Effect of Category Aggregation on Map Comparison

Robert Gilmore Pontius Jr. and Nicholas R. Malizia

Graduate School of Geography, George Perkins Marsh Institute, and
Department of International Development, Community, and Environment
Clark University
950 Main St., Worcester, MA 01610 USA
{rpontius,nmalizia}@clarku.edu

Abstract. This paper investigates the influence of category aggregation on measurement of land-use and land-cover change. To date, research concerning data aggregation has examined primarily the effects of modifying the unit of observation (i.e., the modifiable areal unit problem and the ecological inference problem); here, we examine the effects of changing the categorical definition, such as the conversion from many, detailed Anderson Level II classes to fewer, broader Anderson Level I classes. Cross-tabulation matrices are used to analyze the change between two times for aggregated and unaggregated versions of identical landscapes. A mathematical technique partitions the Total change as the sum of Net (i.e., quantity change) and Swap (i.e., location change). This paper shows that the Total and Net exhibited by maps between two points in time can be substantially reduced through land-use category aggregation, but cannot be increased. Swap, however, can be reduced or increased by the aggregation of categories. We derive five principles that dictate the effect of aggregation and illustrate the principles using both simplified examples and empirical data. The empirical data are from three Human Environment Regional Observatory sites. The principles are mathematical facts that apply to the analysis of any categorical variable.

1 Introduction

1.1 Measuring Change on a Map

Land-use and land-cover change (LUCC) analysis has become an integral component of geographic, economic, and ecological research. Changes in land-use and land-cover are either directly responsible for, or synergistically enhance, many forms of environmental change including biodiversity loss, land degradation, and climatic variation [6, 7]. Scientists study change in landscapes over time to determine its causes and effects as well as to model future landscapes. Such research directly affects conservation and development policy. This paper specifies how a decision in the early stages of a LUCC investigation regarding the definition of land-use and land-cover categories can have a profound effect on subsequent analysis and conclusions. Comparison of maps from an initial time A and a subsequent time B is the most common method to analyze LUCC. A typical first step in this comparison is the calculation of a cross-tabulation matrix. Table 1 demonstrates the format of a typical cross-tabulation matrix, where the rows represent the categories of the land-use map at time A and the columns show the categories at time B.

M.J. Egenhofer, C. Freksa, and H.J. Miller (Eds.): GIScience 2004, LNCS 3234, pp. 251–268, 2004.

Table 1. Cross-tabulation matrix to compare maps from two points in time for three categories.

		Time B				
		Category 1	Category 2	Category 3	Total Time A	Loss
Time A	Category 1	P_{11}	P_{12}	P_{13}	P_{1+}	$P_{12} + P_{13}$
	Category 2	P_{21}	P_{22}	P_{23}	P_{2+}	$P_{21} + P_{23}$
	Category 3	P_{31}	P_{32}	P_{33}	P_{3+}	$P_{31} + P_{32}$
	Total Time B	P_{+1}	P_{+2}	P_{+3}	1	
	Gain	$P_{21} + P_{31}$	$P_{12} + P_{32}$	$P_{13} + P_{23}$		

P_{ij} denotes the proportion of the map that shows a transition from category i at time A to category j at time B. Entries on the diagonal represent persistence on the map between the two points in time, thus P_{jj} identifies the proportion of the map that persists as category j. This matrix also calculates the Total amount of each category for each point in time. Entry P_{i+} sums the amount of category i at time A, while entry P_{+j} sums the amount of category j at time B. To this standard matrix we append an additional row and column to calculate the amount of Gain and Loss for each category between time A and time B. The Loss for category i is calculated by summing the off-diagonal entries for category i at time A. Thus, the amount of Loss for category i is equivalent to $P_{i+} - P_{ii}$. The Gain for category j is calculated by summing the off-diagonal entries for category j at time B, which is equivalent to $P_{+j} - P_{jj}$.

Table 2 shows how these basic statistics are further processed to yield more information that is fundamental to comparing maps of a shared categorical variable [10]. The Loss, Gain, and Total columns of Table 2 show that the sum of the Loss and the Gain for each category between time A and time B is the Total for that category. The left side of equation 1 expresses this relationship for category j.

Table 2. Map change budgets derived from the cross-tabulation matrix in Table 1.

Category	Loss	Gain	Total	Net	Swap
1	$P_{12}+P_{13}$	$P_{21}+P_{31}$	$P_{12}+P_{13}+P_{21}+P_{31}$	$\lvert(P_{12}+P_{13})-(P_{21}+P_{31})\rvert$	$\mathrm{MIN}(P_{12}+P_{13}, P_{21}+P_{31}) * 2$
2	$P_{21}+P_{23}$	$P_{12}+P_{32}$	$P_{21}+P_{23}+P_{12}+P_{32}$	$\lvert(P_{21}+P_{23})-(P_{12}+P_{32})\rvert$	$\mathrm{MIN}(P_{21}+P_{23}, P_{12}+P_{32}) * 2$
3	$P_{31}+P_{32}$	$P_{13}+P_{23}$	$P_{31}+P_{32}+P_{13}+P_{23}$	$\lvert(P_{31}+P_{32})-(P_{13}+P_{23})\rvert$	$\mathrm{MIN}(P_{31}+P_{32}, P_{13}+P_{23}) * 2$
Map	$P_{12}+P_{13}+P_{21}+P_{23}+P_{31}+P_{32}$	$P_{12}+P_{13}+P_{21}+P_{23}+P_{31}+P_{32}$	$P_{12}+P_{13}+P_{21}+P_{23}+P_{31}+P_{32}$	$[\lvert(P_{12}+P_{13})-(P_{21}+P_{31})\rvert + \lvert(P_{21}+P_{23})-(P_{12}+P_{32})\rvert + \lvert(P_{31}+P_{32})-(P_{13}+P_{23})\rvert]/2$	$\mathrm{MIN}(P_{12}+P_{13}, P_{21}+P_{31}) + \mathrm{MIN}(P_{21}+P_{23}, P_{12}+P_{32}) + \mathrm{MIN}(P_{31}+P_{32}, P_{13}+P_{23})$

$$\mathrm{Loss}_j + \mathrm{Gain}_j = \mathrm{Total}_j = \mathrm{Net}_j + \mathrm{Swap}_j \tag{1}$$

The situation is slightly different for the map-level analysis where Total equals the sum of Losses, which is also equal to the sum of Gains for the entire map. This is because a Loss of any category implies a Gain of another category. The bottom row of Table 2 demonstrates the relationship that equation 2 dictates for the map level of analysis, this is denoted with subscript M.

$$\mathrm{Loss}_M = \mathrm{Gain}_M = \mathrm{Total}_M = \mathrm{Net}_M + \mathrm{Swap}_M \tag{2}$$

The Total column of Table 2 shows that the Total at the map level is equal to the sum of the off-diagonal entries of Table 1. The right-hand side of equations 1 and 2 show that at both the category level and at the map level, the Total can be partitioned as the sum of Net and Swap.

