DATA QUALITY IMPLICATIONS OF RASTER GENERALIZATION

Howard Veregin and Robert McMaster
Department of Geography
University of Minnesota
267-19th Ave S, Rm 414
Minneapolis MN 55455

ABSTRACT

This study is concerned with the data quality implications of raster generalization. The study focuses specifically on the effects of neighborhood-based generalization (categorical filtering) on thematic accuracy. These effects are examined empirically using raster land cover maps. Accuracy is defined in terms of changes in class membership between original and generalized maps. Results indicate that changes are concentrated in those portions of the map and for those classes that exhibit high levels of spatial variability.

INTRODUCTION

Generalization in a raster environment is fundamentally different from generalization in a vector environment. In a vector environment the spatial and thematic components can be generalized independently, while in a raster environment generalization is almost always accomplished by manipulating the thematic component alone. Raster generalization changes the thematic content of maps and thus has implications for thematic accuracy and data quality in general. This study examines the effects of raster generalization on thematic accuracy for categorical data.

Raster Generalization

Several authors have developed frameworks for classifying raster generalization operators. According to the framework developed by McMaster and Monmonier (1989) the four fundamental operators are structural generalization, numerical generalization, numerical categorization and categorical generalization. Schylberg (1993) adds a set of area-feature operators which perform generalization on raster objects defined as clumps of contiguous cells with the same class.

Three classes of operators apply to categorical data. Local operators work directly on attribute values and ignore neighborhood effects. Neighborhood operators are based on class frequencies within a neighborhood or kernel. Object-based operators are applied to raster objects.

This study focuses specifically on a neighborhood operator known as modal filtering. Filtering reduces high-frequency variation in order to enhance the clarity of presentation. In "simple" modal filtering, a kernel is centered on a cell and the modal class within the kernel is determined. This modal value replaces the class of the cell at the center of the kernel. The process is repeated for every cell in the map, except those along the edges (Fig. 1). Kernel size can vary, with larger kernels producing higher levels of generalization.

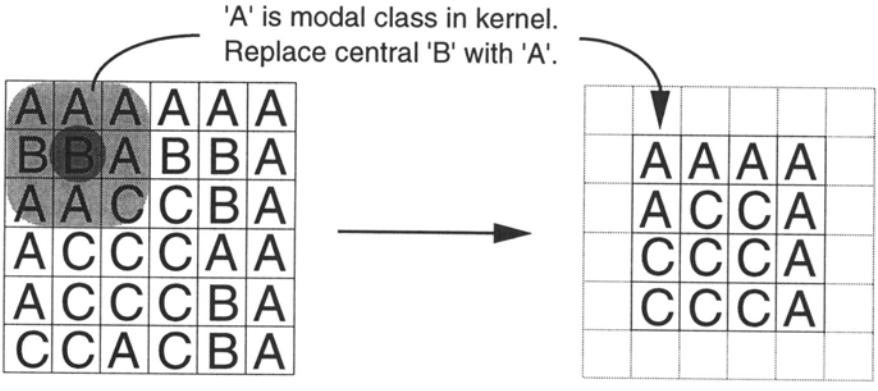

Figure 1. Simple Modal Filtering.

In this example, all classes have equal priority weights. However, unequal weighting is usually required because it is often the case that more than one class has the same frequency in the kernel. When unequal weights are employed, the modal class is defined as the class with the highest weighted frequency. There are several ways to calculate weights. They can be computed based on the area of each class over the entire map. This gives precedence to classes that cover a larger area. Alternatively, weights can be provided by the user. This is useful because it allows for selective enhancement or suppression of certain classes. Finally, weights can be computed by determining class frequencies within a neighborhood outside the kernel. This neighborhood is called a "halo" and its size can vary. A halo "bias factor" can also be defined to give the precedence of frequencies within the halo relative to frequencies within the kernel (Monmonier, 1983).

Whatever the specific method employed, filtering changes the thematic content of the original map by modifying the class memberships of certain cells. These modifications represent a form of thematic error that can be quantified using standard thematic accuracy assessment techniques.

Thematic Accuracy Assessment

Methods of thematic accuracy assessment depend on the measurement scale of the attribute under consideration. For categorical data such as land cover, the most common method is based on the confusion matrix. The matrix, denoted as **C**, has dimensions kxk, where k is the number of classes. Element c_{ij} in the matrix represents the number of cells encoded as class i that actually belong to class j. Correct classifications are those for which i=j. This occurs along the principal diagonal of the matrix. Misclassifications are those for which $i \neq j$. (For a summary of the confusion matrix as applied to classification accuracy assessment in remote sensing see Congalton, 1991).

