

1-1-1981

# Sensitivity of Geographic Information System Outputs to Errors in Remotely Sensed Data

H. K. Ramapriyan

R. K. Boyd

F. J. Gunther

Y. C. Lu

Follow this and additional works at: [http://docs.lib.purdue.edu/lars\\_symp](http://docs.lib.purdue.edu/lars_symp)

---

Ramapriyan, H. K.; Boyd, R. K.; Gunther, F. J.; and Lu, Y. C., "Sensitivity of Geographic Information System Outputs to Errors in Remotely Sensed Data" (1981). *LARS Symposia*. Paper 469.  
[http://docs.lib.purdue.edu/lars\\_symp/469](http://docs.lib.purdue.edu/lars_symp/469)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact [epubs@purdue.edu](mailto:epubs@purdue.edu) for additional information.

Reprinted from

**Seventh International Symposium**

**Machine Processing of**

**Remotely Sensed Data**

with special emphasis on

**Range, Forest and Wetlands Assessment**

**June 23 - 26, 1981**

**Proceedings**

Purdue University  
The Laboratory for Applications of Remote Sensing  
West Lafayette, Indiana 47907 USA

Copyright © 1981

by Purdue Research Foundation, West Lafayette, Indiana 47907. All Rights Reserved.

This paper is provided for personal educational use only,  
under permission from Purdue Research Foundation.

Purdue Research Foundation

# SENSITIVITY OF GEOGRAPHIC INFORMATION SYSTEM OUTPUTS TO ERRORS IN REMOTELY SENSED DATA

H.K. RAMAPRIYAN

NASA/Goddard Space Flight Center  
Greenbelt, Maryland

R.K. BOYD, F.J. GUNTHER, Y.C. LU

Computer Sciences Corporation  
Silver Spring, Maryland

## I. ABSTRACT

The purpose of this paper is to analyze the sensitivity of Geographic Information System outputs to errors in inputs derived from Remotely Sensed Data (RSD). The attention is restricted to outputs of suitability models with "per cell" decisions with gridded Geographic Data Bases (GDB) whose cells are larger than the RSD pixels. The procedure for merging RSD into such GDB's involves classification, registration and aggregation. The first two steps introduce errors at individual pixels and the last step tends to compensate for such errors.

The classification and registration errors are treated independently for the purposes of analysis. Under certain simplifying assumptions, the probability of misaggregation (that is, wrongly assigning a cell after aggregation) is expressed in terms of the probability of misclassification. A Monte Carlo simulation has been performed to show the effects of misregistration on the cell assignments.

Experiments were performed with a data base covering the Harrisburg, Pennsylvania, area. Landsat data covering the same area were classified and registered to the data base.

A baseline data set was prepared as accurately as possible. Perturbations were introduced in the form of (i) classification errors at locations of low confidence in the multispectral classification and (ii) registration errors by selection of subsets of ground control points from those used for the baseline. The errors before and after aggregation and after using the aggregated data in a suitability model were determined using pixel by pixel comparison. For this experiment, combinations of the classi-

fication and registration errors were also used.

It is found that approximately 50% reduction in error occurs due to aggregation when 25 pixels of RSD are used per cell in the GDB. Further reductions in error occur during the modelling process depending on the percentage of the total number of cells affected by RSD.

## II. INTRODUCTION

Geographic Information Systems (GIS) have become increasingly popular for regional planning applications in recent years. Several GIS's in use by various states in this country are listed in a survey by the National Council of State Legislatures (NCSL)<sup>1</sup>. Typical applications of these GIS's are: evaluation of suitability of land for various kinds of development, analysis of erosion potential, inventory of power plant sites, assessment of nonpoint pollution sources and determination of best corridors for highway construction. Each such application has a model associated with it. The model is applied to the data in a Geographic Data Base (GDB) to derive maps to aid in making planning decisions. Clearly, the correctness of the output maps and the consequent decisions will depend on the accuracy of the data in the GDB.

GDBs store various inputs in a common coordinate system at a common resolution. Two commonly used data structures are the "raster" (or "grid") and the "polygon" formats. Raster format requires that each data element be assigned a constant value over a rectangular region, all such regions having the same dimensions. Polygon format assigns a constant value over a polygon instead, where the "polygons" arise from the actual boundaries between the various information or mapping units.

Accordingly no two polygons are likely to have the same size or shape.

