Controlling Data Uncertainty via Aggregation in Remotely Sensed Data

Yohay Carmel

***Abstract*—Aggregation may be used as a means of enhancing remotely sensed data accuracy, but there is a tradeoff between loss of information and gain in accuracy. Thus, the choice of the proper cell size for aggregation is important. This letter explores the change in data accuracy that accompanies aggregation and finds an increase in image thematic accuracy with increasing cell size, resulting from 1) reduction in the impact of misregistration on thematic error and 2) mutual cancelation of inverse classification errors occurring within the same cell. A model is developed to quantify these phenomena. The model is exemplified using a vegetation map derived from an aerial photo. The model revealed a major reduction in effective location error for cell sizes in the range of 3–10 times the size of mean location error; reduction in effective classification error was minor.**

***Index Terms*—Aggregation, classification error, error analysis, misregistration.**

I. Introduction

In RASTER DATA, aggregation (sometimes referred to as image degradation) is a process of laying a grid of cells on the image (cell size $>$ pixel size), and defining the larger cells as the basic units of the new image. When pixels are aggregated into larger grid cells, the information on pixel-specific location is lost. However, the attribute information is retained and can be used to estimate cell composition (e.g., cell-specific percentage cover for each class) [1].

Several studies have suggested that reduction of spatial resolution enhances data accuracy significantly [2]–[5]. On the other hand, the decrease in spatial resolution involves a loss of information that may be valuable for particular applications [6]. Thus, users could benefit from viewing the actual plot of data accuracy as a function of spatial resolution and could choose the specific spatial resolution and its associated uncertainty level that best suits a specific application. The goal of this letter is to explore the relationship between spatial resolution and data accuracy and to develop a model that quantifies this relationship for thematic ("classified") images.

Image thematic error stems from two sources: classification error and location error. The latter component refers to the impact of misregistration on thematic error; it becomes relevant in change detection analyses and in any multilayer GIS analysis. Thus, two types of gain in accuracy are expected when spatial resolution is degraded and cells are aggregated: a gain from reducing the impact of location accuracy on overall thematic accuracy and a gain from canceling out some inverse misclassifications within each cell.

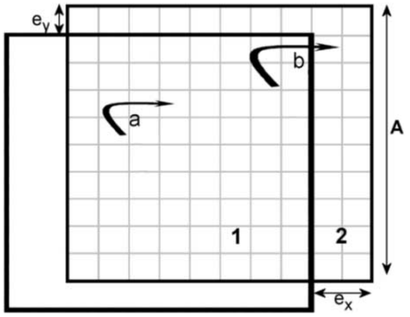

Fig. 1. Location error effect on attribute accuracy. In this example, all pixels within a grid cell are shifted identically, where e_x and e_y denote the x and y components of location error, respectively. Pixels in region 1 would remain within the cell, and attribute accuracy at the cell level would not be affected. Pixels in region 2 would be shifted into neighboring cells and may result in attribute error. The effective location error α is the proportional area of region 2 in the cell (3).

II. Model

A. Location Accuracy

For a single pixel, misregistration is translated into thematic error if its "true" location is occupied by a pixel belonging to a different class. Let us define the probability that a pixel is assigned an incorrect class due to misregistration $p(\mathrm{loc})$

$$p(\mathrm{loc})_{rc} = p(i_{rc} \neq i_{r+e(x),c+e(y)}) \qquad (1)$$

where r and c are pixel coordinates, i_{rc} is the class assigned to the pixel, $e(x)$ and $e(y)$ are the x and y components of cell-specific location error, respectively. Thus, $p(\mathrm{loc})_{rc}$ depends on the magnitude of location error and on image heterogeneity. $p(\mathrm{loc})$ can be estimated empirically for a given image, based on image pattern and the magnitude of location error.

Considering a larger cell size A, let us define a similar probability, $p^A(\mathrm{loc})$, which is the probability that a pixel within the framework of a larger cell was misclassified due to misregistration. For cell sizes larger than location error, this probability would be lower than the original probability $p(\mathrm{loc})$, since misregistration would shift a certain proportion of the pixels only within the grid cell, and for those pixels, thematic error is cancelled at the grid cell level (Fig. 1). This probability is denoted by

$$p^A(\mathrm{loc}) = \alpha \cdot p(\mathrm{loc}) \qquad (2)$$

where α is the proportion of a cell of size A in which location error actually transgresses into neighboring cells and may, thus, result in attribute error (Fig. 1). This proportion (α) is termed

Manuscript received March 2, 2003; revised December 15, 2003. This work was supported by The International Arid Land Consortium (IALC) under Project 02R17 through the University of Illinois.

