J Geograph Syst (2006)
DOI 10.1007/s10109-006-0025-x

ORIGINAL ARTICLE

William W. Hargrove · Forrest M. Hoffman · Paul F. Hessburg

Mapcurves: a quantitative method for comparing categorical maps

Abstract We present Mapcurves, a quantitative goodness-of-fit (GOF) method that unambiguously shows the degree of spatial concordance between two or more categorical maps. Mapcurves graphically and quantitatively evaluate the degree of fit among any number of maps and quantify a GOF for each polygon, as well as the entire map. The Mapcurve method indicates a perfect fit even if all polygons in one map are comprised of unique sets of the polygons in another map, if the coincidence among map categories is absolute. It is not necessary to interpret (or even know) legend descriptors for the categories in the maps to be compared, since the degree of fit in the spatial overlay alone forms the basis for the comparison. This feature makes Mapcurves ideal for comparing maps derived from remotely sensed images. A translation table is provided for the categories in each map as an output. Since the comparison is category-based rather than cell-based, the GOF is resolution-independent. Mapcurves can be applied either to entire map categories or to individual raster patches or vector polygons. Mapcurves also have applications for quantifying the spatial uncertainty of particular map features.

W. W. Hargrove (✉)
Environmental Science Division, Oak Ridge National Laboratory,
P.O. Box 2008, M.S. 6407, Oak Ridge, TN 37831-6407, USA
E-mail: hnw@fire.esd.ornl.gov
Tel.: +1-800-2412748

F. M. Hoffman
Computer Science and Mathematics Division,
Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
E-mail: forrest@climate.ornl.gov

P. F. Hessburg
USDA Forest Service, PNW Research Station, Wenatchee,
WA 98801, USA
E-mail: phessburg@fs.fed.us

Keywords Ecoregion · Goodness-of-fit · Kappa statistic · Landcover · Model validation · Overlap · Spatial concordance · Spatial uncertainty · Vegetation

1 Introduction

The quantitative comparison of two categorical maps seems deceptively simple conceptually, yet proves complex in practice. The identification of categorical differences between maps is the basis for much vegetation cover and change detection research, as well as evaluation of the output from spatially explicit models. While statistical methods exist for comparing categorical maps on a cell-by-cell basis, there are no formal methods for comparing two or more categorical maps based on the categories themselves. This is surprising, given the frequency of practical questions of the form, "Did the amount of deciduous forest change during this time interval?" The importance of categorical map comparison is of growing interest to researchers (Metternicht 1999; Winter 2000; Pontius 2000; Pontius and Schneider 2001; Power et al. 2001).

Conventional categorical comparisons perform a cell-by-cell overlay of the two categorical maps to indicate goodness-of-fit (GOF) and regions of agreement and disagreement. The contingency table, or confusion matrix, in which the columns of the table are categories of one map and the rows are categories of the other, forms the basis for many current categorical GOF statistics. The last row and column give column and row totals. The basis for all GOF statistics generated from a contingency table is cell-by-cell agreement between the two maps.

Chi-square, phi, tau, and kappa statistics are all based upon the contingency table approach (Pontius 2002). The kappa statistic is often used as an overall measure of accuracy which provides a statistical measure of the degree to which cell classification agrees, and has the added advantage of accommodating the effects of chance agreement (Monserud and Leemans 1992). However, kappa does not provide a spatial distribution of the errors (Foody 2002). The functionality and limitations of the kappa statistic have been extensively discussed from its application within several disciplines (Maxwell 1977; Maclure and Willet 1987; Lantz and Nebenzahl 1996).

Contingency table methods fail to distinguish between a near miss and a far miss, and are not designed to account for partial success. For example, two checkerboard maps that are out of phase with each other by one cell width will show total disagreement. Fuzzy sets (Zadeh 1965) and fuzzy classification methods (Hagen 2003; Metternicht 1999; Power et al. 2001) were created as an attempt to account for partial concordance in categorical classifications, but have not been widely adopted. Because they are based on cell-by-cell comparisons, changing the resolution of categorical maps can have a dramatic effect on contingency table-based statistics. Several GOF methods have been derived which depict changes in concordance across multiple resolutions by generating windows at different resolutions, then

plotting agreement as a function of window size (Costanza 1989; Turner et al. 1989; Plotnick et al. 1996).

Because the marginal totals in contingency tables are fixed, accuracy due to quantity is confounded with accuracy due to location (Pontius 2000). Pontius (2000) separates agreement due to quantity versus agreement due to location for categorical maps at a single resolution. In a seminal paper, Pontius (2002) extends these methods for use with both exclusive and fuzzy classification, and at multiple resolutions, but still requires identical number and type of categories in each map being compared.

Most contingency table-based statistics expect the number of rows to equal the number of columns, and to appear in corresponding order, so that the diagonal represents correctly classified cells, and elements above or below the diagonal are incorrectly classified. Many remote sensing analysis methods invalidate an assumption that classes in one map represent the exact same features as corresponding classes in another map. Even if the maps to be compared have the same number of categories, a particular category in one map may not precisely equate to the same category in another, and the most appropriate mapping of the categories in one map to those in the other may be unclear. Such circumstances arise, for example, whenever multiple remotely sensed images are independently subjected to unsupervised classification methods.

Because of the requirement for equal, corresponding categories, current quantitative comparison approaches often start by subjectively developing a translation table which combines the categories in the finely split map to best match the categories in the coarsely split map. This equivalency is made on the basis of a priori assumptions and subjective interpretation of the category legend descriptions, before any examination of the maps themselves. Then the more finely-divided map is translated into the coarser categories by lumping, and the maps are quantitatively compared at the coarsest level of division.

Consider, for example, two alternative ecoregionalizations or vegetation cover maps produced by experts who have different approaches. A ''splitter'' may have simply subdivided essentially the same ecoregions produced by a ''lumper.'' The ''splitter'' might have map categories called ''spruce,'' ''fir'' and ''hemlock,'' whereas the lumper might have a single category called ''evergreen needleleaf forest.'' A typical quantitative comparison approach might start by examining the category legends in each map, before ever examining the maps themselves. A translation- or lookup-table which lumps the categories in the finely split map into the categories in the coarsely split map is developed on the basis of a subjective interpretation of the category legend. Because it is impossible to split the coarser map into the categories of the finer one, the finely split map is translated into the categories of the coarser map, and the maps are quantitatively compared at the coarsest level of division. Such comparisons provide only a quantitative veneer for the subjectively developed translation table. The a priori translation rules are based solely on interpretation of the category legend descriptors. The translation table should be a product resulting from the map comparison process, not the basis for it.

Many of the problems associated with current comparison methods for categorical maps stem from the fact that they are cell-based rather than based on the features of interest. An ideal categorical map comparison method should be independent of both changes in resolution and differences in number of categories between the maps being compared. It should not require that the same number of categories be present in both maps, nor should it make assumptions about the equivalency of those categories. Needed is a method that is based on the objects of interest, e.g., categories, polygons, or patches, rather than individual map cells, which are simply the objects of their depiction.

We present Mapcurves, a quantitative GOF method that unambiguously shows the degree of concordance between two or more categorical maps. It is not necessary to interpret (or even know) legend descriptors for the categories in the maps to be compared, since the degree of fit in the spatial overlay forms the basis for the comparison. Since the comparison is category-based rather than cell-based, the GOF is resolution-independent. The Mapcurves algorithm can be applied to entire map categories, or to individual patches or polygons. The Mapcurves technique can also be applied to quantify the spatial uncertainty of particular map features.

