simplicity of a cell-by-cell comparison with the multiple resolution comparison following the idea of the nearest neighborhood analysis (Burt and Barber, 1996).

The algorithm starts with the cell-by-cell comparison, and in case of a misfit, expands the search to a series of concentric layers around the cell in the second map. A match of two corresponding cells in both maps (same category at the same location) will add a one to the similarity count. A match found in one of the layers adds a $W\epsilon(0,1)$ to the similarity count, in which W is a weighting factor that is a constant for each comparison. If there are more than one match in the layers, the count will be incremented only once per layer and only for the nearest layer. The fit for the entire map is the sum of all weighted matches estimated for each cell:

$$F_{ew}(a,b) = \frac{1}{N}\sum_{r=1}^{N} W^{E_r} \qquad (14)$$

where E_r is one of the characteristic values, the distance that we introduced previously. The smaller the value of W is applied, the higher the penalty for finding a distant match (Figure 5). It should be noted that this algorithm is not symmetric. *F(a,b)* does not equal *F(b,a)*, where *a* and *b* are the two maps. For example, comparing Map 1 to Map 2 in Figure 6 the sum of the weighting points is equal to 8 ($F_{ew} = 0.89$). The other way round, starting with Map 2, we get 8.5 weighting points ($F_{ew} = 0.94$).

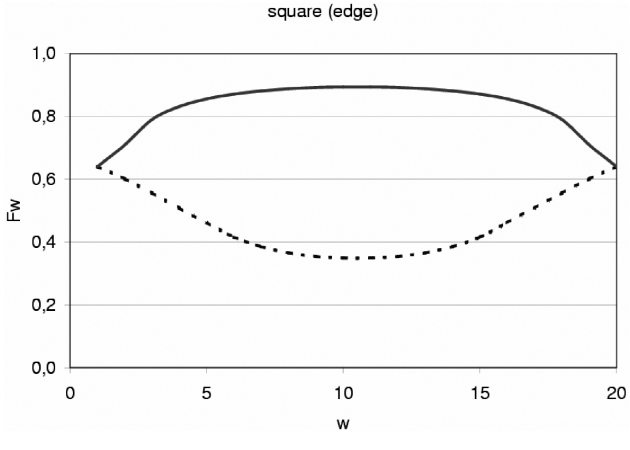

Figure 4. The result of the comparison of the maps in Figure 2. The solid line shows the comparison of Map A with a grey map (first comparison), and the broken line is the comparison of Map B with a grey map (second comparison). For intermediate window sizes, the cells in the middle matter more than the cells at the edge. In the first comparison, the cells in the middle are not changed and the fit is improved. The index for the second comparison shows a drop and then an increase, because the mismatching cells are in the middle.

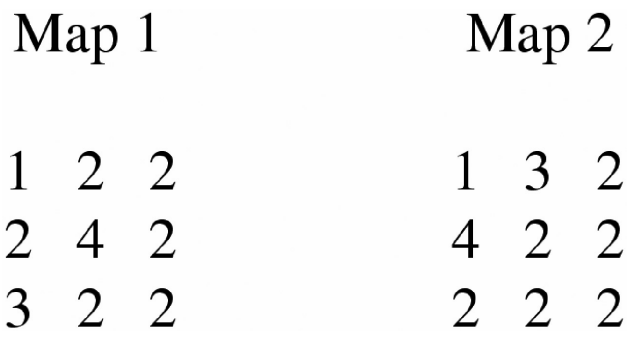

Figure 5. Example for a comparison using the expanding window algorithm: for cell (1,1) with category 1: the maps are identical -> add $W^0 = 1$ to the index; for cell (2,1) with categories 2 and 4: the match is in the first layer -> add $W^1 = W$ (≤ 1) to the index; for cell (3,1) in the Map 1 -to- Map 2 comparison the match is in the second layer add W^2 ($\leq W$) to the index, and in the Map 2 -to- Map 1 comparison, the match is in the first layer -> add W^1 ($\leq W$) to the index.