The author is with Technion—Israel Institute of Technology, Haifa, 32000 Israel (e-mail: yohay@tx.technion.ac.il).

Digital Object Identifier 10.1109/LGRS.2004.823453

1545-598X/04$20.00 © 2004 IEEE

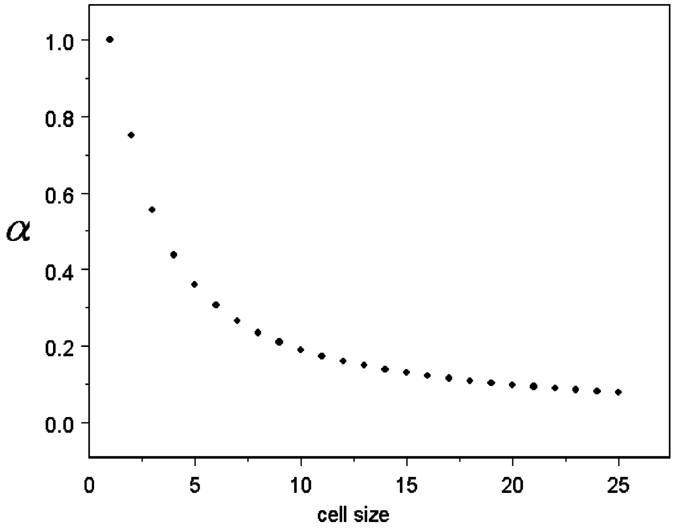

Fig. 2. Effective location error α as a function of cell size. Average location error is set to 1 unit in both x and y directions. Cell size varies between 1 and 25 times the magnitude of location error.

here the effective location error. It is a function of cell size A and of location error magnitude. The effective location error α is the proportional area of region 2 in the cell (see Fig. 1) and is denoted by

$$\alpha = (A \cdot e(x) + A \cdot e(y) - e(x) \cdot e(y))/A^2 \tag{3}$$

where location error components $e(x)$ and $e(y)$ are assumed constant for all pixels within a single cell. Using (3), the reduction in effective location error when cell size increases may be illustrated easily for the special case where $e(x) = e(y)$ (Fig. 2). Effective location error α declines rapidly from 1 for cell sizes less than or equal to the magnitude of location error, to 0.36 for cell sizes five times the magnitude of location error (Fig. 2).

Location error may vary largely across an image. Thus, $\alpha, p(\text{loc})$, and $p^A(\text{loc})$ should ideally be estimated for each grid cell in the image. This requires that location-error components $e(x)$ and $e(y)$ are available at the grid-cell level. Typically, location error information is available for several locations (test points) only. Interpolation methods such as kriging may be used to construct location error surfaces for both $e(x)$ and $e(y)$ [7]. The practice of shifting the image against itself [2], [3], [8] may be used in order to derive cell-specific $p(\text{loc})$. Following (1), a specific cell i_{rc} is shifted by $e(x)_{rc}$ and $e(y)_{rc}$ on the x and y axes, respectively. $p(\text{loc})$ is estimated as the proportion of pixels that are "misclassified" due to the shift. Equations (2) and (3) may then be used to derive cell-specific α and $p^A(\text{loc})$. This process is repeated for each cell in the image, and the global mean of these parameters can then be calculated.

An alternative to this process, which may be less accurate but much simpler to apply, is to assume a constant error across the image. The average location error is typically defined as root-mean-square error (RMSE), decomposed here into its x and y components (only RMSE(x) is presented)

$$\text{RMSE}(x) = \sqrt{\frac{\sum_{g=1}^{n} e(x)_g^2}{n}} \tag{4}$$

where $e(x)_g$ is the x component of the deviation between the true location of a test point g and its location on the image, and n is the number of test points. In order to determine $p(\text{loc})$, the whole image is shifted by RMSE(x) and RMSE(y) on the x and y axes, respectively. $p(\text{loc})$ is estimated for the entire image as the proportion of pixels that are "misclassified" due to the shift. α, and $p^A(\text{loc})$ can then be calculated for the entire image, using (2) and (3).