Net (i.e., quantity change) is the amount of uncompensated change for a category. For category j, the Net is equal to $|P_{+j} - P_{j+}|$. If P_{+j} is greater than P_{j+} then the category is Net gaining, if P_{+j} is less than P_{j+} then the category is Net losing. Table 2 shows that the Net at the map level is one half the sum of the Net for the individual categories [10].

Figure 1 demonstrates this concept, where the urban areas experience Net gain while the forest areas experience Net loss. Urban sprawl emanating from the city of Worcester in the central Massachusetts study area is overtaking the other categories on the map, most notably the eastern forest areas. In Figure 1, urban is a Net gaining category while forest is a Net losing category.

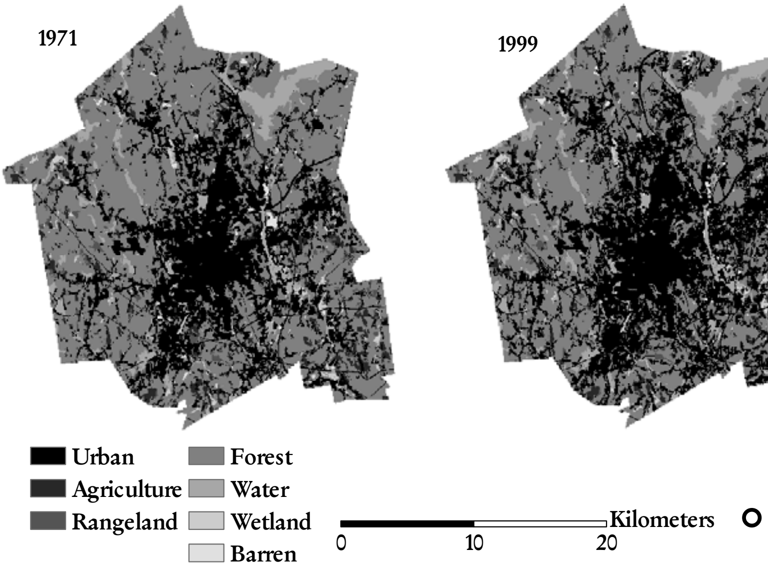


Fig. 1. These maps show the land-use change in the Massachusetts study area between 1971 and 1999. The change on these maps is predominantly Net gain of urban and Net loss of forest.

Swap (i.e., location change) is the amount of compensated change for a category. Swap for a category is equal to twice the minimum of the Gain or Loss for the category (i.e., Swap is equal to twice the minimum of P_{+j} and P_{j+} for category j). This is because the category-level Swap derives from pairing as many gaining pixels with losing pixels of the same category as possible; while the Net is the remaining unpaired pixels. Table 2 shows that Swap at the map level is equal to one half of the sum of the Swap for the individual categories [10].

Figure 2 demonstrates the concept of Swap in a map of the border between Arizona, U.S. and Sonora, Mexico. The rangeland and barren lands retain a relatively similar quantity of their respective categories on the map while the locations of those categories change. The change observed on these maps is due mainly to classification error. The maps show a considerable amount of apparent transition between the rangeland and barren categories probably because of the difficulty in distinguishing between these two land-use/cover categories using remotely sensed data. While both categories lose a large amount between time A and time B, they also gain a substantial amount that compensates for the loss, which yields a set of maps dominated by Swap.

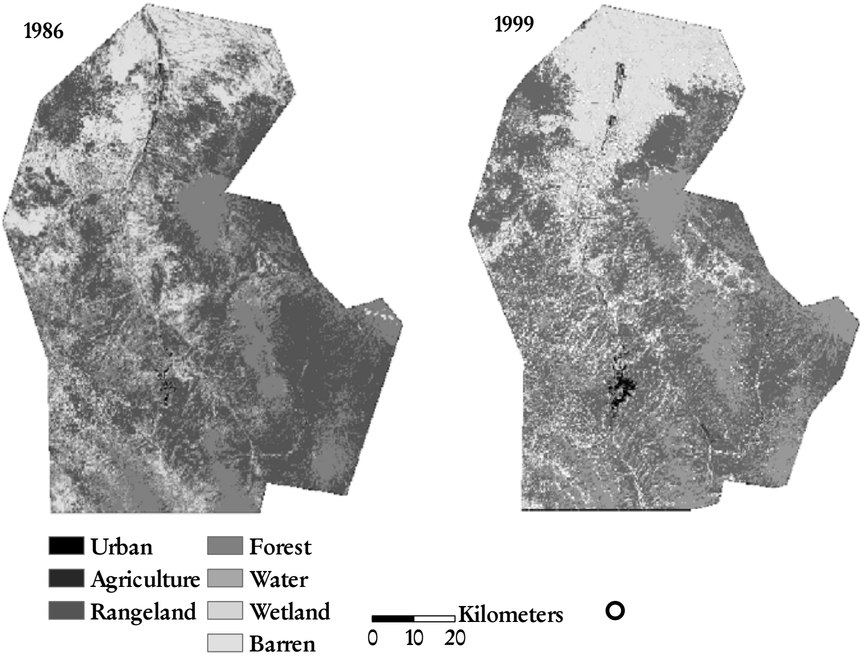


Fig. 2. These maps show the land-use change in the Arizona study area between 1986 and 1999. The change on these maps is predominantly Swap in the rangeland and barren categories.

1.2 The Problem of Aggregation

Figures 3 and 4 illustrate the potentially important effect of category aggregation on the measurement of LUCC. Both figures show maps of the landscape of Kansas' Grey County for 1985 and 2001. Figure 3 shows these maps classified at a modified Anderson Level II [1]; whereas Figure 4 shows the same maps aggregated to Anderson Level I to allow for comparison with the maps in Figures 1 and 2. The Anderson classification system was employed in the cross-site analysis to provide a standard metric with which to compare sites. Level I of this classification system can distinguish a map of 9 classes, which can be sub-divided to make a total of 37 possible categories of land uses or covers at Level II [1].

While the unaggregated maps in Figure 3 show nine categories at Level II, the aggregated maps in Figure 4 show only five categories at Level I. This is due to the amount of categorical detail that is identified within the agriculture category at

Anderson Level II. Many categories at Anderson Level II (confined feeding operations, other cropland, cool season crops, warm season crops, and two season crops) become one category in Anderson Level I (agriculture).