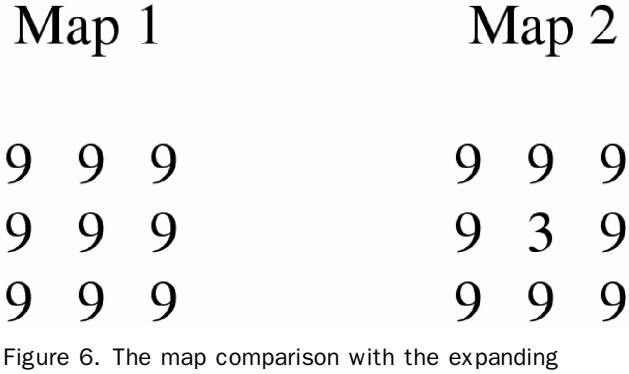

Figure 6. The map comparison with the expanding window algorithm is asymmetric. Comparing Map 1 to Map 2 may produce a different result from comparing Map 2 to Map 1.

To achieve symmetry we do both comparisons, and then take the average:

$$F_{ew} = (F_{ew}(a,b) + F_{ew}(b,a))/2.$$

This algorithm does not assume any special window pattern and does not depend upon the shape of the analyzed area. In this case, we do not have to go beyond the study area, we are only searching the data values for a matching cell.

The results of comparison depend on an arbitrarily chosen weight W. For example, in the chessboard comparison for a factor of $W = 0.5$, the result of the comparison is 0.5. If $W = 1$ the result is 1, which is probably not exactly right, since there is definitely a difference between the two patterns. We suggest that W is used as a calibration parameter that might be different for different applications.

Applying this method to the example in Figure 2, we find that F_{ew}(Map A) = 0.71, and F_{ew}(Map B) = 0.69. Though again the A comparison gets a higher grade than the B one; the difference between them is no longer as dramatic as in the moving window where the edge effect exaggerated the difference. This result can be easily interpreted if we remember that the F_{ew} index is a measure of the distance between similar cells in the two maps: indeed, in Map A there is always a shorter distance to the next gray cell (<3) than in Map B (the distance is greater for the cells in the middle of the black zone in the center): the algorithm detects this. Clearly in real life applications, it is unlikely that we will be comparing such special map designs, however these simple test examples proved to be very useful to understand how different methods work, and what to expect from them in certain extreme situations.

Coupled Indices

To couple the three characteristic values, we can derive an index based on a weighted sum of the quantity, location, and distance measures. We use F_{oall}, F_{cbc}, and F_d as characteristic values of a comparison, and a full characterization of a fit could be then built as a combination of these three indicators:

$$F_{full} = (\lambda_1 F_{oall} + \lambda_2 F_{cbc} + \lambda_3 F_d)/3. \quad (15)$$

Here λ_i are weight factors, $0 \leq \lambda_i \leq 1$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$ that can be used to assign a higher value to the quantity, location, or distance indicator. With the F_{full} index there are two λ_i parameters that need be defined in a calibration procedure. We have tried to perform this calibration using a visual comparison of a set of maps as a baseline. To achieve this baseline, we have set up an Internet survey, where a set of several pairs of maps was offered for comparison. The above-mentioned algorithms were then applied to calculate the map comparison indices over the same set of maps. All the algorithms were programmed as User Code functions within the open source Spatial Modeling Environment (SME) package (Maxwell and Costanza, 1997; Voinov *et al.*, 1999).