2 Methods

2.1 Directionality of map comparisons

We postulate that map comparisons are unidirectional and intransitive; that is, Map 1 may fit better when compared with Map 2 as a reference than Map 2 fits using Map 1 as a reference. The selection of a base or reference map to which another is compared determines the direction of the comparison. If the number of categories in the maps being compared differs widely, the coarser map usually will exhibit a better comparison with the finer one as a reference than vice versa. Coarseness depends on the average size and number of the patches in each category, and may or may not be reflected in the number of categories in the map. The comparison direction that produces the best degree of fit is the one that we intuitively consider to be the level of concordance between the maps.

Our conceptual model for comparison of categories is based on the degree of spatial overlap (Fig. 1). Two categories from two different maps are judged to be a good fit if their degree of spatial overlap is nearly complete (Fig. 1, right side). At this extreme, the two categories have a large degree of positive spatial correlation. This establishes a strong identity between these two categories, which can then be said to express the same feature in the two maps. Similarly, two categories are judged to be a poor fit if they share very little area of spatial overlap (Fig. 1, left side). Little spatial concordance means that these categories are not identical, and therefore describe separate features. An ideal GOF model will be especially responsive to incremental increases at high overlap, since this extra sensitivity will discriminate excellent fit from good fit, while distinguishing both from poor fits.

2.2 The GOF algorithm

The GOF algorithm can be applied equally well to whole map categories, individual patches, or even to vector polygons (although only application to entire map categories is described here). The comparison is restricted to the extent that both maps overlap spatially, and begins by selecting a category from the map which is being compared (Map 1, Fig. 2). All categories from the reference map (Map 2) having any degree of spatial overlap with this category are identified. The map comparison GOF algorithm is based on two values: (1) the proportion of the intersecting area to the total area of the intersecting category from Map 2, and (2) the proportion of the intersecting area to the total area of the category from Map 1. The first term gives the proportion of "insideness" that the reference category shares with the tested category, and itself represents a GOF term. The second term weights this degree of fit by the fractional share of the Map 1 category's area that is intersected. Without such area weighting, the presence of many large, intersecting categories, each of which might share only a small spatial intersection with the category being tested, would result in a high degree of fit.

Summation of the product of "insideness" and the area weighting term over all intersecting categories provides a GOF score for this Map 1 category. Units are area squared over area squared, so that this GOF measure is

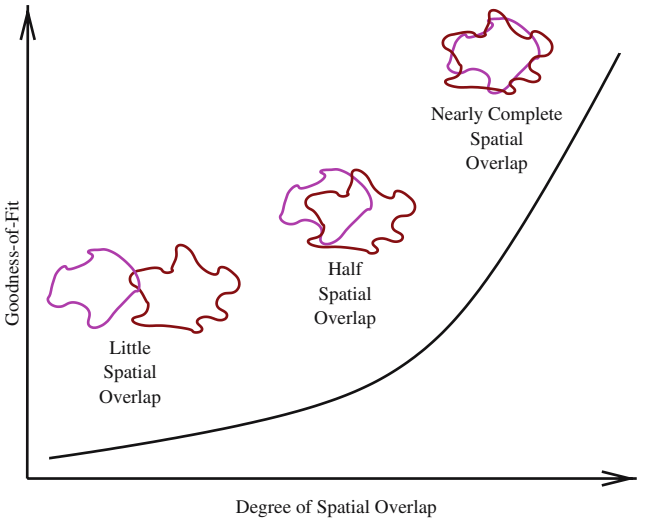

Fig. 1 Conceptual basis for comparing two categorical maps. The conceptual degree of fit is shown for two categories from separate maps as their spatial overlap is increased. When spatial overlap is maximized, the goodness of fit is high, and an identity between the map categories is suggested. When there is little spatial overlap, goodness of fit is low, and identity is unlikely. An ideal GOF model will be especially responsive to incremental increases at high overlap, since this extra sensitivity will discriminate excellent fit from good fit, while distinguishing both from poor fits

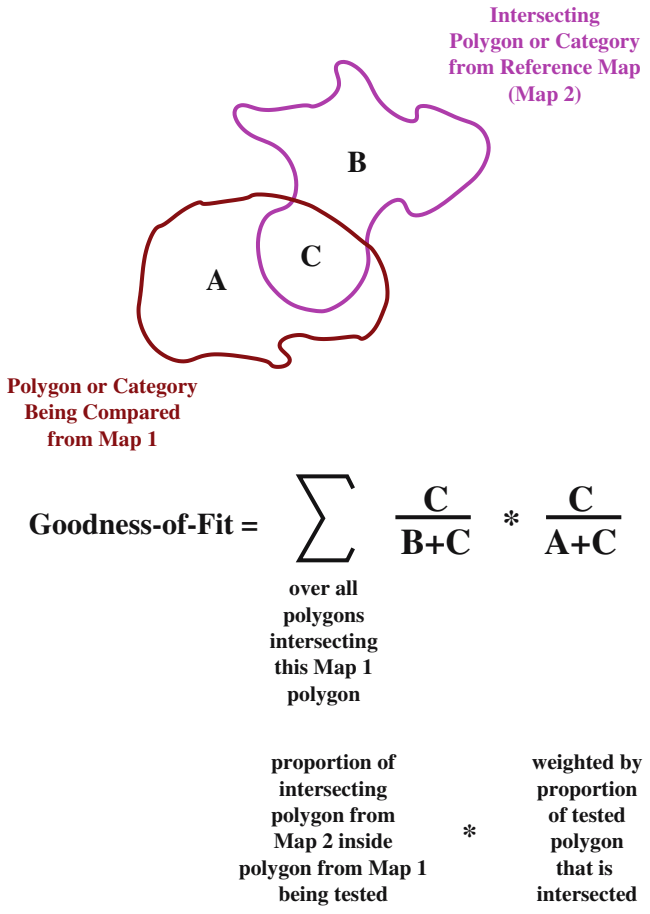

Fig. 2 Goodness-of-fit (*GOF*) algorithm used with Mapcurves. A polygon or category is isolated from the map being compared (Map 1, *brown*), and all intersecting polygons or categories from the reference map (Map 2, *red*) are identified. For each of these, the proportion of their total area contained within the intersection is calculated as an indication of "insideness." The degree of "insideness" is tempered by weighting it by the proportion of total area that the intersection represents of the category or polygon being compared. The sum of the product of each insideness term weighted by its proportion of overlap with the tested polygon gives a GOF term that is unitless, and is scaled so that GOF scores can be compared across multiple categories or polygons

unitless. Expressed as a percentage, the GOF measure is standardized, and can be compared across categories and maps.

Any Map 1 category that can be exactly comprised of a set of Map 2 categories will show a perfect fit with this measure (Fig. 3a). The "insideness" of all completely contained categories is 1, and the weighting factor is the proportion of the area that they represent, which must sum to 1 for a perfect fit. The Map 1 category shown in Fig. 3b is a better fit to the