Inputs to GDB's come from a variety of sources. Traditionally these have primarily been either maps or photography. With the advent of remotely sensed data (RSD) such as Landsat, Heat Capacity Mapping Mission (HCMM), Seasat, and others a new source of relatively inexpensive up-to-date information now exists.

Typically RSD are geometrically rather imprecise in their raw form. In addition, the data must be converted from number sequences into information-bearing categories before they can be utilized in a GDB. Depending on the time and care taken in performing geometric correction and thematic classification, varying degrees of error will remain and be transmitted into the data base, potentially affecting the results of modelling operations carried out on the data base.

In general, RSD are more easily incorporated into gridded GDBs. Gridded GDBs typically have cell sizes that range from 70 meters to 1 kilometer square<sup>2</sup>. RSD from a given sensor system on the other hand has a specific pixel size. Landsat, for example, produces pixels that are roughly 57 by 79 meters in size. Whenever the cell size of the GDB is larger than that of the RSD the RSD must be aggregated to the larger size. Depending on the degree of aggregation lesser or greater impact on the input data will occur. At least one previous study<sup>3</sup> has dealt with the issue of the impact of aggregation on the information content of the input pixels.

The purpose of this paper is to examine the effects of procedures specific to merging RSD into a GDB on the errors in GIS outputs. Throughout the paper the discussion is couched in the context of an actual GIS application. In that regard the impact of certain modelling operations carried out with an actual GDB was also measured experimentally.

### III. THEORETICAL ANALYSIS

The purpose of this section is to develop relationships between the errors in the input remotely sensed data to a GDB and the output of a suitability model applied to the GDB. A suitability model generates a binary suitability map at the resolution of the GDB. It shows whether a particular cell is suitable for a given purpose or not. The decision on each cell, in

general, involves several data planes of which RS data is one.

Let  $X=(x_1, x_2, \dots, x_n)$  be the "feature vector" characterizing a cell in the GDB. Let  $x_1$  correspond to RS data and the others to data from other sources. (We shall use the term feature vector here in its general sense).

The element  $x_1$  could then be the spectral class number in a classification map (if the GDB resolution is the same as that of RSD) or an informational class number derived from the spectral class number(s) of one or more pixels.

Consider a general "per cell" suitability model (i.e., where the decision on a cell depends only on the feature vector for that cell). It can be written as:

$$S=\{f(x)\in C\}$$

That is, the cell is suitable if and only if the function  $f(x)$  of the feature vector is in class  $C$ .

Class numbers can only be combined logically with other data planes to yield meaningful models. Therefore, consider models of the form

$$S=\{x_1 \in C_1\} \wedge \{f(x_2, \dots, x_n) \in C_2\} \quad (2)$$

$$S=\{x_1 \in C_1\} \vee \{f(x_2, \dots, x_n) \in C_2\} \quad (3)$$

Where  $\wedge$  and  $\vee$  mean "and" and "or", respectively.

For example, in the model: "A given cell is suitable for farming if it has a given set of soil types, slope in a given range and soil depth in a given range, is not identified as a historic site and has no urban or water land cover" all features except the land cover are derived from ancillary data planes and landcover is derived from RSD. The constraints on the non-RSD features can be combined into the form  $f(x_2, \dots, x_n) \in C_2$ .

Now, in the "and" model of equation (2), any cell deemed unsuitable by the non-RSD constraints is unaffected by RSD. Similarly, in the model in equation (3) any cell deemed suitable by non-RSD is unaffected by RSD.

Therefore, the actual percentage of cells yielding wrong decisions from the modelling process as a result of errors in the RSD-developed input depends on

- (i) proportion of cells satisfying [not satisfying] non-RSD constraints in model (2) [(3)]

- (ii) cells where the satisfaction of the RSD constraint ( $x_1 \in C_1$ ) is affected by the errors in RSD-developed input.

In fact,  $P$ , the proportion of cells which have erroneous decisions due to RSD can be written as

$$P = P_1 P_2$$

Where  $P_1$  = Proportion of cells affected by RSD

$P_2$  = Proportion of those cells where RSD is in error.

We shall restrict our attention to cells satisfying [not satisfying] the non-RSD constraints with models (2) [(3)]. Then the event  $E$  that "the cell produces a decision error" can be written as

$$E = \{x \in C, y \notin C\} \vee \{x \notin C, y \in C\} \quad (4)$$

Where we have dropped subscript 1 for convenience,  $x$  is the RSD-derived input and  $y$  is the true value of  $x$  (i.e., what  $x$  would have been, had it been derived from ground truth). The error event  $E$  is caused by various steps involved in the generation of  $x$ .