In the case of modal filtering, an error is defined as a cell with a different class on the original and filtered maps. The confusion matrix tabulates these differences. Element c_{ij} in the matrix represents the number of cells with a class of i on the filtered map and a class of j on the original map.

The information contained in the confusion matrix is typically summarized using indices of thematic accuracy. One such index is PCC, or the proportion of cells that are correctly classified. The maximum value of PCC is 1, which occurs when there is perfect agreement. For modal filtering PCC is defined as the level of agreement between the original and filtered maps. A value close to 1 indicates that the original and filtered maps are nearly identical.

It is usual to distinguish between omission and commission errors in the classification error matrix. An omission error occurs when a cell is omitted from its actual class, i.e., a cell that actually belongs to class j is assigned instead to class i. In the classification error matrix, the off-diagonal elements that occur in a given column j are omission errors in that they represent cells that have been erroneously omitted from class j. Commission error refers to the insertion of a cell into an incorrect class, i.e., a cell is assigned to class i but actually belongs to class j. In the classification error matrix, the off-diagonal elements that occur in a given row i are commission errors in that they represent cases that have been erroneously included in class i.

Any error of omission is simultaneously an error of commission and vice versa. In modal filtering, an error is defined as a cell with a different class on the original and filtered maps. This is an omission error since the cell has been omitted from the class assigned on the original map, and a commission error since the cell has been assigned to a different class than that on the original map.

METHODS

Hypotheses

The effects of filtering on thematic accuracy are hypothesized to be non-uniform spatially and thematically. Filtering reduces high-frequency variation,

such that its effects on accuracy will be most significant in those portions of the map and for those classes that exhibit high-frequency spatial variation.

High-frequency variation is characteristic of classes that are fragmented into small, isolated patches or long, narrow ribbons. These classes tend to lack dominance at the neighborhood level and thus tend not to form the modal class within kernels. These classes will therefore tend to be suppressed by filtering, inducing errors of omission in the filtered map. These effects are reversed for classes that exhibit low-frequency spatial variation, such as those that occur as large, homogeneous clumps. These classes tend to be dominant at the neighborhood level and thus frequently form the modal class within kernels. These classes will therefore tend to be enhanced, resulting in errors of commission in the filtered map.

It is probable that these effects will be non-uniform spatially, since spatial variability is itself variable over space. Changes in class membership will tend to occur in those portions of the map in which variability is highest. These effects will also be affected by kernel size, since the degree of spatial variability is dependent on spatial scale.

Data

Data for this study were derived from aerial video imagery of the Mud Run urban watershed in Akron, Ohio. Imagery was acquired in December, 1994, using a color video camera and was post-processed to extract three spectral bands. Post-processing also included resampling to a 1-meter cell size. Supervised classification was performed using a minimum distance to means classifier (Veregin et al, 1996). The original classified map is shown in Figure 2.

Fig. 2. Original Map.

The area contains a mixture of residential and commercial buildings interspersed with transportation features, grass and bare soil. Areas of deep

shadow are common due to the low sun angle at the time of data collection. (These areas were classified as shadow rather than as their true class due to the limited amount of spectral information that could be extracted from shadow areas.) Different classes exhibit different degrees of spatial variability. For example, grass tends to occur in large homogeneous clumps, while transportation classes (especially concrete) are more linear. Other classes such as roofs and shadows occur as small, isolated clusters.

Methods

The effects of filtering were assessed by comparing the original and filtered maps. To facilitate hypothesis testing, various statistics were computed.

- Confusion Matrix. Element c_{ij} of this matrix represents the number of cells with a class of i on the filtered map and a class of j on the original map.

- Agreement. An overall index of agreement was computed as sum of the diagonal elements of the confusion matrix divided by the number of cells. This is analogous to the PCC index discussed above.

- Omission. An index of omission error was computed for each class j by dividing the diagonal element in column j of the confusion matrix by the column total for j. A higher value for the index indicates less omission error. A value approaching 0 means that almost every cell with that class on the original map has been omitted from this class on the filtered map.