Internet Survey

Experimental Set-up

For the survey we used maps that contain five different land-use categories presented by five different colors. The first survey[1] contained ten pairs of maps that are shown in Figure 7. There were five pairs of maps generated by simulation runs (Figure 7: Numbers 1, 3, 7, 9, and 10) for the Hunting Creek watershed (Seppelt and Voinov, 2003). They were different only in the quantity of the categories. Maps in pairs 2, 4, 5, 6, and 8 were put together by hand, making them different in the location of cells, to trigger all three characteristic values. The maps in the pairs with the numbers 4, 5, and 8 (Figure 7) were only different in location, and maps in two pairs number 2 and 6 (Figure 7)

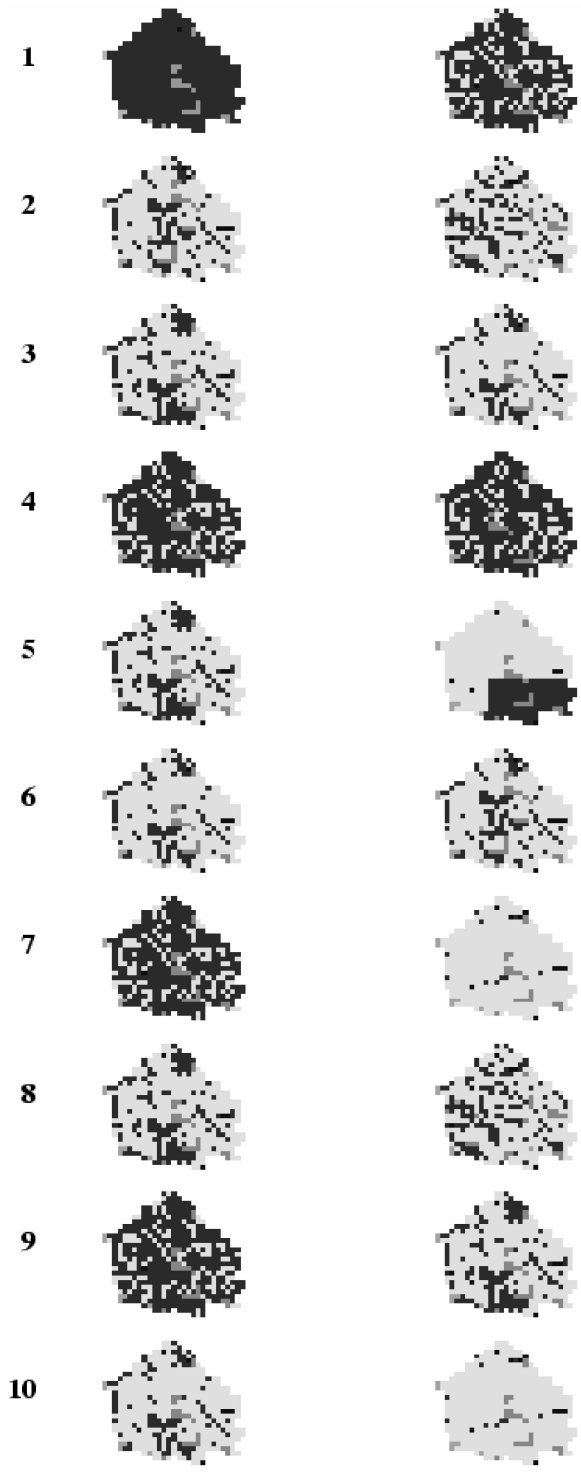

Figure 7. Pairs of maps offered for comparison in the Internet survey.

[1]The survey can be accessed at http://www.likbez.com/AV/Maps/ (last date accessed: 20 April 2005).

were different in quantity and location. The second survey contained ten comparisons, five of them were similar to the ones in the first survey (1, 3, 7, 9, and 10 maps in Figure 7). The second survey contained only comparisons that are different in the number of cells in the same categories, but not in location. The experimental setup was the same as in the first survey.

Participants in the test were asked to rank each comparison with a value between 0 and 1 with increments of 0.1.

Results and Discussion

In total, 186 forms were submitted from students and scientists (two independent groups of about the same size for the two surveys). This provided us a good sampling of how human perception works for the comparison. We used the results of the survey to compare them with the results of the different algorithms of map comparison described above.