These steps are listed below:

- (i) Collection and transmission of data
- (ii) Radiometric preprocessing
- (iii) Pixel classification into spectral classes
- (iv) Geometric correction to the GDB's coordinate space
- (v) Merging spectral classes into informational classes
- (vi) Aggregation of informational classes from several pixels to derive  $x$ .

The user of RSD usually receives the data after steps (i) and (ii). The steps (iii) through (vi) do not necessarily have to be performed in that order, but do represent a typical analysis sequence. Steps (iii) and (iv) are prone to errors, but the user has some control over them. The effects of steps (iii) and (iv) are, in general, correlated. However, to simplify the analysis we shall treat them independently. (Some comments will be made about their joint effects in connection with the experimental results).

The steps (v) and (vi) tend to compensate for errors from (iii) and (iv). The following two subsections will demonstrate the relations among steps (iii), (v), (vi) and (iv), (v), (vi).

We shall first introduce the following definitions. The coordinate system relative to which the remotely sensed input image lines and samples are defined will be called the I-space (or image space). The coordinate system of the GDB will be called the G-space. The image pixel sizes are determined by the resolutions in the line and sample directions (approximately 79m and 57m for Landsat MSS). The geometric correction process (step (iv)) involves a mapping from the I-space to the G-space. We shall refer to the pixels resulting from such a mapping as G-space pixels. Nearest neighbor resampling is assumed in the generation of G-space pixels. It is common to generate square pixels in G-space with area comparable (or equal) to that of pixels in I-space (e.g., resolution in both line and sample directions equal to  $\sqrt{79 \times 57} = 67$  meters). The aggregation in step (vi) will be assumed to be the combination of information from an integral number of G-Space pixels. For the purposes of geometric analysis each pixel will be treated as a rectangle (or square) in the respective space.  $\Delta l$  and  $\Delta s$  are the line and sample resolutions, the  $(m,n)^{th}$  pixel is the region in the respective space covered by  $[m\Delta l, (m+1)\Delta l] \times [n\Delta s, (n+1)\Delta s]$ .

#### A. CLASSIFICATION ERRORS

To treat classification errors independently of geometry, assume an ideal sensor which generates G-space pixels. Assume a per pixel classifier which produces  $K$  spectral classes. For the suitability model these are merged into two informational classes - suitable and unsuitable. Then, for a given cell in the GDB, the RSD-derived input  $x$  is defined as the number of pixels in the cell which are in the "suitable" class. Let there be  $M$  pixels per cell. Then the cell is considered suitable if

$$x \geq M/2.$$

In the notation of equation (4),

$$C = [M/2, M]$$

The "true" classification of the cell is obtained by checking  $y$ , the number of pixels in the cell which truly belong to the "suitable" class. The classification errors at the pixel level will be called

misclassifications. The classification error at the cell level (due to the condition indicated by equation (4)) will be called misaggregation. The purpose of this subsection is to relate probabilities of misclassification and misaggregation.

Let

$$P_{ij} = \Pr \{ \text{A pixel is classified as } i \mid \text{true class} = j \} \quad (5)$$

for  $i=1,2$  and  $j=1,2$ . Let 1 be the suitable and 2, the unsuitable class. The probability of misaggregation can then be expressed as

$$\Pr(E) = \Pr\{x \geq M/2 \mid y < M/2\} \Pr\{y < M/2\} + \Pr\{x < M/2 \mid y \geq M/2\} \Pr\{y \geq M/2\} \quad (6)$$

Now, assuming independent classification of individual pixels we get

$$\begin{aligned} P_{k\ell} &= \Pr \{x=k \mid y=\ell\} \\ &= \Pr \{k \text{ pixels are assigned to 1 given that } \ell \text{ pixels are in 1}\} \\ &= \sum_{r=0}^k \Pr \{r \text{ pixels from 1 and } (k-r) \text{ pixels from 2 are assigned to 1 given that } (\ell \text{ pixels are in 1 and } M-\ell \text{ are in 2})\} \\ &= \sum_{r=0}^k \binom{\ell}{r} \binom{M-\ell}{k-r} P_{11}^r P_{21}^{\ell-r} P_{12}^{k-r} P_{22}^{M-\ell-k+r} \quad (7) \end{aligned}$$

$$Q_{\ell} = \Pr \{x \geq M/2 \mid y=\ell\} \Pr\{y=\ell\} = \sum_{k \geq M/2} P_{k\ell} q_{\ell} \quad (8)$$