- Commission. An index of commission error was computed for each class i by dividing the diagonal element in row i of the confusion matrix by the row total for i. A higher value for the index indicates less commission error. A value approaching 1 for a given class indicates that almost every cell labeled as that class on the filtered map is that same class on the original map. A value close to 0 indicates that almost every cell labeled as that class on the filtered map is in fact some other class on the original map.

- Change in Area. A simple area change index was computed for each class as the row total divided by the respective column total. A value greater than 1 indicates that the class has more cells on the filtered map than on the original. A value less than 1 indicates the opposite.

- Dissimilarity. A dissimilarity index was computed for each class. This index is a measure of local variability or "texture" for categorical data. Texture can be computed for numerical data as the variance of the cell values in the kernel (Haralick et al, 1973). For categorical data, dissimilarity is defined as the proportion of cells in the kernel that have a class that is different from the class of the cell at the center of the kernel. A higher dissimilarity value means that more variability is present. To maintain consistency in spatial scale, dissimilarity was computed using a kernel of

the same size as that used for generalization. For analysis purposes, mean dissimilarity was computed for each class.

RESULTS

Simple Modal Filtering

The first set of results apply to simple modal filtering using a 3x3 kernel (Fig. 3). Overall agreement between the original and filtered maps is 0.87. As hypothesized, differences between the original and filtered maps are associated with cells having high dissimilarity. Mean dissimilarity is 0.68 for cells that change class and only 0.22 for cells that do not change. Thus, those parts of the map that exhibit thematic error tend to be areas with high spatial variability. This effect is clearly evident in Figures 4 and 5, which show the spatial pattern of dissimilarity and the spatial pattern of error, respectively.

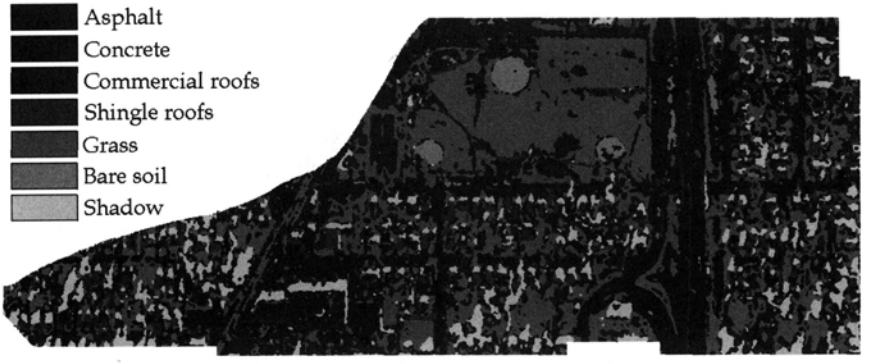

Fig. 3. Filtered Map.

Thematic accuracy statistics for each class are graphed in Figure 6 as a function of mean class dissimilarity. As this figure shows, classes exhibit different levels of dissimilarity. Classes that tend to occur in large, homogeneous clumps (such as grass) have the lowest mean dissimilarity. Classes that tend to occur as isolated patches (such as shingle roofs and commercial roofs) or in long, narrow ribbons (such as concrete) tend to have higher mean dissimilarity values.

As hypothesized, there is a thematic component associated with the effects of generalization. As shown in Figure 6, high dissimilarity is associated with a tendency for classes to be suppressed (area change < 1). Only two classes, grass and asphalt, exhibit growth (area change > 1), and both of these classes have low dissimilarity values. Bare soil and shadow appear to be anomalies, as they have low dissimilarities but tend to be suppressed by the generalization operator.

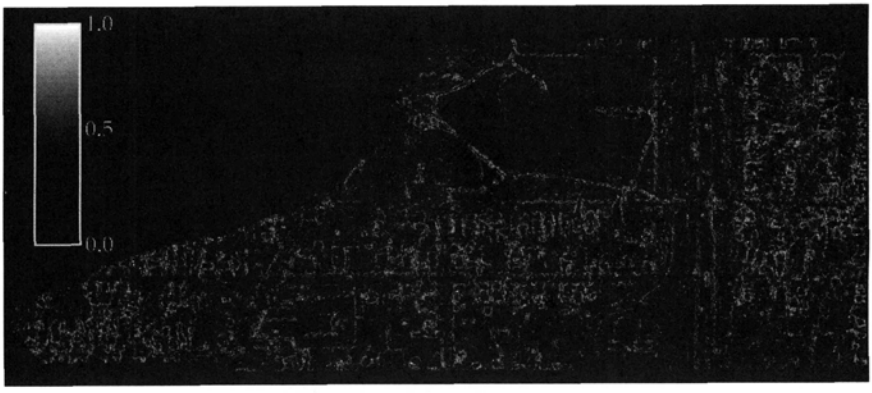

Fig. 4. Spatial Pattern of Dissimilarity Index.