This change in category definition has a substantial influence on the amount of change measured on the map. At Anderson Level II, there is 61% Total in the maps between 1985 and 2001. Of that change, 48% is Swap while 13% is Net. Almost all of this change is due to the seasonal variation within the detailed agricultural categories. Analysis of the same landscape at Anderson Level I yields substantially different results. At Anderson Level I, 13% change is observed on the maps. Swap accounts for 10% of that change while Net claims the small remaining portion. This demonstrates the dramatic effect that the categorical scale can have on analysis, a problem we christen the "category aggregation problem" (CAP).

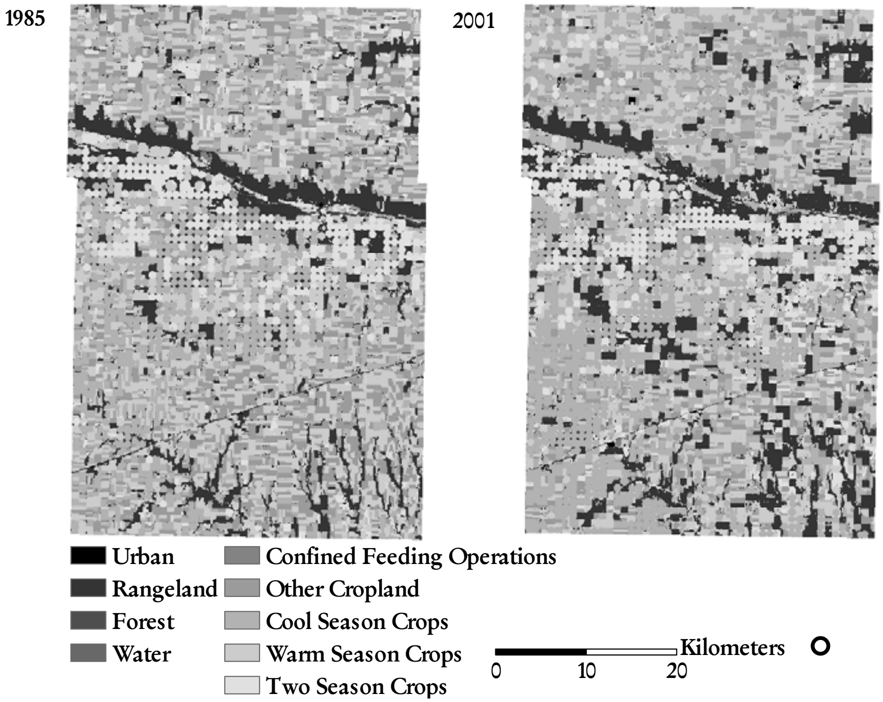


Fig. 3. These maps show the land-use change in the Kansas study area between 1985 and 2001. These maps are classified using Anderson Level II classification. Most of the change on the maps is Swap.

While some scientists might have an intuitive idea of how an aggregation scheme can change the information in a dataset, this paper provides scientists with a mathematical explanation for the effects of the manipulation of the categorical definition. It is important that scientists know exactly how category aggregation influences the measurement of difference between two maps, because category aggregation is an extremely common practice and the effect of some aggregation schemes are not intuitive. Scientists aggregate categories for many reasons. One of the most important reasons is to allow for comparison across diverse sites, times, and datasets. It is usually easiest to compare various maps when each map has the same categories; the most common way to attain this is to aggregate the data from available maps to a common set of broad categories. A second important reason for aggregation is to allow for simplification and reduction of the data. Usually the scientist wants to focus on the dynamics of the most important categories, while the available data may contain many more categories than just the most important ones. Scientists do not want the aggregation step to eliminate some potentially important information or to introduce some spurious signals. If scientists know exactly how an aggregation scheme influences the information in the maps, then they can reformat the data with confidence, ultimately in order to use a larger array of maps and statistical techniques.

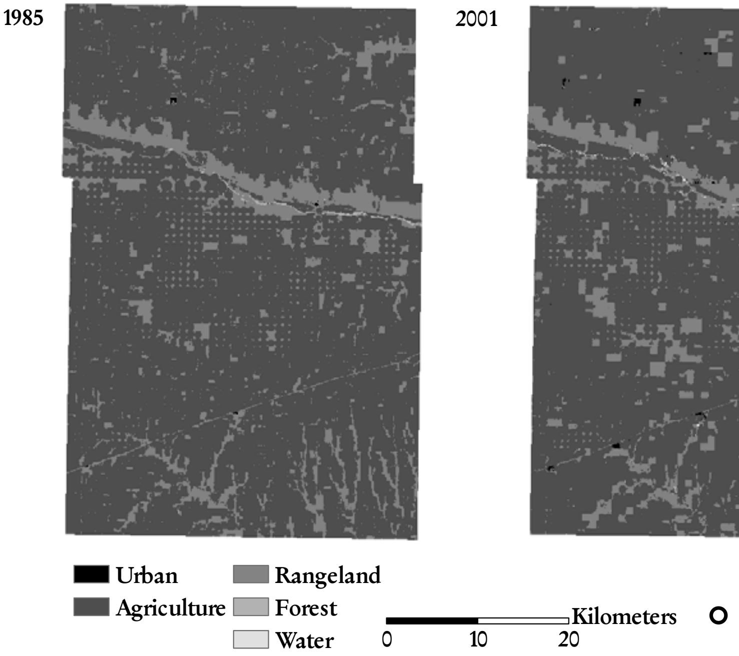


Fig. 4. These maps show the land-use change in the Kansas study area between 1985 and 2001. These maps are classified using Anderson Level I classification. Most of the change is Swap; however, there is considerably less change here than on the same maps classified using Anderson Level II.

1.3 Literature Review

Scale has been identified as a research priority within the field of GIS [11, 12]. The CAP falls under this broad issue and should be a concern for any investigation of categorical information, spatially explicit or otherwise. This problem is related to other issues of scale, including the modifiable areal unit problem and ecological inference problem.

The modifiable areal unit problem (MAUP) has been a concern in geography and related fields for over half a century [2, 14]. A detailed investigation of the problem began during the late 1970s. Openshaw [9] defined two components of the MAUP, the scale problem and the aggregation problem. The aggregation problem realizes there are numerous ways of defining the boundaries of zones, and as a result, differing conclusions are often drawn from these differing delineations [8]. The scale problem, however, is more relevant to the CAP. Openshaw [9] describes the scale problem as "the variation in results that can often be obtained when data for one set of areal units are progressively aggregated into fewer and larger units for analysis" (p. 8). The CAP can be considered a cousin of the MAUP because both problems examine the effect of aggregating from detailed information to generalized information. However, the CAP differs fundamentally from the MAUP because the MAUP modifies the unit of observation, while the CAP modifies the variable definition. Openshaw and Taylor [8] showed that researchers can obtain nearly any measurement of association between the two maps by modifying the unit of analysis. However, this is not the case for the CAP. This paper derives the mathematical principles that constrain the effect of categorical aggregation on statistical results.