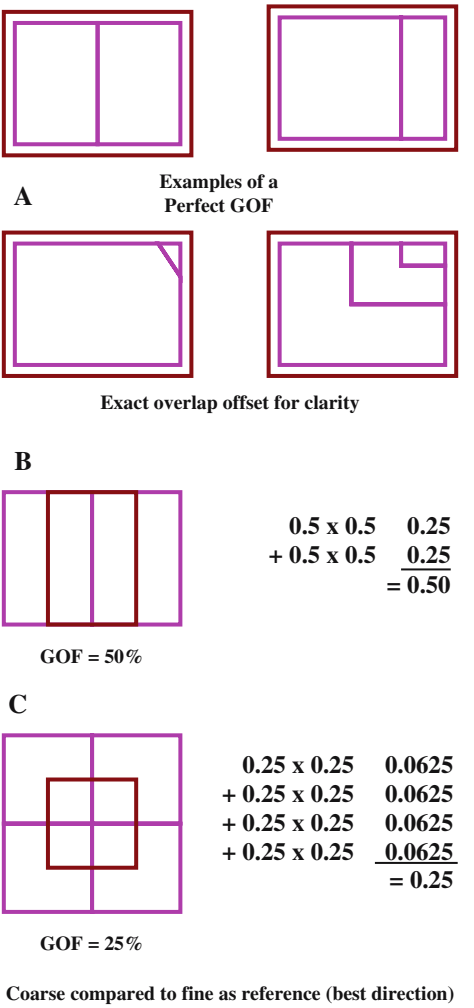

Fig. 3 Examples of goodness-of-fit (*GOF*) scores used with Mapcurves. **a** No matter what the spatial configuration or type of division, sets of polygons or categories in one map that exactly constitute a polygon or category in another map will show perfect GOF. All wholly included polygons have maximum "insideness" and, when multiplied by the area of overlap and summed, equal a perfect score, irrespective of the number of intersecting components. This design allows GOF of maps created by "splitters" and those created by "lumpers" to be compared regardless of the level of division that has been used. Outer boundaries of these groupings should coincide exactly, but have been depicted as adjacent for clarity. Examples shown in **b** and **c** indicate how this GOF changes for polygons sharing different degrees of spatial overlap with polygons from a reference map. The polygon shown in **b** has a higher GOF score than the one shown in **c**, as intuitively expected

intersecting reference polygons than the one shown in Fig. 3c, and this is intuitive, since there is more common overlap in the former.

GOF is tested for each of the categories in Map 1 to estimate the directional fit of Map 1 compared to reference Map 2. GOF is calculated

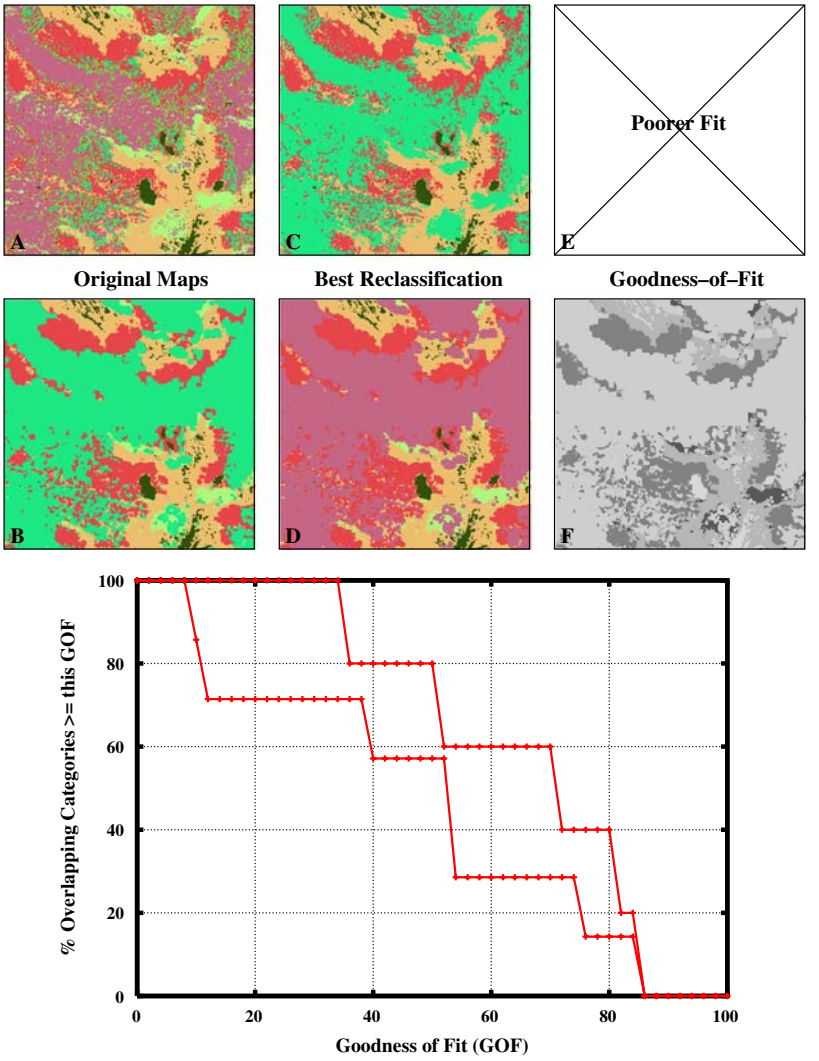

Fig. 4 Comparison of a First Pair of Test Maps using Mapcurves. Maps A and B are being compared, and maps in each row are derived from these. The middle maps in each row (Maps C and D) are reclassified by the translation table that maximizes the resemblance to the reference map by changing the label of the entire category and applying the reference map's color table. The rightmost maps in each row (Maps E and F) show the goodness-of-fit (*GOF*) for each category. White is the highest GOF, and black is the lowest. Mapcurves resulting from both possible comparison directions are shown below. The uppermost Mapcurve reflects the comparison of Map B to Map A as a reference, and this is the best fit (GOF score = 0.6470). The lower Mapcurve shows the opposite comparison (Map B score = 0.4621), and is disregarded. These relative scores indicate a slightly greater degree of resemblance of Map D to Map A relative to the resemblance of Map C to Map B. A color version of this figure is available at http://www.geobabble.ornl.gov/JGS