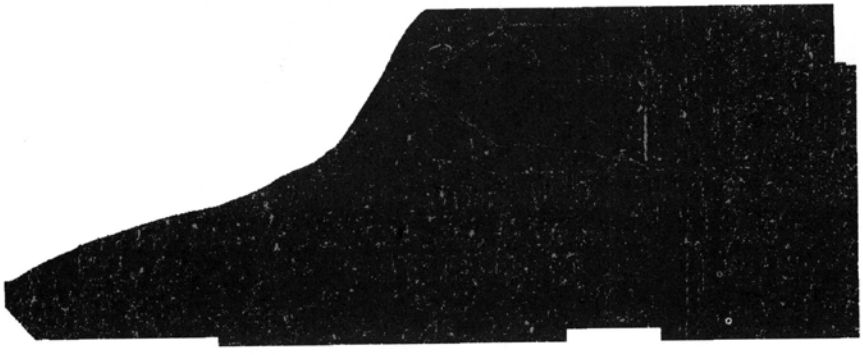

Fig. 5. Spatial Pattern of Thematic Error (Gray areas are cells with different classes on original and filtered maps).

Omission error varies across classes. Classes with high dissimilarity tend to have high levels of omission error (low omission error index). This observation reflects the fact that classes with high dissimilarity tend to be suppressed by classes with low dissimilarity, which are dominant enough to be able to form modal classes. As in the case of area change, bare soil and shadow appear to be anomalies. Figure 6 also shows that omission error and commission error are inversely related. However, there is not a clear relationship between commission error and dissimilarity.

These results support the hypothesis that filtering has the greatest impact on those portions of the map and on those classes that exhibit high-frequency spatial variation. However, mean dissimilarity seems to be an imperfect predictor of this effect. There are several reasons for this.

A high dissimilarity value for a cell means that, within the kernel centered on that cell, a large proportion of the cells are of a different class than the

center cell. However, this does not imply that these neighboring cells are all of the same class, a prerequisite for forming the modal class in the neighborhood. Thus high dissimilarity is not always correlated with a change in class membership.

- Dissimilarity may exhibit significant spatial variations that are masked by the use of mean class values. Dissimilarity for a particular class may depend on proximity to other classes. For example, bare soil might have a high level of dissimilarity when interspersed with grass, but a lower level of dissimilarity when adjacent to transportation features.

- Dissimilarity has no direct implications for commission error. High mean dissimilarity for a class indicates that the class tends to occur in proximity to other classes. This suggests a tendency for classes with high dissimilarity to be suppressed when filtering is performed. However, dissimilarity cannot be used to predict which classes will replace the suppressed classes, since it does not take into account the classes that tend to dominate in the proximity of classes with high dissimilarity values.

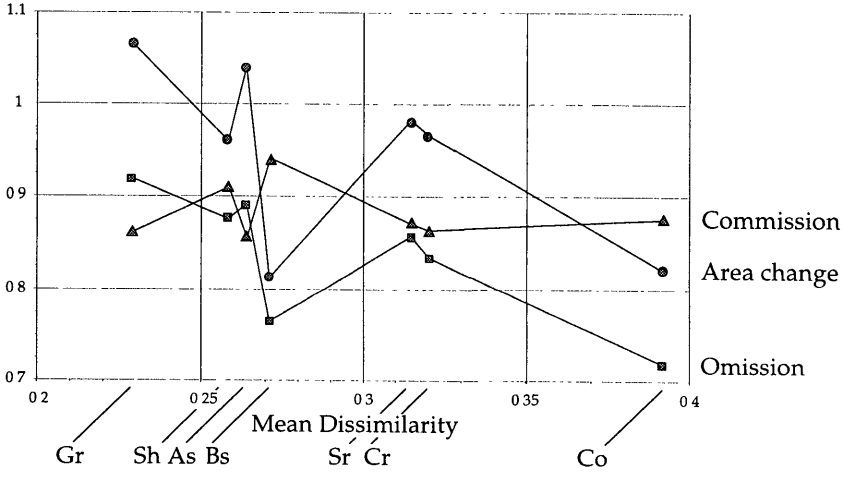

Fig. 6. Statistics for Original and Filtered Maps.

