

# 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.

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

		Time B			Total Time A	Loss
		Category 1	Category 2	Category 3		
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	$P_{12} + P_{13}$	$P_{21} + P_{31}$	$P_{12} + P_{13} + P_{21} + P_{31}$	$ (P_{12} + P_{13}) - (P_{21} + P_{31}) $	$\text{MIN}(P_{12} + P_{13}, P_{21} + P_{31}) * 2$
2	$P_{21} + P_{23}$	$P_{12} + P_{32}$	$P_{21} + P_{23} + P_{12} + P_{32}$	$ (P_{21} + P_{23}) - (P_{12} + P_{32}) $	$\text{MIN}(P_{21} + P_{23}, P_{12} + P_{32}) * 2$
3	$P_{31} + P_{32}$	$P_{13} + P_{23}$	$P_{31} + P_{32} + P_{13} + P_{23}$	$ (P_{31} + P_{32}) - (P_{13} + P_{23}) $	$\text{MIN}(P_{31} + P_{32}, P_{13} + P_{23}) * 2$
Map	$P_{12} + P_{13} + P_{21} + P_{23} + P_{31} + P_{32}$	$P_{12} + P_{13} + P_{21} + P_{23} + P_{31} + P_{32}$	$P_{12} + P_{13} + P_{21} + P_{23} + P_{31} + P_{32}$	$[ (P_{12} + P_{13}) - (P_{21} + P_{31})  +  (P_{21} + P_{23}) - (P_{12} + P_{32})  +  (P_{31} + P_{32}) - (P_{13} + P_{23}) ] / 2$	$\text{MIN}(P_{12} + P_{13}, P_{21} + P_{31}) + \text{MIN}(P_{21} + P_{23}, P_{12} + P_{32}) + \text{MIN}(P_{31} + P_{32}, P_{13} + P_{23})$

$$\text{Loss}_j + \text{Gain}_j = \text{Total}_j = \text{Net}_j + \text{Swap}_j \quad (1)$$

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.

$$\text{Loss}_M = \text{Gain}_M = \text{Total}_M = \text{Net}_M + \text{Swap}_M \quad (2)$$