Figure 8 presents the distribution of scores that were obtained for each of the compared pairs of maps. The graphs do not offer a clear Gaussian distribution; most of the distributions have more than one local maximum. However, the global maximum is quite well-pronounced showing that in general, there is some agreement between the respondents. To make sure that this is not a random effect, we have conducted a second independent survey with a different group of respondents. Figure 9 shows the same trends in the curves, and the maxima are in similar places. Very likely, we are detecting some real phenomena related to the human perception of the maps. Perhaps some respondents may be attaching a larger weight to background comparisons, while others may value different colors differently, or think that certain patterns or clusters are more important. All these aspects may influence the results of the test. This is clearly an interesting topic for further analysis, however it is beyond the scope of this paper. For now, it was important to conclude that there is a clear preference in people's judgment that we can use for further reference. We used the result of the first survey to determine if some agreement is possible between the algorithmic methods and the visual comparisons presented by the higher peak in the distributions.

We used the following indices for the comparison study. The moving window was represented by the integrated value F_t (with $k = 0.1$), and for the expanding window, we used a weighting factor of $W = 0.5$. In addition to these two algorithms, we correlated F_{cbc}, F_{oall}, F_K, $F_{Location}$, and F_3 with the survey. F_3 stands for F_w with $w = 3$. The window size of 3×3 was chosen to test the impact of small changes in location upon the overall comparison. The results show that the F_3 index seems to match very well with the visual comparisons obtained in the survey, which leads to a hypothesis that the window of size 3×3 is what people normally employ in making the judgment about maps similarity.

The scatter plots in Figure 10 show the relationship between the survey and the different algorithms. The relationship between the Kappa values and the survey seems to be quite random, while the distribution for the other indices

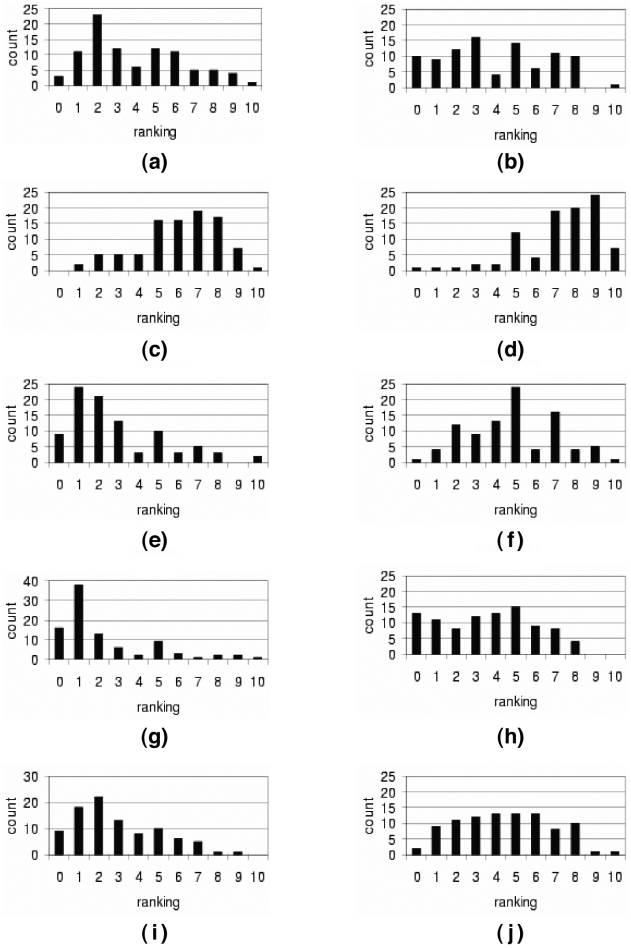

Figure 8. Distribution of the grading for each of the ten pairs of maps that were compared during the survey: (a) Pair 1, (b) Pair 2, (c) Pair 3, (d) Pair 4, (e) Pair 5, (f) Pair 6, (g) Pair 7, (h) Pair 8, (i) Pair 9, and (j) Pair 10.