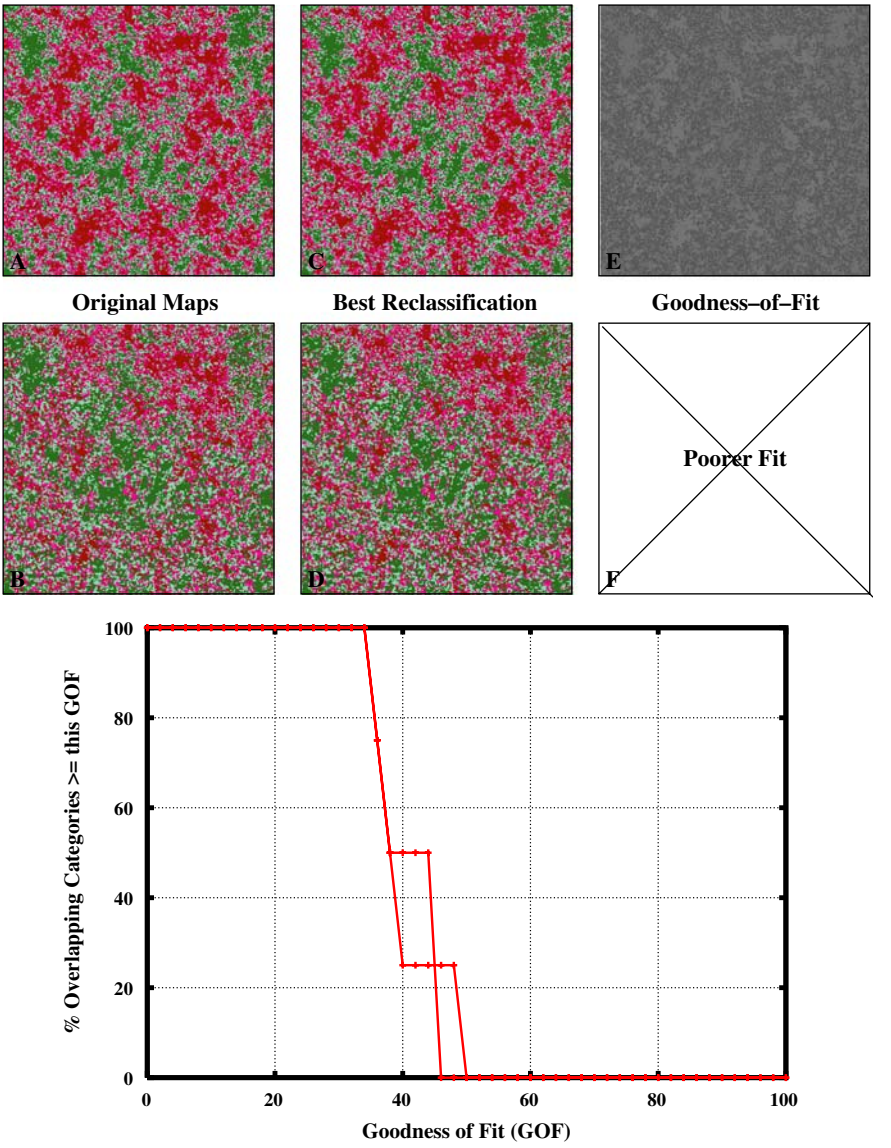

Fig. 5 Comparison of a second pair of test maps using Mapcurves. Figure components as explained in Fig. 4. The identity of the first and second maps in each row indicates that each test map is already as much like the other as simple category reassignment can make it. The Mapcurves indicate that both maps have four categories (four possible tiers in the graphs), and actually cross over each other. Although nearly equal, Map A (GOF score = 0.4030) is a slightly better fit than Map B (GOF score = 0.4028), and is the uppermost Mapcurve. The Mapcurves and GOF score indicate that this pair of maps has a poorer GOF than the first test pair compared in Fig. 4. A color version of this figure is available at http://www.geobabble.ornl.gov/JGS

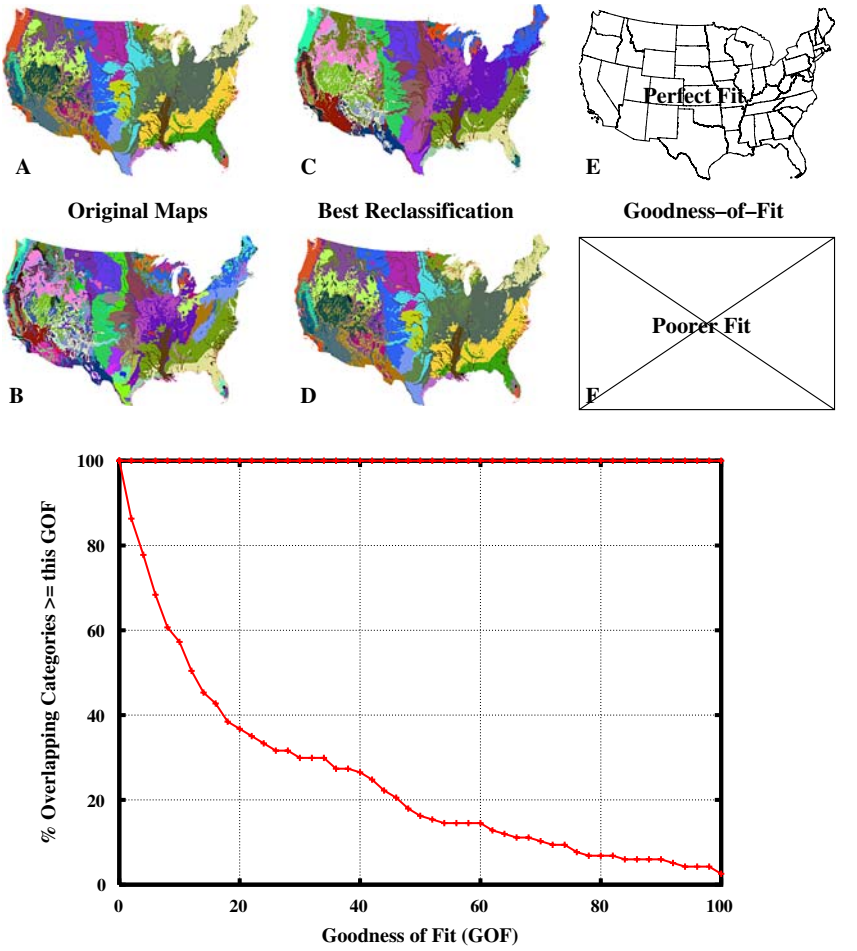

Fig. 6 Comparison of Kuchler's national vegetative types and vegetative forms maps. Figure components as explained in Fig. 4. Although not clear from simple inspection, Kuchler's Types map (Map B) is a subdivision of his Forms map (Map A). This is shown by the fact that the reclassified Types map (Map D) exactly matches the original Forms map (Map A), and also by the fact that the goodness-of-fit (*GOF*) for all categories in the Forms map is perfect (Map E, empty state borders shown to outline the all white map). The Mapcurves also reflect this exact nesting; the comparison of the coarse map to the fine map as reference is a perfect fit (GOF score = 1.0, horizontal Mapcurve across top of graph). The fine map to coarse map comparison is poorer (bottom Mapcurve, GOF score = 0.2479). This is intuitive, since it will always be more difficult to make a coarse map look like a finer one. A color version of this figure is available at http://www.geobabble.ornl.gov/JGS

separately for each category from Map 1 according to the algorithm shown in Fig. 2. All categories in reference Map 2 sharing any spatial overlap are involved in the GOF summation for that Map 1 category. Since the map comparison is based on spatial overlap, it is not necessary for both maps to