B. Classification Accuracy

The probability that a pixel is misclassified [$p(\text{cls})$] may be estimated as the proportion of misclassified pixels in the image. The probability that class i pixel is assigned to class j due to misclassification $p(\text{cls})_{ij}$ can be estimated from the error matrix as

$$p(\text{cls})_{ij} = n_{ij}/n_i \tag{5}$$

where n_i is the number of class i pixels in a cell and n_{ij} is the number of class i pixels misclassified as j in that cell. This simple method to calculate $p(\text{cls})_{ij}$ follows the common practice in classification accuracy assessments and ignores the heterogeneous nature of classification error (e.g., error is more likely to occur near edges between classes). An alternative to this method was recently suggested [9], [10], where kriging is used to construct classification error surface. The same method may be used here to estimate cell-specific $p(\text{cls})_{ij}$.

Considered within the framework of grid cells, classification error may be reduced when cell size increases. Consider the case of two pixels within the same grid cell, class i pixel misclassified as j and class j pixel misclassified as i. At the cell level, where pixel information is reduced to proportion cover of each class in the cell, both errors cancel out each other. The probability for a pixel to be misclassified within the frame of a larger cell A, $p^A(\text{cls})$, can be calculated as

$$p^A(\text{cls}) = \beta \cdot p(\text{cls}) \tag{6}$$

where β is the proportion of misclassified pixels in the cell that were not canceled out at the grid-cell level. β is dependent on cell size and on the spatial pattern of the image (since it is a function of the number of pixels of each class in each grid cell). Thus, β should be estimated for each classification error pair ij separately. β_{ij} is dependent on the number of both ij and ji misclassification types. The abundance of these misclassifications is denoted by n_{ij} and n_{ji}, respectively. In what follows, β, n_i, and n_{ij} are cell-specific. Consider a cell that contains many class i pixels and many class j pixels. It is expected that some misclassified ij pixels (n_{ij}), as well as several ji misclassified pixels (n_{ji}), will be present in that cell. n_{ij} is calculated locally as the product of the number of class i pixels in the cell n_i and $p(\text{cls})_{ij}$

$$n_{ij} = n_i \cdot p(\text{cls})_{ij}. \tag{7}$$

In order to calculate β_{ij}, we need to know the spatial relationship between ij and ji misclassified pixels in each grid cell. If $n_{ij} < n_{ji}$, then all ij misclassifications are canceled, and an equal quantity of ji misclassified pixels is canceled as well. In that case, the effective ij misclassification rate β_{ij} is 0, and the

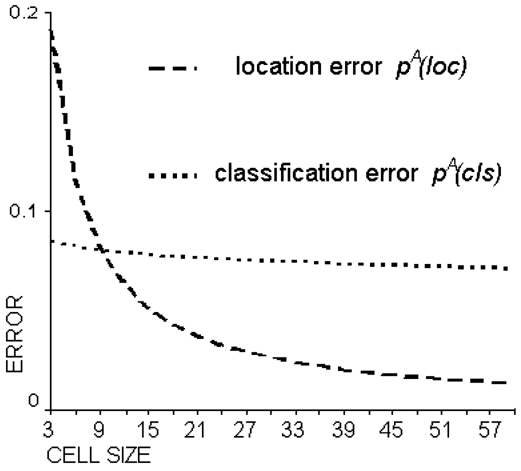

Fig. 3. Cell-level error probabilities $p^A(\mathrm{loc})$ for location error and $p^A(\mathrm{cls})$ for classification error, as a function of cell size. Cell size is in meters.

effective ji misclassification rate β_{ji} is $(n_{ji} - n_{ij})/n_{ji}$. Thus, β_{ij} is denoted by a conditioned term as follows:

$$\begin{cases} \beta_{ij} = 0, & \text{if } n_{ij} \leq n_{ji} \\ \beta_{ij} = \frac{n_{ij} - n_{ji}}{n_{ij}}, & \text{if } n_{ij} > n_{ji}. \end{cases} \tag{8}$$

Using (4), (6), and (7), β_{ij} can be calculated for each aggregated cell in the image. Next, β can be determined as the weighted average of all β_{ij}

$$\beta = \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\beta_{ij} \cdot \frac{n_{ij}^A}{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}^A} \right) \quad (i \neq j). \tag{9}$$

Average β can be calculated for the whole image, for a range of or cell sizes, and the reduction in effective classification error that accompanies the aggregation process can be illustrated.

C. Model Application

The process is exemplified using a vegetation map derived from a 1995 aerial photo of Carmel Valley, CA, for which extensive information on both types of error is available [3]. In the original image, pixel size is 0.6 m. Location error components RMSE(x) *and* RMSE(y) *are* 1.86 and 1.68 m, respectively, and the proportion classified correctly *PCC* is 0.91. Here, model parameters were estimated assuming error homogeneity for both error types.