Research on the ecological inference problem also has implications for the category aggregation problem. The ecological inference problem, also referred to as the ecological fallacy problem, occurs when conclusions drawn from aggregate data are applied at the individual level [4]. The ecological inference problem acknowledges that the relationships observed for groups do not always hold for individuals. Therefore, conclusions regarding a landscape established through the use of aggregated categorical land-use information may not hold true when the same landscape is analyzed using more detailed categories. In this analogy, we regard the detailed categories on the unaggregated maps as the individuals and the broader categories on the aggregated maps as the groups.

Scientists have made only limited progress on the modifiable areal unit problem and the ecological inference problem [4, 13]. For the CAP, we have been able to distill the fundamental concepts that dictate the effect of category aggregation, which the methods and results sections describe.

2 Methods

2.1 Strategy

This section of the paper demonstrates the category aggregation problem through the aggregation of various combinations of three categories that compose a simple example map. We focus on the effect of a single step of aggregating two of the three categories in order to give insight into more complex multi-category aggregations. Aggregation of several categories can be considered a step-wise process where two categories are aggregated per step. The order in which categories are combined does not affect the analysis of the final aggregated maps. Therefore, the principles gleaned from dissecting a single step of aggregation can be extrapolated to apply to any combination of any number of categories. Below, the mathematics that govern the process of a single step of aggregation are illustrated with simple example maps. These principles are then applied using empirical data in an examination of the three study landscapes mentioned above, located in Massachusetts, Arizona, and Kansas.

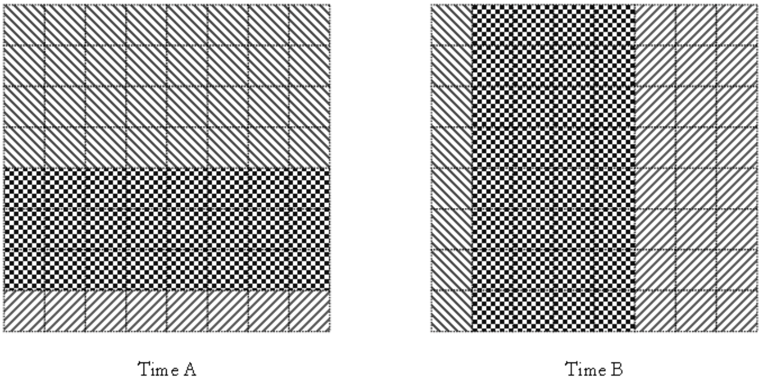
2.2 Examination of Illustrative Examples

Figure 5 shows maps of the example landscape at time A and time B. Through the transition on these unaggregated maps, category 1 loses while categories 2 and 3 gain. Just over 70% of the map undergoes a transition between time A and time B. Net accounts for 37% of the map change while Swap accounts for 33%. There are three possible ways to combine the categories such that the aggregated maps are composed of exactly two categories.

One possible combination is the union of categories 2 and 3. Both these categories experience a Net gain from time A to time B in the pre-aggregated maps, thus the resulting category also shows Net gain. The post-aggregation maps are shown in Figure 6. By combining categories 2 and 3, the amount of Total on the map is reduced from 70% to 50%. The Net is maintained at 37% while the Swap on the maps is reduced to 13%.

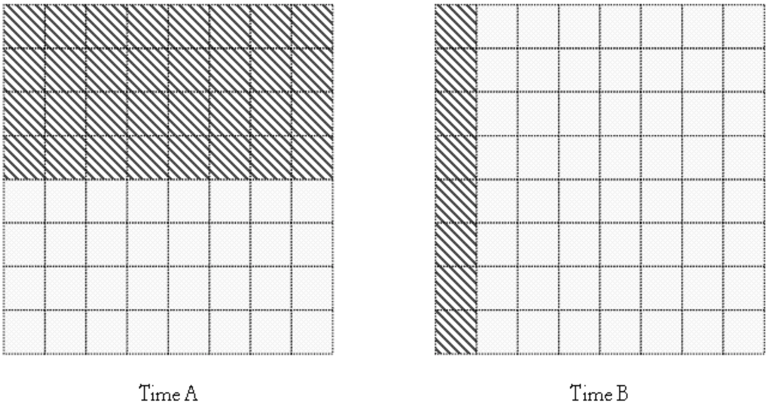
Another possible aggregation is the union of categories 1 and 2, a Net losing and Net gaining category, respectively. The resulting category, $1 \cup 2$, is Net losing (Figure 7). The combination of these categories results in the decrease of the Total observed on the map from 70% to 41%. The Net is reduced to 25% while the Swap on the maps is reduced to 16%.

The final possible aggregation with these maps is the union of categories 1 and 3. Between time A and time B category 1 loses, while category 3 gains. Category $1 \cup 3$ experiences Net loss (Figure 8). The Total on the aggregated maps is reduced to 50% and the Net decreases to 12%, while the Swap actually increases to 38% of the study area.



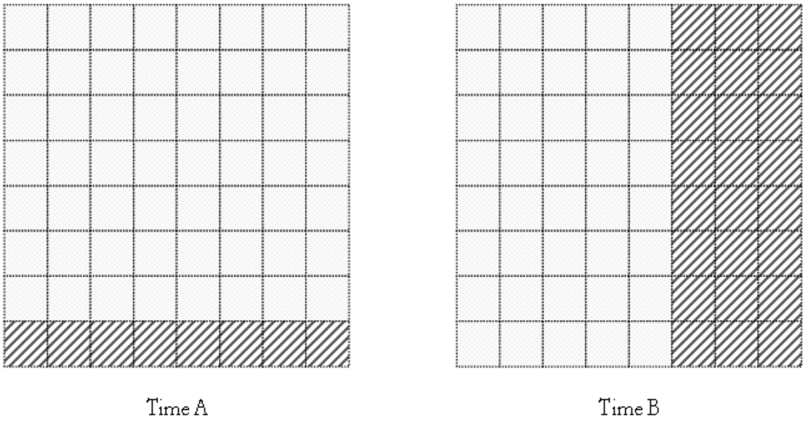
		Loss	Gain	Total	Net	Swap
▧	Category 1	28	4	32	24	8
▩	Category 2	12	20	32	8	24
▨	Category 3	5	21	26	16	10
	Map	45	45	45	24	21

Fig. 5. These example maps illustrate the principles of aggregation. The numbers in this budget and those in Figures 6-8 refer to map pixels, not percentages.