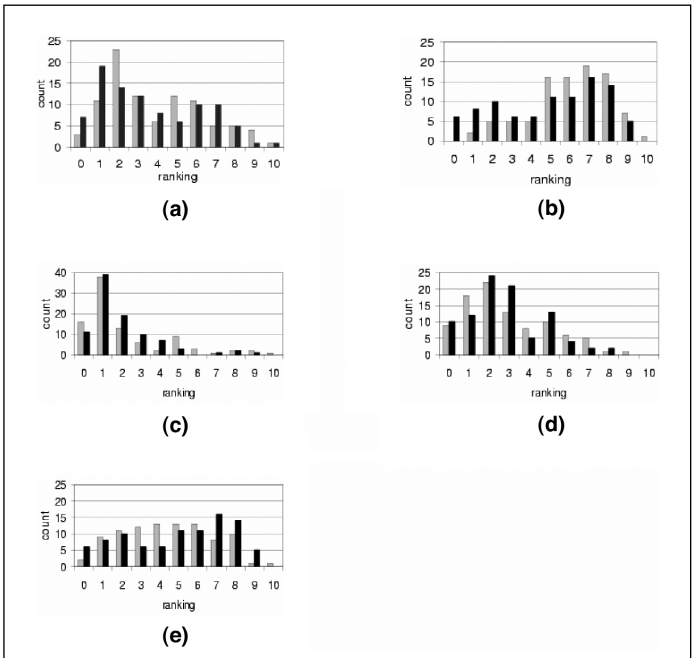

Figure 9. The comparison of the results in two independent surveys (grey the first, and black the second survey). The numbers of the comparisons correspond to the numbers in Figure 7 and Figure 8. (a) Pair 1, (b) Pair 3, (c) Pair 7, (d) Pair 9, and (e) Pair 10.

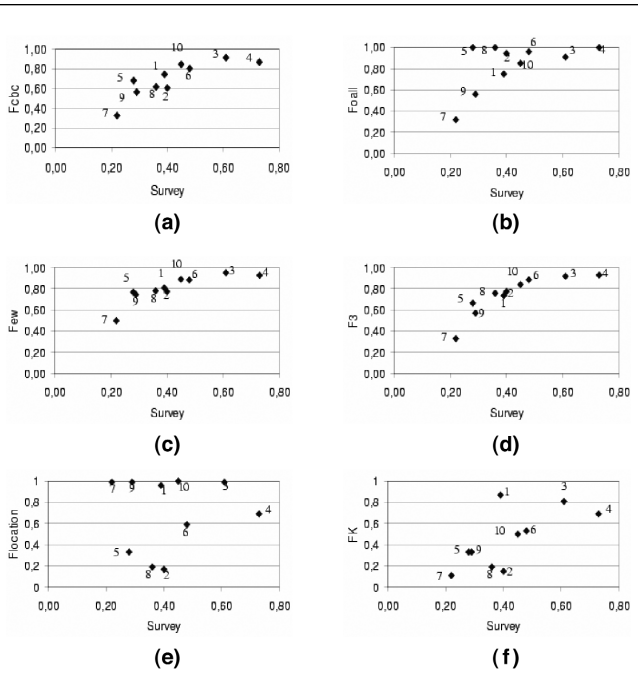

Figure 10. The scatterplots show the different relationships between the algorithms and the survey: (a) and (b) are the cell-by-cell and the overall comparison, (c) and (d) the moving window value (F_t) and the moving window with fixed window size 3 × 3 (F_3), and (e) and (f) are the Kappa for location and Kappa standard. The number of the points constitutes the corresponding pair in Figure 7.