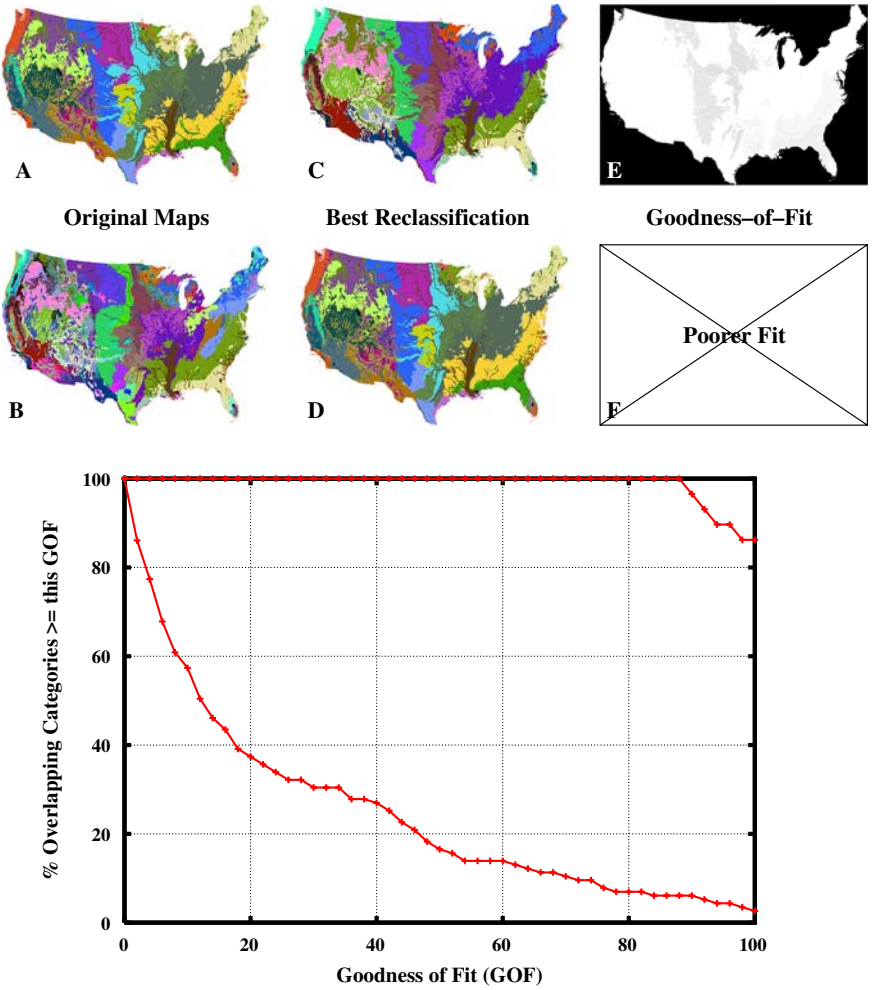

Fig. 7 Altered Version of Kuchler's national vegetation types map to show effect on mapcurves. Figure components as explained in Fig. 4. Two of the 118 categories in Kuchler's vegetation types map were eliminated by combining them with neighboring categories, in order to slightly degrade the perfect nested hierarchical fit, and the Mapcurves analysis from Fig. 6 was repeated. The GOF score of Map A is now reduced to 0.9899, and the uppermost Mapcurve now deviates from perfect horizontal by descending in four discrete steps. These four steps correspond to the four categories that were altered in the map (two combined with two others). The GOF map (Map E) is no longer pure white, but shows the same four altered categories in shades of light gray. The GOF score for Map B increases slightly to 0.2496, bringing the two Mapcurves slightly closer together since the difference in their number of categories has been slightly reduced. Altering one map results in slight changes to both Mapcurves. A color version of this figure is available at http://www.geobabble.ornl.gov/JGS

have the same number of categories in order to be compared. A single GOF value is produced for each category present in Map 1, regardless of the number of categories from the reference Map 2 that overlap with it.

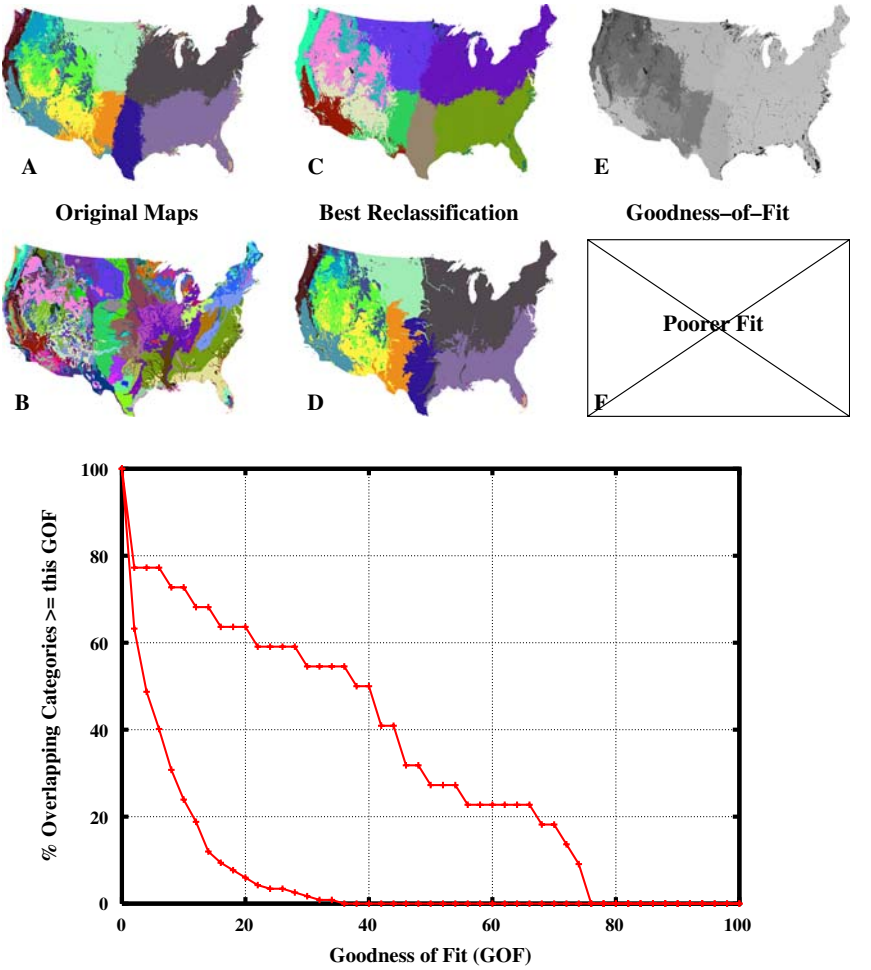

Fig. 8 Comparison of Hargrove/Hoffman statistical ecoregions, with 25 ecoregion divisions, with Kuchler's national vegetation types map using Mapcurves. Figure components as explained in Fig. 4. The Hargrove/Hoffman 25 ecoregion map (Map A) has the best fit (GOF = 0.3442) using the Kuchler map as reference. Reclass Map D shows the best comparison with original Map A. Major rivers and wetlands are responsible for biggest differences between the maps, along with the mountainous regions of the western US (Map E). With a GOF score of 0.4578, the 10 ecoregion Hargrove/Hoffman map is an even better fit than the 25 ecoregion version shown here, and corresponds more closely with Kuchler Types than the second pair of test maps do with each other (GOF score = 0.4029). A color version of this figure is available at http://www.geobabble.ornl.gov/JGS

Conversely, calculating the fit of each category from Map 2 using Map 1 as a reference provides the quantitative comparison in the opposite direction.

Translation tables are produced that show the best possible recoding of categories in one map to maximize the fit to the other map. Each entry in the translation table shows the single map category in the reference map having the greatest amount of spatial overlap with each category in the translated

map. Different translation tables exist for each direction of the pairwise comparison. Using the translation tables, categories in each map can be reclassified such that one map resembles the other map as much as possible when entire categories are re-assigned. Each category can also be colored by the goodness-of-fit score to show parts of the map where agreement with the reference map is relatively good or poor. In this way, each map in a pairwise comparison can be reclassified to show the spatial locations of categories where fit is good and categories where fit is poor.

2.3 Mapcurves

A Mapcurve is a GOF power curve showing the decline in percentage of map categories on the *y*-axis that still satisfy an increasing GOF threshold on the *x*-axis (Figs. 4, 5, 6, 7, 8, bottom). A cumulative frequency distribution is plotted from each directional comparison showing the percentage of categories in one or multiple maps that meets or exceeds a particular sliding threshold of GOF. As the GOF threshold is increased, a smaller percentage of map categories satisfy or surpass that level of fit.

All Mapcurves start at the top left corner of the graph, since in all map comparisons, 100% of the intersecting categories have a 0% or greater match (Fig. 4). The Mapcurve resulting from a perfect match is a straight horizontal line along the top of the graph. Each Mapcurve is monotonically decreasing. If a Mapcurve intersects the right edge of the graph, this point indicates the percentage of overlapping categories within the comparison map that wholly contain categories within the reference map. Thus, a perfect fit Mapcurve running along the top of the graph indicates that 100% of comparison map categories completely contain reference map categories. The poorest possible fit would be indicated by a steep, rapid plunge to the *x*-axis.

The area under the Mapcurve can be used as a single index for the GOF of the entire map to the reference map. Mapcurves that are higher and integrate more area indicate better matches between maps. Since they are plotted on standardized axes, all Mapcurves are comparable, and reveal which map comparisons represent closer matches. A flat Mapcurve across the top of the graph resulting from a perfect comparison represents an integrated area of 1.0, which is $100 \times 100\%$.

Two Mapcurves are produced from each pairwise map comparison (one for each direction). Whichever of these Mapcurves integrates more area indicates the comparison of the coarser map to the finer map as a reference, and is the relevant optimal direction of comparison. The other Mapcurve of the pair can be ignored. The direction of the most favorable comparison usually switches as the number of categories in one map exceeds those in the other, although artificial maps can be designed for which this is not the case. The most favorable direction of comparison cannot easily be determined before the full Mapcurves analysis is performed.