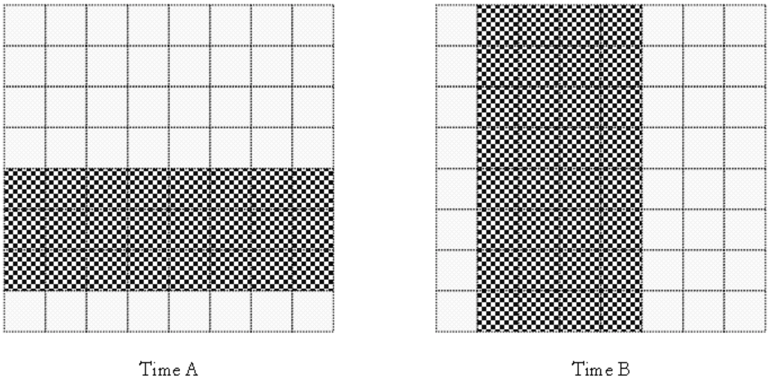
	Loss	Gain	Total	Net	Swap
▨ Category 1	28	4	32	24	8
░ Category 2 U 3	4	28	32	24	8
Map	32	32	32	24	8

Fig. 6. These maps represent the same example landscapes in Figure 5, however the two Net gaining categories (categories 2 and 3) have been aggregated. Net is maintained yet Swap decreases.



	Loss	Gain	Total	Net	Swap
░ Category 1 U 2	21	5	26	16	10
▨ Category 3	5	21	26	16	10
Map	26	26	26	16	10

Fig. 7. These maps represent the same example landscapes in Figure 5, however a Net losing category and a Net gaining category (categories 1 and 2, respectively) have been aggregated. Both Net and Swap decrease.



		Loss	Gain	Total	Net	Swap
□	Category 1 U 3	20	12	32	8	24
▨	Category 2	12	20	32	8	24
	Map	32	32	32	8	24

Fig. 8. These maps represent the same example landscapes in Figure 5, however a Net losing category and a Net gaining category (categories 1 and 3, respectively) have been aggregated. Net decreases and Swap increases.

2.3 Effect of Aggregation on Change Budgets

Table 3 demonstrates how the aggregation of categories affects the amount of change on the maps. The original maps are composed of three categories, as in Table 1; however, categories 2 and 3 are aggregated to form a single category called $2 \cup 3$ in Table 3.

Table 3. Cross-tabulation matrix to examine the effect of aggregating categories 2 and 3.

		Time B			
		Category 1	Category 2 ∪ 3	Total Time A	Loss
Time A	Category 1	P_{11}	$P_{12} + P_{13}$	P_{1+}	$P_{12} + P_{13}$
	Category 2 ∪ 3	$P_{21} + P_{31}$	$P_{22} + P_{23} + P_{33} + P_{32}$	$P_{2+} + P_{3+}$	$P_{21} + P_{31}$
	Total Time B	P_{+1}	$P_{+2} + P_{+3}$	1	
	Gain	$P_{21} + P_{31}$	$P_{12} + P_{13}$		

Comparison of the cross-tabulation matrices associated with the original maps and the subsequent aggregated maps yields substantial information about the effects of the manipulation. The amount of persistence between time A and B on the unaggregated maps (i.e., P_{11} , P_{22} and P_{33}) remains as persistence in the aggregated maps. The amount of the maps that transitions between the two aggregated categories becomes persistence after aggregation. That is, entries P_{23} and P_{32}, which were off-diagonal in the cross-tabulation table for the unaggregated maps, become part of the diagonal, and therefore persistence, in the aggregated maps. This movement to the diagonal eliminates P_{23} and P_{32} from the calculation of Gain and Loss and thereby from the Total and its components of Net and Swap. Thus, Total can only be decreased or maintained through aggregation.

Therefore, the effect of aggregation is more pronounced where there are large transitions between the aggregated categories. If the aggregated categories do not transition (i.e., $P_{23} = P_{32} = 0$), then aggregation has no effect on the Total. The amount of persistence remains the same in the pre and post-aggregation maps for category 1. The amount of category 1 that transitions between time A and B also remains the same. Category 1 can transition to only category $2 \cup 3$ in the post-aggregation maps. Therefore, the sum of P_{12} and P_{13} yields a new entry that represents the amount of category 1 that transitions to the new aggregated category (Table 3). Conversely, entries P_{21} and P_{31} are summed to yield an entry representing the amount of the maps that transitions from the new aggregated category to category 1. Thus, the distribution of the Net and Swap on these aggregated maps is determined by the relative sizes of the Loss and Gain of category 1. Table 4 illustrates the effect of the aggregation on the change budget for a situation where categories 2 and 3 are aggregated.

Table 4. Map change budgets derived from the cross-tabulation matrix of Table 3.

Category	Loss	Gain	Total	Net	Swap
1	$P_{12}+P_{13}$	$P_{21}+P_{31}$	$P_{12}+P_{13}+$ $P_{21}+P_{31}$	$\lvert(P_{12}+P_{13})-(P_{21}+P_{31})\rvert$	$\mathrm{MIN}(P_{12}+P_{13}, P_{21}+P_{31}) * 2$
$2 \cup 3$	$P_{21}+P_{31}$	$P_{12}+P_{13}$	$P_{21}+P_{31}+$ $P_{12}+P_{13}$	$\lvert(P_{21}+P_{31})-(P_{12}+P_{13})\rvert$	$\mathrm{MIN}(P_{21}+P_{31}, P_{12}+P_{13}) * 2$
Map	$P_{12}+P_{13}+$ $P_{21}+P_{31}$	$P_{12}+P_{13}+$ $P_{21}+P_{31}$	$P_{12}+P_{13}+$ $P_{21}+P_{31}$	$\lvert(P_{12}+P_{13})-(P_{21}+P_{31})\rvert$	$\mathrm{MIN}(P_{12}+P_{13}, P_{21}+P_{31})$ + $\mathrm{MIN}(P_{21}+P_{31}, P_{12}+P_{13})$

2.4 Application to Kansas, Massachusetts, and Arizona Sites

To examine the effect of aggregation in practice, we use empirical data of three Human Environment Regional Observatory (HERO) study sites. The HERO Network is funded by the National Science Foundation and is composed of four sites located in Arizona, Kansas, Massachusetts, and Pennsylvania. The network addresses three core research themes: land-cover change, greenhouse-gas emissions, and the impact of these activities on climate. One explicit purpose of the HERO research is to compare land cover across its various sites, therefore we must create datasets that facilitate cross-site comparison, and hence we want to aggregate the available data to a common set of land categories. We investigate the effect of aggregation using data from Grey County in Kansas, central Massachusetts, and the Sonoran region of Arizona and Mexico.

Section 1.2 describes the Kansas data. In the central Massachusetts region, land-cover information was acquired from the Commonwealth [5], which delineates twenty land categories for the ten town region at two times: 1971 and 1999. This information is then aggregated to the Anderson level I classification, leaving seven categories. The southern Arizona data for 1986 and 1999 were classified through remote sensing to Anderson level I; however, we suspect that the method confused two categories, rangeland and barren. In this instance, we combine these two Anderson level I categories to create an aggregated classification with only six categories.