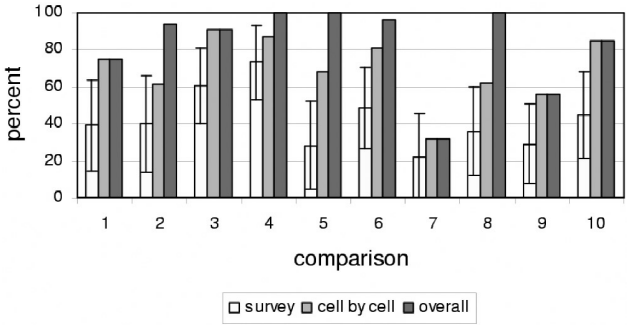

Figure 11. The survey results (n = 93) and the values for the quantity (F_{cbc}) and location (F_{oall}) indices. The results of the survey are consistently lower than the indices for changes in quantity and location. The standard deviations in all comparisons are 20 to 26 percent.

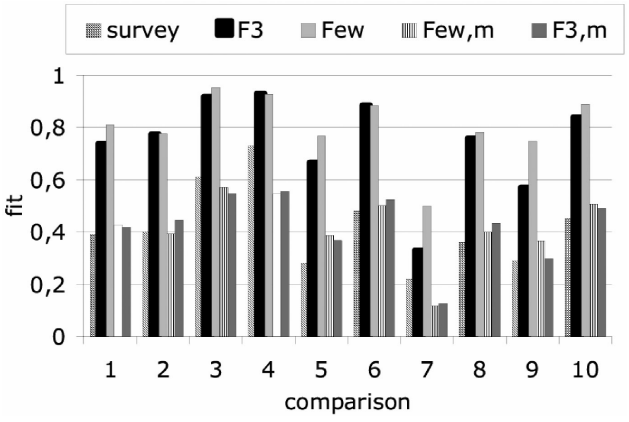

Figure 12. Comparison of the expanding window (F_{ew}) and moving window (F_3) algorithms with the survey results. $F_{ew,m}$ and $F_{3,m}$ are similar to F_t and F_3, but shifted down for a better match with the survey.

are closer to a straight line, meaning that there are better correlations between the comparisons.

The results of all map comparisons are presented in Table 2. The average values from the survey turned out to be lower than the values from all methods (Table 2 and Figure 11). Figure 12 shows the comparison of the expanding window and the moving window with a fixed window size w = 3. These two techniques have beside kappa the best correlation with the survey (Table 3). The F_3 index does not distinguish between a direct match and a match over a distance of 3 cells.

The results of the survey were used to calibrate the algorithms, assuming that the survey is the correct comparison. For example, for the expanding window the weighting factor W was calibrated. We ran the algorithm with ten different weighting factors in steps of 0.1 from 0.1 to 1.0. Table 4 presents the correlations with the survey that were produced; based on these results, we may assume that the best weighting factor is W = 0.3.

TABLE 2. INDICES OF MAP COMPARISONS CALCULATED BY DIFFERENT METHODS

Comparison	1	2	3	4	5	6	7	8	9	10
F_{cbc} (cell by cell)	0.75	0.61	0.91	0.87	0.68	0.81	0.32	0.62	0.56	0.85
F_{oall} (overall)	0.75	0.94	0.91	1.00	1.00	0.96	0.32	1.00	0.56	0.85
F_t	0.75	0.91	0.91	0.99	0.92	0.94	0.32	0.95	0.56	0.85
F_3	0.74	0.77	0.92	0.93	0.67	0.89	0.33	0.76	0.57	0.84
$F_{ew}(W = 0.5)$	0.81	0.78	0.95	0.93	0.77	0.88	0.50	0.78	0.75	0.89
F_K	0.27	0.15	0.81	0.69	0.33	0.53	0.11	0.19	0.33	0.5
$F_{location}$	0.96	0.17	0.99	0.69	0.33	0.59	0.99	0.19	0.99	1
Survey	0.39	0.40	0.61	0.73	0.28	0.48	0.22	0.36	0.29	0.45

Note: The cell-by-cell-comparison compares the two maps cell by cell, the overall gives the sum of the difference between the numbers of cells in each category in the two maps. F_t is the integrated result for the moving window and F_3 is the value for the moving window, using only window size of three. $F_{ew}(W = 0.5)$ is the value for the expanding window calculated with the weighting factor W = 0.5. F_K and $F_{location}$ are Kappa (Cohen, 1960) and Kappa of location (Pontius, 2000), respectively.