Where  $q_{\ell} = \Pr\{y=\ell\}$ .

$$\begin{aligned} R_{\ell} &= \Pr\{x < M/2 \mid y=\ell\} \Pr\{y=\ell\} \\ &= \sum_{k < M/2} P_{k\ell} q_{\ell} \quad (9) \end{aligned}$$

Now,

$$\Pr(E) = \sum_{\ell < M/2} Q_{\ell} + \sum_{\ell \geq M/2} R_{\ell} \quad (10)$$

Thus,  $\Pr(E)$  can be evaluated using equations (7) through (10) given  $M$ ,  $P_{11}$ ,  $P_{22}$ , and  $q_{\ell}$  for  $\ell = 0, 1, 2, \dots, M$ . (Note that  $P_{12} = 1 - P_{22}$  and  $P_{21} = 1 - P_{11}$ ).

For the case where all values of  $y$  from 0 to  $M$  are equally likely (that is, the number of pixels in a cell which are "suitable" is uniformly distributed),  $q_{\ell} = 1/(M+1)$ . Table 1 shows the values of  $\Pr(E)$  for this case with the further assumption that  $P_{12} = P_{21}$  = Probability of misclassification.

## B. REGISTRATION ERRORS

To treat registration errors independently of classification errors, we assume a perfect classifier which classifies each pixel into the two informational classes "suitable" and "unsuitable" according to the class with the larger area within the pixel. This is assumed to be true regardless of the geometry (orientation or sampling resolution) of the image space. These assumptions are approximated by a classification which uses a large number of spectral classes and merges them into the two informational classes.

Now, consider the  $(i,j)^{\text{th}}$  pixel in G-space. Let  $(u,v)$  be the coordinates in I-space corresponding to  $(i\Delta\ell, j\Delta s)$  where  $\Delta\ell, \Delta s$  are the line and sample resolutions of pixels in G-space. Then the nearest neighbor resampling assigns the  $(m,n)$  pixel in I-space to the  $(i,j)^{\text{th}}$  in G-space where  $(m,n)$  are the integers nearest to  $(u/\Delta\ell, v/\Delta s)$  and  $\Delta\ell, \Delta s$  are the line and sample resolutions in the I-space.

The class assigned to a cell in G-space depends on the majority class among the  $M$  G-space pixels in that cell. These, in turn, are uniquely determined by the  $M$  corresponding I-space pixels.

The "true" class assignment of a cell is defined as the majority class among the  $M$  G-space pixels, if they had been sensed and classified in G-space. (An alternative is to measure the actual areas occupied by each class within the cell and assign the class with the larger area). An error may result when the assignment using the I-space pixels differs from the true assignment. This is because a boundary separating the two classes may separate a G-space pixel and the corresponding I-space pixel differently.

This error is inherent in the process of using RSD with a GDB and occurs regardless of the accuracy with which the G-space to I-space transformation is determined. In practice, there are errors in finding the transformation also. Both these errors can be estimated by the procedure which is described below.

Over a small neighborhood around a cell, (say, less than  $10 \times 10$  pixels), a linear approximation is valid for the transformation between the geographic and image coordinates. Let

$$\begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = B \begin{bmatrix} U \\ V \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} \quad (11)$$

Where  $u, v$  are the image coordinates and  $U, V$  are the corresponding geographic coordinates. Assume that the sampling resolutions in the two spaces are included in the matrix  $B$  such that, given the  $(i, j)^{th}$  pixel in G-space, the corresponding line and sample numbers in I-space are given by

$$k=[u(i, j)], \quad l=[v(i, j)] \quad (12)$$

where

$$\begin{bmatrix} u(i, j) \\ v(i, j) \end{bmatrix} = B \begin{bmatrix} i \\ j \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} \quad (13)$$

and  $[n]$  denotes the integer nearest to  $x$ . Denote the class associated with the  $(i, j)^{th}$  pixel in G-space by  $A(i, j)$ . Let the classification of the  $(k, l)^{th}$  pixel in I-space be  $C(k, l)$ .