3 Results

3.1 Generalizable Effects of Aggregation

Our investigation reveals that five important principles govern the effects of category aggregation. Each principle is a mathematical fact that applies to any categorical variable and expresses the effect of the aggregation as a function of the transitions among the categories on the unaggregated maps. The post-aggregation Total, Net, and Swap are dependent upon whether the categories being aggregated transition among each other in the unaggregated maps and whether they exhibit Net loss or Net gain when transitioning from time A to time B on the unaggregated maps. All five principles are based on the fact that the Total observed on the maps is equal to the sum of the Net and Swap (equation 2). Thus, the difference in the Total due to aggregation is equal to the sum of the differences in Net and Swap as expressed in equation 3.

$$\Delta \mathrm{Total}_{\mathrm{M}} = \Delta \mathrm{Net}_{\mathrm{M}} + \Delta \mathrm{Swap}_{\mathrm{M}} \tag{3}$$

The first principle applies to all cases and is the foundation for the other principles. Principles 2 and 3 apply to cases where a losing category is aggregated with another losing category or where a gaining category is aggregated with another gaining category. Principle 2 describes the effect of aggregation on Net while principle 3 describes the effect on Swap. Principles 1 and 2 are used to prove principle 3. Alternatively, principles 4 and 5 apply to cases where a losing category is being aggregated with a gaining category. Principle 4 describes the effect of this aggregation on Net while principle 5 describes the effect on Swap. We use principles 1 and 4 to prove principle 5.

3.2 Principle 1

The first principle states "if any categories that transition between each other are aggregated, then the amount of Total on the map is reduced by the amount of their transitions." Mathematically speaking, equation 4 explains the principle where categories i and j are aggregated.

$$\Delta \mathrm{Total}_{\mathrm{M}} = -\left(\mathrm{P}_{\mathrm{ij}} + \mathrm{P}_{\mathrm{ji}}\right) \tag{4}$$

When two categories are combined, the transitions between them move to the diagonal. Thus the aggregation of categories i and j will decrease the amount of Total by P_{ij} + P_{ji} because these values will be converted to persistence and be brought onto the diagonal. Consequently, the aggregation of categories cannot increase the amount of Total observed on the maps. This principle is evidenced in Tables 1 and 3. These tables show the aggregation of categories 2 and 3; as a result of this aggregation, entries P_{23} and P_{32}, which were off the diagonal in Table 1, are moved onto the diagonal in Table 3. The only instance of aggregation in which the Total is not reduced is when categories that do not transition to each other are combined, i.e., when $\mathrm{P}_{\mathrm{ij}} + \mathrm{P}_{\mathrm{ji}} = 0$.

3.3 Principle 2

The second principle states "if either a Net losing category is aggregated with a Net losing category or a Net gaining category is aggregated with a Net gaining category, then the Net on the map is maintained". Mathematically, if ($P_{+i} - P_{i+} \leq 0$ and $P_{+j} - P_{j+} \leq 0$) or ($P_{+i} - P_{i+} \geq 0$ and $P_{+j} - P_{j+} \geq 0$) then equation 5 applies.

$$\Delta \text{Net}_{\text{M}} = 0 \tag{5}$$

This is shown by Tables 2 and 4, where categories 2 and 3 are being aggregated and the resulting Net of category $2 \cup 3$ is equal to the sum of the individual Nets of categories 2 and 3. The proof of this is as follows. In Table 2, the Net for category 2 is equal to $|(P_{21}+P_{23}) - (P_{12}+P_{32})|$ while the Net for category 3 is equal to $|(P_{31} + P_{32}) - (P_{13} + P_{23})|$. If both categories experience a Net loss, then the absolute value signs are irrelevant, thus the Net for the unaggregated categories is equal to $(P_{21}+P_{23}) - (P_{12}+P_{32})$ for category 2 and $(P_{31} + P_{32}) - (P_{13} + P_{23})$ for category 3. The two Net losing categories are added together during the aggregation, so the values for P_{23} and P_{32} cancel. The resulting Net is $(P_{21}+P_{31}) - (P_{12}+P_{13})$, which is equal to $|(P_{21}+P_{31}) - (P_{12}+P_{13})|$, which is the Net for the aggregated category according to Table 4. Alternatively, if the two aggregated categories each demonstrate a Net gain, then a similar situation results. The Net for category 2 is $|(P_{21}+P_{23}) - (P_{12}+P_{32})|$, which would be equivalent to $(P_{12}+P_{32}) - (P_{21}+P_{23})$. Similarly, the Net for category 3 is $|(P_{31} + P_{32}) - (P_{13} + P_{23})|$, which would be equal to $(P_{13} + P_{23}) - (P_{31} + P_{32})$. During the aggregation, again cells P_{32} and P_{23} cancel and the Net for the aggregated categories is $(P_{12}+P_{13}) - (P_{21}+P_{31})$, which is equal to $|(P_{21}+P_{31}) - (P_{12}+P_{13})|$. Therefore, $\text{Net}_{2 \cup 3} = \text{Net}_2 + \text{Net}_3$ so the pre-aggregation Net equals the post-aggregation Net, thus $\Delta \text{Net}_{\text{M}} = 0$.

3.4 Principle 3

The third principle establishes "if either a Net losing category is combined with a Net losing category or a Net gaining category is combined with a Net gaining category, then the Swap on the map decreases by the amount of their transitions". Mathematically, if ($P_{+i} - P_{i+} \leq 0$ and $P_{+j} - P_{j+} \leq 0$) or ($P_{+i} - P_{i+} \geq 0$ and $P_{+j} - P_{j+} \geq 0$)then equation 6 results.

$$\Delta \text{Swap}_{\text{M}} = -\left(P_{ij} + P_{ji}\right) \tag{6}$$

The aggregation of categories i and j decreases the amount of Swap by P_{ij} + P_{ji} when the categories being combined exhibit a similar direction of Net. Equation 6 results when equation 4 and equation 5 are substituted into equation 3. Thus principle 3 is a direct consequence of principles 1 and 2. This third principle applies to all cases where the second principle applies because both principles involve the aggregation of categories that have a similar direction of Net. Principles 2 and 3 apply also to instances where a category that undergoes Net on the unaggregated maps is combined with one that has zero Net, but still transitions with other categories.