Prediction of commission error requires a measure of the tendency for different classes to exist in the vicinity of each other. One such measure is co-occurrence, which refers to the frequency with which different classes combinations occur. Co-occurrence is computed by counting the number of cells of each class within each kernel location. This yields a co-occurrence matrix, **O**, in which element o_{ij} is the number of cells with a class of i that occur within all kernels centered on cells with a class of j.

In this study, data from the co-occurrence matrix for the original map (Fig. 2) was used to predict the off-diagonal elements in the confusion matrix. The derived regression equation is as follows:

$$c_{ij} = -26.5 + 0.059\ (o_{ij} \times d_j / d_i) \qquad r^2 = 0.96$$

In this equation, c_{ij} the element in row i and column j of the confusion matrix, o_{ij} is the element in row i and column j of the co-occurrence matrix, and d_j and d_i are the mean dissimilarities for classes j and i, respectively. A large ratio of the two dissimilarity values implies that class j is more dissimilar than class i, which means that class i will tend to dominate. This implies that as the ratio increases in value, there is a greater tendency for cells of class j to be assigned a class of i on the generalized map. The high r^2 value for the regression equation indicates that class dissimilarities coupled with co-occurrence data permit reliable prediction of the off-diagonal elements of the confusion matrix.

Other Effects

Results indicate that class suppression and enhancement effects are magnified as kernel size increases. Those classes with the highest dissimilarity are all but eliminated on filtered maps when a large kernel size is used. The effects of filtering are also impacted by the selection of class weights based on the frequencies of class occurrence in a halo surrounding the kernel. Class membership is tabulated in the kernel and separately in the halo. Each of these two vectors of frequencies is then weighted by a bias factor. In this study, it was observed that the use of such weights has essentially the same effect as using a larger kernel. This is simply because the classes that tend to dominate in the halo are the same as those that dominate in large kernels.

Weights can also be defined by the user to selectively enhance or suppress certain classes. Use of these weights has a mitigating effect on the relationships between dissimilarity and thematic accuracy. In general, it is not possible to predict the effects of filtering using dissimilarity if arbitrary weights are employed. However, dissimilarity can be used to select appropriate values for these weights. High mean dissimilarity for a class implies a greater tendency for the class to be suppressed. Hence class weights that are proportional to mean class dissimilarities should ensure that classes are suppressed more evenly.

CONCLUSIONS

The results of this analysis support the hypothesis that modal filtering has the greatest impact on those classes and those parts of the original map where spatial variability is greatest. Thus thematic error introduced by filtering varies over space and theme. To our knowledge this is the first attempt to quantify the effects of raster generalization operators on thematic accuracy. Future work needs to consider the limitations of mean dissimilarity as an index of variability in an effort to enhance understanding of generalization effects and better predict

the degree of thematic error that is introduced. This would facilitate the creation of filtered maps containing low levels of thematic error and minimal omission and commission error for all classes. Future work must also consider local and object-based operators, and should focus attention on issues of visualization of generalization effects. A longer-term goal is to define rules to ascertain the types of generalization that are appropriate in different contexts in order to assure that a minimum threshold of accuracy is maintained.

ACKNOWLEDGMENTS

We wish to acknowledge the assistance of Larry Davis and Jim Jenkins in the creation of the original classified map.

REFERENCES

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment, 37*: 35-46.

Haralick, R.M., K.S. Shanmugam & I. Dinstein (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics, SMC-3*: 61-622.

McMaster, R.B. & M. Monmonier (1989). A conceptual framework for quantitative and qualitative raster-mode generalization. *GIS/LIS '89*: 390-403.

Monmonier, M. (1983). Raster-mode area generalization for land use and land cover maps. *Cartographica, 20*(4): 65-91.

Schylberg, L. (1993). *Computational Methods for Generalization of Cartographic Data in a Raster Environment.* Doctoral thesis, Royal Institute of Technology, Department of Geodesy and Photogrammetry, Stockholm, Sweden.

Veregin, H., P. Sincak, K. Gregory & L. Davis (1996). Integration of high-resolution video imagery and urban stormwater runoff modeling. *Proceedings, Fifteenth Biennial Workshop on Videography and Color Photography in Resource Assessment*, pp. 182-191.