TABLE 3. THE CORRELATION OF EACH METHOD WITH THE SURVEY RESULTS

	F_{cbc}	F_{oall}	F_t	F_{ew}	F_3	F_{int}	F_K	$F_{location}$
Correlation	0.822	0.565	0.820	0.826	0.845	0.577	0.844	0.090

Note: F_t is the integrated value for the moving window with k = 0.1, for the expanding window we used the weighting factor of 0.5. F_3 is the result of the moving window using only window size w = 3. F_{int} is the integral index of the expanding window divided by the result for a perfect match.

TABLE 4. CORRELATION BETWEEN THE SURVEY AND THE RESULTS OF THE COMPARISONS WITH THE EXPANDING WINDOW WITH DIFFERENT WEIGHTING FACTORS

W	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Correlation	0.825	0.828	0.829	0.828	0.824	0.814	0.792	0.748	0.658	0.463

Finally, we used the survey results to calibrate the λ_I weight coefficients for the characteristic values by minimizing the error between the survey results and the coupled index F_{full}:

$$\hat{F} = \sum_{1,\ldots,10} \left(\lambda_1 F_{cbc} + \lambda_2 F_{oall} + \lambda_3 F_d - F_{survey}\right)^2 \longrightarrow \min!$$

with the constraints $0 \leq \lambda_i \leq 1$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$. The result of a global optimization was $\lambda_1 = 0.85$, $\lambda_2 = 0$ and $\lambda_3 = 0.15$. The correlation between the linear combination of the combined distance measures $\hat{F}$ and the survey results was $R^2 = 0.66$. Apparently with the visual comparison, what matters most is the estimate of total number of cells in the same category. We do not know how people interpret the count by simply looking at the map, but it appears that humans are pretty good at that judgement. The loecation of change *per se* does not seem to matter at all; however, the distance of the shift between cells that moved to a different location does matter. As one would expect, it does not matter where on the map the cells got shifted. As long as they did not get too far apart, the fit stands. The fact that the F_3 index works so well also speaks in favor of this assumption.

Conclusions

The differences between two categorical grid maps can be characterized by three values: the quantity of the categories in the two maps, the location of the mismatched categories, and the distance between two corresponding cells that changed location. There are a number of analytical methods available for map comparisons, however, there is little understanding about how to compare these methods among themselves and what are the advantages and drawbacks of each method.

A web-based survey was an attempt to create some kind of a reference index. The human eyes seem to be quite good in finding similarities and differences, but humans often fail in assigning a number to their comparison results. Assuming that the survey value is the "correct" value for the comparison, we have derived a criterion to compare the algorithms.

The traditional Kappa techniques worked very well when pattern and location of change was not involved (Figure 7: Numbers 1, 3, 7, 9, and 10), because it is based on the error matrix, and these data are detached from the correct locations of the grid cells. Only the calculation of direct matches (F_{cbc}) and the number of mismatches are taken into account, but not the distance between corresponding cells that changed location. This is a disadvantage compared to the other discussed algorithms. However, the values in the moving and the expanding windows showed about the same correlation with the survey that the Kappa test (Table 2). The moving window algorithm worked very well except for some special cases.