3.5 Principle 4

The fourth principle states "if a Net losing category is aggregated with a Net gaining category, then Net on the map decreases by the smaller of the two Nets of the individual categories being aggregated". Mathematically, if $P_{+i} - P_{i+} \leq 0$ and $P_{+j} - P_{j+} \geq 0$, then equation 7 applies.

$$\Delta \mathrm{Net}_M = -\mathrm{MIN}\left(\left|P_{+i} - P_{i+}\right|, \left|P_{+j} - P_{j+}\right| \right) \tag{7}$$

To prove principle 4, we examine a general case where categories 1 and 2 are aggregated to form category $1 \cup 2$. Net for an unaggregated map is calculated in equation 8.

$$\text{Pre -Aggregation Net}_M = \frac{\sum_{j=1}^{J} \left|P_{+j} - P_{j+}\right|}{2} = \frac{\left|P_{+1} - P_{1+}\right| + \left|P_{+2} - P_{2+}\right|}{2} + \frac{\sum_{j=3}^{J} \left|P_{+j} - P_{j+}\right|}{2} \tag{8}$$

To complete the proof, we must show that equation 9 is true. Equation 9 expresses the pre-aggregation Net as the sum of the • Net and the post-aggregation Net. The • Net is located just to the right of the equal sign in equation 7, while the post-aggregation Net constitutes the remainder of the right-hand side of equation 9 in square brackets.

$$\text{Pre-Aggregtion Net}_M = \mathrm{MIN}\left(\left|P_{+1} - P_{1+}\right|, \left|P_{+2} - P_{2+}\right|\right) + \left[\frac{\left|P_{+1\cup 2} - P_{1\cup 2+}\right|}{2} + \frac{\sum_{j=3}^{J} \left|P_{+j} - P_{j+}\right|}{2} \right] \tag{9}$$

Without loss of generality, call the Net losing category 1 and call the Net gaining category 2. There are two cases that need to be considered. The first case is where category 1 Net loses the same as or less than category 2 Net gains, i.e., $\left|P_{+1} - P_{1+}\right| \leq \left|P_{+2} - P_{2+}\right|$. If this case occurs, then we examine the behavior of categories 1 and 2 in equation 8. All of the Net loss of category 1 is paired with a subset of the Net gain of category 2 such that equation 10 holds. Equation 10 can then be substituted into equation 8 to yield equation 9.

$$\frac{\left|P_{+1} - P_{1+}\right| + \left|P_{+2} - P_{2+}\right|}{2} = \frac{2 \times \left|P_{+1} - P_{1+}\right| + \left|P_{+1\cup 2} - P_{1\cup 2+}\right|}{2} = \mathrm{MIN}\left(\left|P_{+1} - P_{1+}\right|, \left|P_{+2} - P_{2+}\right|\right) + \frac{\left|P_{+1\cup 2} - P_{1\cup 2+}\right|}{2} \tag{10}$$

The second case is where category 2 Net gains less than category 1 Net loses, i.e., $\left|P_{+2} - P_{2+}\right| < \left|P_{+1} - P_{1+}\right|$. If this case occurs, again we examine the behavior of categories 1 and 2 in equation 8. All the Net gain of category 2 is paired with a subset of the Net loss of category 1 such that equation 11 holds. Equation 11 can be substituted into equation 8 to yield equation 9.

$$\frac{\left|P_{+1} - P_{1+}\right| + \left|P_{+2} - P_{2+}\right|}{2} = \frac{2 \times \left|P_{+2} - P_{2+}\right| + \left|P_{+1\cup 2} - P_{1\cup 2+}\right|}{2} = \mathrm{MIN}\left(\left|P_{+1} - P_{1+}\right|, \left|P_{+2} - P_{2+}\right|\right) + \frac{\left|P_{+1\cup 2} - P_{1\cup 2+}\right|}{2} \tag{11}$$

3.6 Principle 5

The fifth principle states that “if a Net losing category is aggregated with a Net gaining category, then the Swap on the map can decrease, increase, or be maintained.” Stated mathematically, if $P_{+i} - P_{i+} \leq 0$ and $P_{+j} - P_{j+} \geq 0$ then equation 12 results.

$$\Delta \text{Swap}_M = \text{MIN}\left(\left|P_{+i} - P_{i+}\right|, \left|P_{+j} - P_{j+}\right|\right) - \left(P_{ij} + P_{ji}\right) \tag{12}$$

Equation 12 is the consequence of substituting equation 4 and equation 7 into equation 3. Thus principle 5 is the direct consequence of principles 1 and 4. Principle 1 establishes the decrease in Total and principle 4 establishes the decrease in Net. Equation 3 implies the difference in Swap on the maps is equal to the difference in Total minus the difference in Net. If the decrease in Total is greater than the decrease in Net, then the Swap decreases. If decrease in Net is greater than the decrease of Total, then the Swap increases. If decrease in Total equals the decrease in Net, then the Swap is maintained.

3.7 Empirical Results

Figure 9 shows the effect of aggregation on land-use maps of study areas in Massachusetts, Arizona, and Kansas.

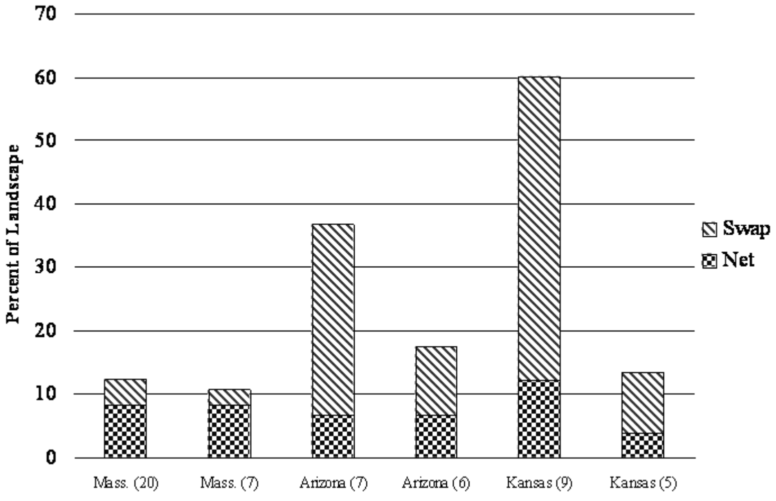


Fig. 9. This figure shows the amount of Total, Net, and Swap for each of the three study areas investigated at two different levels of categorical scale to illustrate the effect of aggregation on the maps. The height of the bar corresponds to the amount of Total on the maps.

In central Massachusetts, the twenty-category, unaggregated maps show 12% Total between 1985 and 1999; 8% is Net while the remaining 4% is Swap. After the aggregation to Anderson Level I, 11% of the maps show change where again 8% is Net and the remainder is attributable to Swap. In the Arizona maps, the rangeland and barren categories are combined, both of which are losing categories. The amount of Total on the mapped landscape decreases from 37% to 17%. The Net on the maps remains unchanged at 7% while the Swap shrinks from 30% to 11%.