We compared two sets of test maps, as well as other well-known vegetation and ecoregion maps, in order to explore and demonstrate the behavior of Mapcurves. While several of these comparisons have expected outcomes, some do not. Finally, we use Mapcurves to rank pairwise comparisons

among a number of popular landcover and ecoregion maps, including some comparisons anticipated to display a poor GOF.

3 Results

3.1 Comparison of test maps using Mapcurves

Figure 4a and b show the first pair of maps to be compared, each with the same random color table assignment (although this does not imply correspondence). Each map in the same row of Fig. 4 stems from one of the original maps. Map A has seven categories, while Map B has only five. Map C shows Map A reclassified to match Map B as well as possible, and assigned Map B's color table. Similarly, Map D shows the best reclassification of Map B to match Map A.

The pair of Mapcurves resulting from the two directions of this comparison are shown at the bottom of Fig. 4. The higher of the two Mapcurves represents the comparison of Map B to Map A as a reference, and this is the most favorable comparison direction. Five descending steps can be seen in this Mapcurve, corresponding to the five categories in Map B (the lower Mapcurve has seven potential descending steps). Map B's score when compared to Map A, calculated by integrating the area under the higher Mapcurve, is 0.6470. Therefore, the best comparison is reclassified Map D with Map A. Map F shows Map B with each category colored by its GOF with Map A (lighter colors indicate a better fit). In Map F, GOF is shown for whole categories, not individual patches. The GOF for particular patches may be good, but the category is assigned a single GOF value and gray scale that represents the fit across the entire map. It is not necessary to draw Map E, since this represents the poorer comparison direction.

The second pair of maps to be compared, their derivatives, and Mapcurves are shown in Fig. 5. That the first and second maps shown in each row are identical indicates that each Test Map is already as much like the other as simple category reassignment can make it. The GOF scores for these maps are nearly equivocal, but Map A has a slightly higher GOF of 0.4029, making the reclassified Map C and Map B comparison slightly better. The Mapcurves show that both maps have four categories, and the curves actually cross over each other. Comparison of their GOF scores shows that the second pair of maps is a much poorer fit with each other than the first pair.

3.2 Comparison of Kuchler vegetation maps

We also used Mapcurves to compare Kuchler's Vegetation Forms and Kuchler's Vegetation Types maps, from coverages digitized at the United States Environmental Protection Agency from the 1979 Physiographic Regions Map produced by the Bureau of Land Management, which added 10 physiognomic types to Kuchler's 1964 United States Geological Survey (USGS) Potential Natural Vegetation map (Kuchler 1964) [and similarly

differs from the 1985 USGS map revised by Kuchler and others, Kuchler (1993)]. Each of these maps has a resolution of 5 km^2.

Although not obvious by simple inspection, Kuchler's Vegetation Types map (118 categories, Fig. 6 Map B) is a subdivided version of his Vegetation Forms map (29 categories, Fig. 6 Map A). The Mapcurves method immediately shows this to be the case, since the comparison is perfect (upper flat horizontal line, Fig. 6 bottom), and the Map A GOF score is 1.0. The reclassified Map D is identical with Map A, and, when colored by GOF, all categories in Map E are white (empty state boundaries are shown in Fig. 6e to outline the perfect fit of the all-white map). The flat horizontal Mapcurve represents the coarser Map A compared to the finer Map B as a reference. The lower Mapcurve (fine to coarse as reference) intersects the right edge of the graph, indicating that about 2% of the categories in the fine map completely contain categories in the coarse map. Indeed, category 48, California steppe, is identical with category 9 from Map A, California grassland. Similarly, category 52 from Map B, Alpine meadows and barren, is identical with category 11 from Map A, Alpine meadow. Since these two categories are not subdivided, 1.7% of the categories (2 of 118) are completely contained.

As a demonstration and test, we altered Kuchler's finer Vegetation Types map by combining two small spatially contiguous categories with their neighboring categories. In this new test map, we eliminated the sandhills in Nebraska (category 89) by combining it with Oak/hickory/pine (category 111), and we re-labeled the Blackbelt in Mississippi and Alabama (category 75) to now become Grama/Buffalo grass (category 65). These changes were designed to slightly degrade the perfectly nested, hierarchical fit of these two maps.

The Mapcurves comparison of this new map with Kuchler's original Vegetation Forms map is shown in Fig. 7. As before, reclassified Map D is the best comparison with Kuchler's original Forms Map A, but now the two changes in Nebraska and Mississippi can be seen. The altered Map B's GOF score is now 0.9899. The GOF Map E shows four categories in light gray, the two combined categories and the two categories with which they were combined. The uppermost Mapcurve (Fig. 7, bottom) now deviates from the perfect horizontal fit shown in Fig. 6. At the upper right, the altered Mapcurve descends four steps, corresponding to the two categories that were blended with two others. Both Mapcurves were altered by the changes to the finer map only, although the change in the lower curve is subtle due to the large number of categories in the finer map. The gap separating the two Mapcurves narrows as the difference between the number of categories in the two maps decreases.

3.3 Comparison of Hargrove/Hoffman ecoregions with Kuchler types

Hargrove and Hoffman (1999, 2004a, b) have experimented with ecoregionalizations created using Multivariate Geographic Clustering (MGC). MGC uses non-hierarchical multivariate clustering, employing the iterative *k*-means algorithm of Hartigan (1975) to produce national ecoregions statistically at a resolution of 1 km^2, based on a number of abiotic environmental variables. Normalized variable values from each map raster

cell are used as coordinates to plot each map cell in a data space with as many axes as there are multivariate environmental descriptors. Similarity is inversely related to separation distance in this data space. The MGC process iterates on a parallel supercomputer until it converges on a particular classification structure.

The user can specify the number of clustered ecoregions which result from the process, making it possible to divide the map into a few large, coarsely-defined ecoregions or a larger number of small, finely-resolved ones. All large ecoregions produced by MGC have a similar upper limit on within-group variance. This control on heterogeneity across ecoregions prevents delineation of highly variable regions in the same map with ones that are more homogeneous.

Hargrove and Hoffman (2004a) have produced as many as 5,000 US ecoregions on the basis of 25 environmental factors, including elevation, mean and extremes of annual temperature, mean monthly precipitation, soil nitrogen, organic matter, and water capacity, frost-free days, soil bulk density and depth, and solar aspect and insolation. Ecoregions created with MGC are useful for characterizing regional borders (Hargrove and Hoffman 1999), predicting species ranges (Hargrove and Hoffman 2003), statistically designing large networks of sensors or samples (Hargrove et al. 2003; Hargrove and Hoffman 2004b; White et al. 2005) and detecting trends in other complex multivariate phenomena, such as simulation output from global circulation models (Saxon et al. 2005; Hoffman et al. 2005).

Because maps generated by MGC represent a way to vary the number of division categories present in the map, they offer a unique chance to test the Mapcurve comparison method. A series of national ecoregions can be produced at different levels of division, from fine to coarse, all based on the same set of multivariate environmental descriptors. Because MGC is non-hierarchical, all borders between ecoregions are re-drawn for each separate level of division.

We compared a Hargrove/Hoffman map containing 25 ecoregions, created using our MSTC process based on the 25 environmental variables described above, with Kuchler's Vegetation Types map (Fig. 8a, b). With the finer Kuchler Vegetation Types map serving as the reference, the Hargrove/Hoffman 25 ecoregions map has the higher map score of 0.3442. Reclassed Map D is the best version for comparison with Map A.