The expanding window algorithm circumvents the problem of the edge effect. The algorithm can search for similarities independent of the area or pattern of the map. The flexibility of the algorithm makes it a useful tool in map comparison. The variable weighting factor allows the user to adjust it for specific questions of map comparison, but for the presented comparisons it did not produce the correlation significantly better than for of Kappa. Apparently this is because the distances between cells in different location in the survey were quite small. The longest distance is chosen for the fifth comparison (14 cells) and only the eighth comparison has another distance longer than 2 cells (6 cells). The survey turned out to be poorly designed to account for the location changes. Nevertheless, is the survey a good hint of the functionality of the algorithms.

The best correlation with the survey came from the moving window with fixed window size of $w = 3$, which makes us think that human perception penalizes small changes less than changes over long distances.

The Internet survey will stay on-line at **http://www.likbez.com/AV/Maps**, and hopefully generate more comparison results from our readers. This may help us identify more important factors in the comparisons and improve our quantitative methods in the future.

Acknowledgment

We gratefully thank two anonymous referees for their valuable remarks and suggestions.

References

Borenstein, D., 1998. Towards a practical method to validate decision support systems, *Decision Support Systems*, 23:227–239.

Bishop, Y., S. Fienberg, and P. Holland, 1975. *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Massachusetts.

Burt, J.E., and G.M. Barber, 1996. *Elementary Statistics for Geographers*, 2nd Edition, The Guilford Press, London, New York.

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

Congalton, R., 1993. A practical look at the source of confusion in error matrix generation, *Photogrammetric Engineering & Remote Sensing*, 59:641–644.

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

Everitt, B.S., 1977. *The analysis of the contingency tables*, John Wiley & Sons, Inc., New York.

Hagen, A., 2003. Fuzzy set approach to assessing similarity of categorical maps, *International Journal for Geographical Information Science*, 17(3):235–249.

Kok, K., 2001. A method and application of multi-scale validation in spatial landuse models, *Agriculture, Ecosystems and Environment*, 85:223–238.

Maxwell, T., and R. Costanza, 1997. A language for modular spatio-temporal simulation, *Ecological Modeling*, 103:105–113.

Pontius, R.G., 1994. *Modeling Tropical Land Use Change and Assessing Policies to Reduce Carbon Dioxide Release from Africa*, Graduate Program in Environmental Science, SUNY-ESF, Syracuse, New York, 177 p.

Pontius, R.G., 2000. Quantification Error Versus Location Error in Comparison of Categorical Maps, *Photogrammetric Engineering & Remote Sensing*, 66(8):1011–1016.

Pontius, R.G., 2001. Land-cover change model validation by ROC method for the Ipswich watershed, Massachusetts, USA, *Agriculture, Ecosystems and Environment*, 85:239–248.

Pontius, R. G., 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at

multiple resolutions, *Photogrammetric Engineering & Remote Sensing*, 68(10):1041–1049.

Pontius, R.G., E. Shusas, and M. McEachern, 2004. Detecting important categorical land changes while accounting for persistence, *Agricultural, Ecosystems and Environment*, 101(2–3):251–268.

Riitters, K.H., R.V. O'Neill, C.T. Hunsacker, 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.

Seppelt, R., and A. Voinov, 2003. Optimization methodology for landuse patterns - Evaluation based on multiscale habitat pattern comparison, *Ecological Modelling*, 168(3):217–231.

Stehman, S.V., 1996. Estimation the kappa coefficient and its variance under stratified random sampling, *Photogrammetric Engineering & Remote Sensing*, 62(4):401–407.

Turner, M.G., and R. Costanza, 1989. Methods to evaluate the performance of spatial simulation model, *Ecological Modelling*, 48:1–18.

Verbyla, D.L., and T.O. Hammond, 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids, *International Journal of Remote Sensing*, 16(3):581–587.

Voinov A., R. Costanza, L. Wainger, R. Boumans, F. Villa, T. Maxwell, and H. Voinov, 1999. Patuxent landscape model: integrated ecological economic modeling of a watershed, *Environmental Modeling and Software*, 14(5):473–491.

(Received 11 December 2003; accepted 20 February 2004; revised 04 May 2004)