The probabilities of error at the pixel level, derived from the error matrices constructed for both location error and classification error (data from Carmel *et al.* [6, Table 4]) were $p(\mathrm{loc}) = 0.23$ and $p(\mathrm{cls}) = 0.09$. When estimated for a range of cell sizes, α decreased notably from 0.83–0.07, when cell size changed from 3–60 m. For the same range, β decreased moderately from 0.95–0.8. Accordingly, $p^A(\mathrm{loc})$ diminished from 0.19 to ~0.01 in the same range, while the decrease in $p^A(\mathrm{cls})$ was negligible (Fig. 3).

III. Discussion

Several studies have noted the large impact of misregistration on data accuracy [2], [3], [11]. Moreover, Carmel *et al.* [6] found that its contribution to overall thematic error is larger than that of classification error. Thus, estimating the effective location error α would yield a crude approximation of the impact of aggregation on thematic accuracy. This procedure is simple (especially if RMSE is taken to represent e): solve (3) for a range of relevant cell values, and portray α as a function of cell size (Fig. 2).

The impact of aggregation on classification accuracy can be viewed by drawing β, the effective classification error, as a function of cell size. Estimating β is more complex, while this letter finds that the impact of aggregation on classification error is much smaller than that of location error. In highly fragmented images, β may be more prominent.

Further information can be gained by estimating the actual probabilities of error, $p^A(\mathrm{loc})$ and $p^A(\mathrm{cls})$, for various aggregation levels. However, this stage requires additional calculations and spatially explicit simulations that manipulate the actual image.

IV. Conclusion

The methodology developed here provides an effective tool for assessing the impact of aggregation on thematic accuracy and evaluating it against information loss, in order to decide on a proper level of image aggregation. Current results show that the most effective reduction in error is achieved when cell size is in the range of 3–10 times the size of average location error, but image-specific error rates may somewhat alter this conclusion.

Acknowledgment

The author is grateful for the comments of T. Ellsworth, O. Lahav, R. Adler, and three anonymous reviewers. In particular, thanks are due to reviewer #1, whose comments on an error in his equations induced notable improvements. A. Bar Massada helped with model verification.

References

[1] Y. Carmel and R. Kadmon, “Grazing, topography, and long-term vegetation changes in a Mediterranean ecosystem,” *Plant Ecol.*, vol. 145, pp. 239–250, 1999.

[2] J. R. G. Townshend, C. O. Justice, C. Gurney, and J. McManus, “The impact of misregistration on change detection,” *IEEE Trans. Geosci. Remote Sensing*, vol. 30, pp. 1054–1060, Sept. 1992.

[3] X. L. Dai and S. Khorram, “The effects of image misregistration on the accuracy of remotely sensed change detection,” *IEEE Trans. Geosci. Remote Sensing*, vol. 36, pp. 1566–1577, Sept. 1998.

[4] A. J. J. Van Rompaey, G. Govers, and M. Baudet, “A strategy for controlling error of distributed environmental models by aggregation,” *Int. J. Geograph. Inform. Sci.*, vol. 13, pp. 577–590, 1999.

[5] P. Fieguth, W. Karl, A. Willsky, and C. Wunsch, “Multiresolution optimal interpolation and statistical analysis of TOPEX/POSEIDON satellite altimetry,” *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 280–292, Mar. 1995.

[6] Y. Carmel, D. J. Dean, and C. H. Flather, “Combining location and classification error sources for estimating multi-temporal database accuracy,” *Photogramm. Eng. Remote Sens.*, vol. 67, pp. 865–872, 2001.

[7] P. Fisher, “Improved modeling of elevation error with geostatistics,” *GeoInformatica*, vol. 2, pp. 215–233, 1998.

[8] D. A. Stow, “The role of GIS for landscape ecological studies,” in *Landscape Ecology and GIS*, S. Cousins, Ed. London, U.K.: Taylor & Francis, 1993, pp. 11–22.

[9] A. Cherrill and C. McClean, “An investigation of uncertainty in field habitat mapping and the implications for detecting land cover change,” *Landscape Ecol.*, vol. 10, pp. 5–21, 1995.

[10] P. Kyriakidis and J. Dungan, “A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions,” *Environ. Ecol. Statist.*, vol. 8, pp. 311–330, 2001.

[11] D. A. Stow, “Reducing the effects of misregistration on pixel-level change detection,” *Int. J. Remote Sens.*, vol. 20, pp. 2477–2483, 1999.