Now, let  $\hat{B}, \hat{u}_0, \hat{v}_0$  define the estimated approximation to the transformation of equation (11). These could, for example, be derived using ground control points.

$$\hat{k}=[\hat{u}(i, j)], \quad \hat{l}=[\hat{v}(i, j)] \quad (14)$$

$$\begin{bmatrix} \hat{u}(i, j) \\ \hat{v}(i, j) \end{bmatrix} = \hat{B} \begin{bmatrix} i \\ j \end{bmatrix} + \begin{bmatrix} \hat{u}_0 \\ \hat{v}_0 \end{bmatrix} \quad (15)$$

Then, the class assignment  $A(i, j)$  made to the  $(i, j)$  pixel in G-space is given by

$$\hat{A}(i, j)=C(k, l) \quad (16)$$

[Note that even in the absence of errors in estimating  $(B, u_0, v_0)$ ,  $\hat{A}(i, j)$  may not equal  $A(i, j)$ ].

Now, suppose a boundary between the suitable and unsuitable classes in and around a cell is known. We can then find  $A(i, j)$  for all  $(i, j)$  covering the cell. Also, if  $B, u_0, v_0$  are known, we can transform the boundary to the I-space and compute  $C(k, l)$  for all  $(k, l)$  of interest. Next, from  $B, u_0, v_0$ , we can find  $\hat{A}(i, j)$  using equations (14) through (16).

The class assigned to the cell is then the majority class in the array  $\hat{A}(i, j)$ , whereas the true class is the majority in the array  $A(i, j)$ . Analytical expressions for  $\hat{A}(i, j)$  or  $A(i, j)$  in terms of the boundary are very difficult (if not impossible) to derive. Therefore, a Monte Carlo simulation using various piecewise linear random boundaries has been made to estimate the probability of misaggregation.

The simulation procedure is as follows:

- (i) choose a starting point at random on one of the edges of the neighborhood.
- (ii) choose the first line with random length between 1 and 2 and a random orientation pointing into the neighborhood. The end of this line is the second boundary point.
- (iii) Given  $M$  boundary points we have  $(M-1)$  lines in the piecewise linear boundary. Find the  $M^{th}$  line with random length between 1 and 2 and at a random angle in the interval  $(-3\pi/8, 3\pi/8)$  with respect to the  $(M-1)^{th}$  line.
- (iv) stop boundary computation when the boundary line intersects an edge of the neighborhood.

Only values between  $-0.5$  and  $0.5$  are used for  $(u_0, v_0)$  with no loss of generality since other values can be accounted for by a shift of the origin. Values of  $(\hat{u}_0, \hat{v}_0)$  are used in equal steps with  $|u_0 - \hat{u}_0| < 1.9$  and  $|v_0 - \hat{v}_0| < 1.9$ .

The error in estimating  $B$  results in rotation, skew and scale changes. However, for the local neighborhood around a cell, these can be neglected and taken into account by the shift errors  $|u_0 - \hat{u}_0|$  and  $|v_0 - \hat{v}_0|$ . (If the errors over a cell due to estimating  $B$  are not negligible, then over an image of several hundred cells, the errors will be several pixels, which would not be acceptable).

Tables 2 through 4 show the effects of various shifts in the  $u$  and  $v$  directions. The entries in these tables show estimates of the probability of misassignment of cells based on 200 boundaries.

#### IV. EXPERIMENTS

The previous section considered the effects on errors from a theoretical point of view with certain simplifying assumptions. A thorough verification of the results which will be applicable to all types of databases used in practice will involve either a general characterization of such databases and simulations thereof or experimentation on a large number of databases. This section reports sample experiments on a typical database and shows results of the classification and registration errors both independently and

jointly. In these experiments, a baseline data set using RSD is produced with minimal errors. Varying degrees of error are introduced as perturbations to the baseline. Both the baseline and the perturbed data sets are aggregated. Pixel by pixel and cell by cell comparisons are made to determine the effects of the induced errors. The aggregated RSD are combined with the GDB data and used in a suitability model. The model outputs show the impact of errors on the GIS-derived decisions.