4 Discussion

4.1 Interpretation of Kansas, Massachusetts, and Arizona

The empirical data demonstrate how the principles manifest on maps composed of more than three categories (Figure 9). The Arizona case provides the simplest of the empirical cases. The two most dynamic of the original Anderson Level I categories, rangeland and barren, are aggregated. Both are Net losing categories, so principles 1, 2, and 3 apply. There is a substantial amount of Swap in each of the two categories and the two categories transition between each other because of confusion in their classification. When they are aggregated, Total decreases according to principle 1, the Net is maintained according to principle 2, and Swap decreases from 30% to 11% according to principle 3.

The maps of the Kansas study site also show drastic effects of aggregation, where nine categories are aggregated to five. The aggregated categories are a mix of Net losing and Net gaining where there are substantial transitions among the categories. This leads to a dramatic reduction in the Total observed on the maps, from 61% to 13% based on principle 1. The Net shrinks from 13% to 4% by principle 4 and the Swap is reduced from 48% to 10% through principles 3 and 5.

Of the empirical examples, the maps of the central Massachusetts site have the most categories aggregated, from twenty categories to seven. There is a mix of Net gaining categories and Net losing categories; however none of the transitions among the aggregated categories are extremely large so the effect of the aggregation is small. Where the original maps show 12% Total, composed of 8% Net and 4% Swap, the aggregated maps exhibit 11% Total, where Net remains nearly unaltered at 8%. Forest is one of the original twenty categories. It accounts for 8% of the Total on the unaggregated maps; nearly all of this change is Net loss. It is not aggregated with any other categories, so its 8% change is maintained. Most of the other categories gain parts of the map surrendered by forest. The majority of the aggregations in these maps are gaining categories with other gaining categories; therefore little effect is seen on the Net as dictated by principle 2. Swap is not reduced as dramatically as in the other study sites because Swap is already small on the unaggregated maps. Thus, the Massachusetts example demonstrates that the number of categories aggregated is not necessarily important; instead, it is more important how the aggregated categories transition on the unaggregated maps.

4.2 Broader Applications

The category aggregation problem extends well beyond analysis of LUCC. It is encountered commonly in accuracy assessment. Aggregation of land categories can have a tremendous effect on the error matrices used to assess the accuracy of categorical maps. For example, Helmer et al. [3] demonstrate how this problem manifests in the accuracy assessment of a vegetation cover map. Initially, Helmer et al. classified the land cover in Puerto Rico using a 26-category classification. Their map had a classification accuracy of 79%. They reduced the total number of categories in their initial map to 19 through aggregation of categories that were being confused in the classification. As a result, the classification accuracy of their map rose to 83%. The process of aggregating categories during the production of categorical maps, especially those derived from satellite imagery, is extremely common, however it is unusual for scientists to document the effects of such manipulation in the early stages of analysis, as these authors did. Instead, most authors offer only the final classification, which makes readers blind to one of the scientist's most important decisions concerning the production of the maps. This example shows that the CAP is not limited to analysis of change. It demonstrates that researchers should investigate the effects of the problem as it applies to analysis of any categorical data.

The five principles extend to the aggregation of any categorical data because all the principles apply to any cross-tabulation analysis (Tables 1-4). While we examine the effect of this problem in terms of a spatially-explicit variable, i.e., land-use/cover, these principles apply also to the aggregation of non-spatial categorical data such as census data or labor statistics where job category definitions are analyzed by detailed sub-sectors nested within broader economic sectors. Just as the MAUP and ecological inference problem are general and transdisciplinary, the CAP is important for analysis that goes far beyond applications in geography.

5 Conclusions

A decision in the early stages of investigation regarding aggregation of categories can have a profound effect on subsequent analysis and conclusions. This paper distills five mathematical principles that dictate these effects. Principle 1 states that an aggregation can not increase the Total change. The behavior of unaggregated categories on the original maps must be considered to examine the effect of aggregation on Net and Swap. If the aggregation is of either a Net losing category with another Net losing category or a Net gaining category with another Net gaining category, then principles 2 and 3 apply. If a Net losing category is combined with a Net gaining category, then principles 4 and 5 apply. Scientists should consider the influence of category aggregation on the analysis of any categorical variable because the effects can be substantial.

Acknowledgments

The National Science Foundation supported this research through the grant "Infrastructure to Develop a Human-Environment Regional Observatory (HERO) Network" (Award ID BCS 9978052). John Harrington at Kansas State University and Cynthia Sorrenson at the University of Arizona provided land-use/cover information. Sam Ratick provided valuable insights throughout the course of this project. Mang Lung Cheuk contributed to this work through his development of the GLS macro. Clark Labs created the GIS and image processing software, Idrisi$^{®}$, used in this analysis.

References

1. Anderson, J., Hardy, E., Roach, J., and Witmer, R.: A land use and land cover classification system for use with remote sensor data. USGS Professional Paper 964, Souix Falls, SD, (1976)
2. Gehlke, C. and Biehl., K.: Certain effects of grouping on the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association Supplement* **29** (1934) 169-170
3. Helmer, E., Ramos, O., Lopez, T., Quinones, M., and Diaz, W.: Mapping the forest type and land cover of Puerto Rico, a component of the Caribbean biodiversity hotspot. *Caribbean Journal of Science* **38** (2002) 165-183
4. King G.: *A solution to the ecological inference problem.* Princeton University Press, Princeton, NJ (1997)
5. Massachusetts Geographic Information Systems (MASSGIS).: Land Use. (2002) http://www.state.ma.us/mgis/lus.html
6. Meyer, W. and Turner, B. eds.: *Changes in land use and land cover: A global perspective.* Cambridge University Press, Cambridge, (1994)
7. National Research Council (NRC).: *Grand challenges in the environmental sciences.* National Academy of Sciences Press, Washington, DC (2000)
8. Openshaw, S. and Taylor, P.: A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley, N. (ed): *Statistical Applications in the Spatial Sciences.* Pion, London (1979) 127-144
9. Openshaw, S.: *The Modifiable Areal Unit Problem.* GeoBooks, Norwich (1984)
10. Pontius Jr., R., Shusas, E., and McEachern, M.: Detecting important categorical land changes while accounting for persistence. *Agriculture, Ecosystems and Environment* **101** (2004) 251-268
11. Quattrochi, D. and Goodchild, M.: (eds.): *Scale in Remote Sensing and GIS.* CRC Press, Boca Raton, FL (1997)
12. University Consortium for Geographic Information Science (UCGIS): Scale: Research White Paper (1998)
http://www.ucgis.org/priorities/research/research_whi-te/1998%20Papers/scale.html
13. Wong, D. and Amrhein., C.: Research on the MAUP: old wine in a new bottle or real breakthrough? *Geographical Systems* **3** (1996) 73-76.
14. Yule, G. and Kendall, M.: *An Introduction to the Theory of Statistics.* Griffin, London (1950)