When the Hargrove/Hoffman map is colored by GOF to Kuchler's Vegetation Types (Fig. 8, Map E), major river systems are darkly highlighted as strong differences. Wetlands and swamps also differ between the two maps, and the Everglades, the Okefenokee, the Dismal Swamp, and the Mississippi Delta show as poor GOF, since Kuchler's Vegetation Types map does not contain river or wetland features. Other differences exist between these two maps, particularly in the highly dissected Pacific Northwest (PNW). At this level of ecoregion division, the MGC method does not subdivide the PNW, while the Kuchler Types map does. Nevertheless, a Hargrove/Hoffman 10 ecoregion map is an even better fit with Kuchler Vegetation Types, having a Mapcurves score of 0.4578.

GOF maps like Map E in Fig. 8 are not area-weighted in any way. Instead, the map GOF score is obtained as the mean of the single gray level

taken from each category in the map. Thus, a category covering a large portion of the map could have a poor GOF and be dark, but, if sufficient numbers of other categories with high GOF exist, the GOF score for the map could be high. Mapcurves are based on the proportion of the map's categories exceeding a particular GOF threshold.

3.4 Comparison of common ecoregion and landcover schemes

We compared GOF of several common ecoregion and landcover schemes for the conterminous United States using Mapcurves. Table 1 shows the rank order of GOF when selected pairs of these well-known categorical maps are compared using Mapcurves. Because Mapcurves are standardized, the relative GOF between any two maps can be compared to the GOF between other pairs.

Three distinct types of maps are compared in Table 1. Landcover maps, as descriptions of the type of extant vegetation, represent maps of the realized environment. Ecoregion schemes based on abiotic conditions alone, like Kuchler and Hargrove/Hoffman, predict potential vegetation. Ecoregion schemes that include biotic interactions and human and natural disturbances, like Bailey Aggregated ecosystems, are intermediate between realized and potential, and should be correlated with actual vegetation.

Comparisons between maps within any one of these three types tend to produce higher GOF scores than comparisons across types. Comparisons of landcover maps to other landcover maps show a strong tendency to be at the top of the list, with high GOF (Table 1). Disagreement about the type of vegetation presently existing at each location may be more about nomenclature than disagreement about what type of vegetation is actually there. This may be a more well-defined problem than delineating abstract ecoregions, leading to the higher GOF values.

The comparison of the 200- and 300-ecoregion Hargrove/Hoffman maps, as a potential versus potential comparison, shows a high GOF. Although they are both based on the same data, all ecoregion borders are redrawn each time the non-hierarchical analysis is repeated. Similarly, the comparisons of Bailey Aggregated with Major Land Resource Areas (MLRAs), a realized versus realized comparison, is fairly high in the GOF list.

The GOF is expected to be less than perfect when potential vegetation ecoregions are compared to realized vegetation ecoregions or landcover. Many locations will not have their potential vegetation, since they have been reset or altered by histories of natural disturbance or anthropogenic manipulation. These expected differences place potential versus realized comparisons further down in Table 1.

Table 1 also includes comparisons involving state borders, a well-known map that might be expected to demonstrate poorer fits with ecoregion and landcover schemes. State borders are not drawn without regard to ecological features, and often coincide with rivers or mountain ranges. Each spatially disjoint area in this map was assigned a separate category, so that the Upper Peninsula of Michigan, for example, is labeled differently from the Lower

Table 1 Rank-ordered goodness-of-fit (GOF) scores when several common ecoregion and landcover schemes for the conterminous United States are compared using Mapcurves

First map	Number of categories	Second map (reference)	Number of categories	Best GOF from Mapcurves
BATS landcover	19	Olson landcover	67	0.9853
USGS landcover	27	Olson landcover	67	0.8523
IGBP landcover	17	USGS landcover	27	0.7824
Hargrove/Hoffman	200	Hargrove/Hoffman	300	0.7770
IGBP landcover	17	BATS landcover	19	0.7561
Bailey aggregated	11	MLRAs	221	0.7410
BATS landcover	19	USGS landcover	27	0.7153
Bailey aggregated	11	Hargrove/Hoffman	5,000	0.7097
Bailey aggregated	11	Hargrove/Hoffman	500	0.5822
Bailey aggregated	11	Kuchler forms	29	0.4606
Bailey aggregated	11	State borders	59	0.4490
State borders	59	MLRAs	221	0.4088
Bailey aggregated	11	Olson landcover	67	0.3879
Hargrove/Hoffman	10	Kuchler forms	29	0.3858
Hargrove/Hoffman	300	Kuchler forms	29	0.3843
Bailey aggregated	11	Hargrove/Hoffman	25	0.3417
Hargrove/Hoffman	12	State borders	59	0.3412
Hargrove/Hoffman	10	Olson landcover	67	0.3249
Hargrove/Hoffman	10	Bailey aggregated	11	0.3022
Kuchler types	116	MLRAs	221	0.2685
Kuchler types	116	IGBP landcover	17	0.2632
Kuchler forms	29	State borders	59	0.2550
Hargrove/Hoffman	25	Kuchler forms	29	0.2332
Olson landcover	67	MLRAs	221	0.2049
IGBP landcover	17	Kuchler forms	29	0.1985
State borders	59	Olson landcover	67	0.1204
State borders	59	Random spray	59	0.0172

Because Mapcurves GOF scores are standardized and eliminate the effects of different numbers of categories in the maps being compared, Mapcurves permit a quantitative ranking of GOF scores from comparisons of different maps. The Mapcurve algorithm was applied at the level of map categories. All maps compared at 1 km^2 resolution. Comparisons with a map of the borders of the states within the conterminous United States, and with a map of random cells are also included. The GOF score shown in each case is the area integrated under the highest Mapcurve (resulting from the most favorable comparison direction). The best possible GOF score is 1.0 ($100 \times 100\%$)

Peninsula. Most comparisons with the map of state borders show intermediate to low GOF. Finally, state borders were compared with a random pattern containing an equal number of categories, as an example of an expected poor GOF.

4 Discussion

Mapcurve comparisons behave intuitively, and are interpretable. In contrast to cell-by-cell contingency table analysis, two checkerboard maps shifted laterally by one cell width will show a perfect fit using Mapcurves, since black in one map will equal white in the other.

There are two gaps in GOF in Table 1. The first, between fits of 0.7 and 0.5, separate the comparisons of maps within the same category from

comparisons of maps across categories. The second gap is below 0.1, which isolates the random map comparison from all others. Relative to this GOF spectrum, the first pair of compared maps (Fig. 4) falls into the first gap at a GOF of 0.6470. The second pair of compared maps (Fig. 5) falls into the middle of the pack with a GOF of 0.4029.

It is difficult to tell whether the Mapcurves algorithm has removed all of the confounding effects when comparing maps with differences in the number of categories. The finer the map, the better its categories can fit into any coarse map. This is the reason why high quality video monitors strive for more and smaller pixels. Any image can be displayed in higher quality on a video monitor that contains more pixels. Because of this pixelation effect, it may be impossible to completely remove the artifacts associated with extreme differences in numbers of categories in maps being compared.

However, many of the Mapcurve comparisons involving families of Hargrove/Hoffman ecoregions show peaks of fit that do not necessarily correspond to the ecoregion map that most closely matches the number of categories in the map to which they are being compared. It is not the case that the Hargrove/Hoffman map having the closest match in number of categories has the highest GOF score. This suggests that much of the difference between maps created by differences in their numbers of categories has been successfully eliminated by the Mapcurves analysis. It is not uncommon, however, for the Hargrove/Hoffman ecoregion maps having the largest and/or the smallest numbers of divisions to show the best fit overall.

The Mapcurves algorithm has been applied here to entire categories within maps. Mapcurves could also be applied to each spatially separated patch in each raster map, or they could be applied to each individual polygon in a vector environment. Such application would represent a much more stringent test, and would show not just which categories fit best or most poorly overall, but would instead show exactly in which patches or polygons the fit was best and worst. If several patches were superimposed exactly in both maps, but another was not, the GOF map would show a dark spot only for the single patch that did not spatially agree. Such a patch-based Mapcurves application would likely have detected the sharp straight-line "seam" present in the second pair of maps that was found and reported by several of the other papers in this volume. Applying the Mapcurves algorithm by patch or polygon would be much more computationally intensive than comparison by categories. This would be especially true for maps like the second pair, whose categories have a much greater number of disjunct patches.

While it can identify sets of categories in one map that are a close or perfect match with a single category in another, Mapcurves cannot find sets in both maps whose unions are spatially equivalent. Such higher-order many-to-many matches will be scored as having only mediocre GOF, even if the spatial match of the two sets is perfect. Thus, maps created by two "splitters" who split the same mother categories on the basis of different criteria will not show a high GOF. State-based maps of, say, geology and land cover would likely show low GOF, even though a certain combination of categories or polygons in each would equal the outline of any given state border. A tool that would test for such higher-order fits would be

computationally intensive, since it would need to examine all possible combinations of sets of categories in both maps, which would require on the order of $n \log(n)$ comparisons.

Mapcurves are not limited to pairwise comparisons; large numbers of maps could be compared using the same GOF algorithm. Such a multiple comparison would sequentially compare each map in the suite to all others as reference maps, generating a sum like that shown in Fig. 2 for each reference map in turn. The sums for each reference map in the set would be added together, and then divided by the number of reference maps $(n-1)$. This summation would provide a GOF for each polygon, averaged over all reference maps against which it has been compared.

Such multi-way comparisons would produce a separate Mapcurve for each map being compared, and would show the GOF of that map with all other maps used as references. As with pairwise comparisons, each feature in each map would receive a GOF score, and a GOF score would be calculated for each map. The Mapcurve subsuming the greatest area under the curve would describe the best fit among all of the maps in the group, when the rest of the maps are used as references.

GOF feature scores across multiple maps could be used to quantify the degree of spatial uncertainty for each patch across a larger set of map realizations. Such realizations could be created synthetically (e.g., Hargrove et al. 2002; Jager et al. 2005), or could represent maps using location information from different sources. The single map exhibiting the highest GOF with respect to all others would be the map in which features were "centered" in the most probable locations, considering the locations of these features displayed in all other maps. Such a map, once identified, might be the one best selected for use. The GOF of each feature in this map would show its degree of spatial certainty, with respect to all of the other maps in the suite.

Mapcurves are a general approach for examining goodness of fit when comparing two or more maps, and are not tied to any particular goodness of fit measure, or even to spatial data. The results of any GOF measure could be portrayed as a set of power curves in the same way. Similarly, Mapcurves could be used to portray the degree of concordance between two or more group classifications of non-spatial data, using a GOF method based on something other than spatial overlay.

Mapcurves analysis requires nothing more sophisticated than standard GIS tools, yet allows quantitative comparison of multiple categorical maps. Results of the comparison are given in a single quantitative score, as well as shown in a spatially explicit way via GOF maps. Mapcurves could be included as a standard analysis feature in future releases of commercial GIS packages.

Acknowledgments This manuscript was substantially improved by comments from John Bell, Rebecca Efroymson and three anonymous reviewers. Research partially sponsored by the USDA Forest Service under Agreement Number PNW 03-IA-11261927-532 with Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE–AC05–00OR22725. The authors wish to express their extreme sadness at the untimely death of Ferko Csillag, not only for the loss to his colleagues and family, but also

for the loss which we feel his passing represents to the discipline of statistical ecology.

References

Costanza R (1989) Model goodness of fit: a multiple resolution procedure. Ecol Modell 47:199–215

Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sensing Environ 80:185–201

Hagen A (2003) Fuzzy set approach to assessing similarity of categorical maps. Int J Geogr Inf Sci 17(3):235–249

Hargrove WW, Hoffman FM (1999) Using multivariate clustering to characterize ecoregion borders. Comput Sci Eng 1(4):18–25

Hargrove WW, Hoffman FM (2003) An analytical assessment tool for predicting changes in a species distribution map following changes in environmental conditions. In: Proceedings, GIS/EM4 conference, Banff, Alberta, Canada, Sept. 2–8, 2000. CD-ROM, ISBN: 0-9743307-0-1

Hargrove WW, Hoffman FM (2004a) The potential of multivariate quantitative methods for delineation and visualization of ecoregions. Environ Manage 34(5):S39–S60

Hargrove WW, Hoffman FM (2004b) A flux atlas for representativeness and statistical extrapolation of the AmeriFlux Network. ORNL Technical Memorandum ORNL/TM–2004/112. Available at http://www.geobabble.ornl.gov/flux-ecoregions

Hargrove WW, Hoffman FM, Schwartz PM (2002) A fractal landscape realizer for generating synthetic maps. Conserv Ecol 6(1):2. [online]: http://www.consecol.org/vol6/iss1/art2

Hargrove WW, Hoffman FM, Law BE (2003) New analysis reveals representativeness of the AmeriFlux network. Eos 84(48):529–535

Hartigan JA (1975) Clustering algorithms. Wiley, New York

Hoffman FM, Hargrove WW, Erickson DJ III, Oglesby R (2005) Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models. Earth Interact (in press)

Jager HI, King AW, Schumaker NH, Ashwood TL, Jackson BL (2005) Spatial uncertainty analysis of population models. Ecol Modell (in press)

Kuchler AW (1964) Potential natural vegetation of the conterminous United States, American Geographical Society, Special Publication No. 36

Kuchler AW (1993) Potential natural vegetation of the conterminous United States. Digital vector data in an Albers Equal Area Conic polygon network and derived raster data on a 5 km^2 Albers Equal Area 590× 940 grid. In: Global Ecosystems Database Version 2.0. Boulder CO: NOAA National Geophysical Data Center

Lantz CA, Nebenzahl E (1996) Behavior and interpretation of the K statistic: resolution of two paradoxes. J Clin Epidemiol 49(4):431–434

Maclure M, Willet WC (1987) Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126(2):161–169

Maxwell WE (1977) Coefficients of agreement between observers and their interpretation. Br J Psychiatry 130:79–83

Metternicht G (1999) Change detection assessment using fuzzy sets and remotely sensed data: an application of topographic map revision. ISPRS J Photogram Remote Sensing 54(4):221–233

Monserud RA, Leemans R (1992) Comparing global vegetation maps with the Kappa statistic. Ecol Modell 62:275–293

Plotnick RE, Gardner RH, Hargrove WW, Prestegaard K, Perlmutter M (1996) Lacunarity analysis: a general technique for the analysis of spatial patterns. Phys Rev E 53(5):5461–5468

Pontius RG Jr (2000) Quantification error versus location error in comparison of categorical maps. Photogram Eng Remote Sensing 66(8):1011–1016

Pontius RG Jr (2002) Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. Photogram Eng Remote Sensing 68(10):1041–1049

Pontius RG Jr, Schneider LC (2001) Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agric Ecosyst Environ 85(1–3):239–248

Power C, Simms A, White R (2001) Hierarchical fuzzy pattern matching for the regional comparison of land use maps. Int J Geogr Inf Sci 15(1):77–100

Saxon E, Baker B, Hargrove WW, Hoffman FM, Zganjar C (2005) Mapping environments at risk under different global climate change scenarios. Ecol Lett 8:53–60

Turner MG, Costanza R, Sklar FH (1989) Methods to evaluate the performance of spatial simulation models. Ecol Modell 48:1–18

White MA, Hoffman FM, Hargrove WW, Nemani RR (2005) A global framework for monitoring phenological responses to climate change. Geophys Res Lett 32(4):L04705 Doi:1029/2004GL021961

Winter S (2000) Location similarity of regions. ISPRS J Photogram Remote Sensing 55(3):189–200

Zadeh L (1965) Fuzzy sets. Inf Control 8:338–353