#### A. DATA SETS

The RSD used in the experiment were a subset of the Landsat Scene 83009915071, dated June 12, 1978, covering Harrisburg, Pennsylvania. The data were cloud-free and of good quality. Color infrared aerial photography flown in February 1974 and USGS 7 1/2 minute topographic maps were used as reference data. The GDB was a part of the Environmental and Land Use Data System also covering the Harrisburg area. The data were obtained from the Pennsylvania Power and Light Company (PP&L) in grid format with a cell size of 22.9 acres and consisted of 43 layers. Table 5 shows the categories associated with each of these layers.

#### B. EQUIPMENT

The experiments were carried out using the Interactive Digital Image Manipulation System (IDIMS) at the Eastern Regional Remote Sensing Applications Center (ERRSAC), Goddard Space Flight Center. This system consisted of several components of which a COMTAL image display terminal, a TALOS coordinate digitizer table, and the associated software were used extensively in these experiments.

#### C. BASELINE DATA SET PREPARATION

The GDB was referenced on the Universal Transverse Mercator (UTM) coordinate system. Therefore, the Landsat dataset was geometrically corrected to UTM coordinates. Seventy evenly distributed Ground Control Points GCP's were selected. Their I-space and G-space coordinates were carefully determined. The origin of the G-space was chosen to coincide with the Northwest corner of the GDB. The pixel size in G-space was taken to be 61m squared, to yield 25 pixels per cell. A third order polynomial was used for the geometric transformation. The coefficients of the polynomial were determined using a least squares fit and discarding five of the GCP's with the largest residual errors. The average residual errors using the remaining 65

GCP's was 0.6 pixel. The geometric correction was performed using nearest neighbor resampling. The resulting image size was 595x775 pixels.

A semi-supervised approach was used for the classification of the Landsat data. The clustering algorithm ISOCLS was used on a random sample of the data to define 24 spectral classes. By comparing the cluster map with ground truth (as determined by aerial photographs, topographic maps, and personal knowledge of the area), several classes were identified as mixtures. A second clustering run was used to resolve these problems. The result was a set of 29 spectral classes. A maximum likelihood classification was then performed on the entire data set.

The geometrically corrected classification map with 29 classes was used as the "baseline" data set. A "confidence map" was also produced by the maximum likelihood classifier showing the probability of the assigned class at a pixel being correct.

#### D. PERTURBATION

The most likely classification errors arise due to confusion between classes whose spectral characteristics are similar. Therefore, the following procedure was used to simulate such errors. The classes in the baseline classification map were numbered such that, to the extent possible, nearest neighbors in spectral space had adjacent class numbers. A confidence threshold was chosen for defining a particular perturbation. All pixels with higher values than the threshold in the confidence map were left unchanged. The class numbers of the other pixels were increased or decreased by one, at random.

The most likely geometric errors arise due to imperfections in GCP selection. The following procedure was used to simulate these. Only a subset of the GCP's used for the baseline data set were used. A pair of perturbed data sets was produced with a random selection of 1/2 and 1/4 of the GCP's. Another pair was produced with 34 and 14 GCP's with the highest residual errors. Table 6 shows the residual errors for the various cases.

The two perturbation types were combined to examine the joint contributions of the classification and the registration errors. Table 7 shows the perturbed data sets generated. In this table, nonblank entries are the numeric designations of these data sets.



## E. AGGREGATION

Aggregation involved combining 5x5 pixel regions of the RSD into cells to match the GDB resolution. Two techniques, Systematic Aligned Sampling (SAS) and Dominant Land Use (DLU) were used. The SAS technique assigns the thematic category of the central pixel of a cell to the cell. The DLU method assigns the dominant category instead. The spectral classes were merged into two information classes at the pixel level before aggregation. For the purposes of the farmland suitability model, these classes are "available" and "unavailable" for farming. Spectral classes labelled residential, commercial or water were merged into the "unavailable" class and all others into the "available" class. (With SAS aggregation, the order of class-merging and aggregation does not affect the aggregated result, while with DLU aggregation it does. The two aggregation methods applied to the ten data sets discussed above resulted in twenty aggregate data sets.

## F. MODELLING

Nine of the ten DLU aggregated data sets were used in the farmland suitability model. The soil and slope information from the GDB were used as indications of agricultural potential as defined by the Soil Conservation Service<sup>5</sup>. A binary output was generated for each case showing cells with "high agricultural potential and available for farming (i.e., suitable)" and "low agricultural potential or unavailable for farming (i.e., unsuitable)."

## G. RESULTS

The perturbed data sets before aggregation, after aggregation and after modelling were compared pixel by pixel (or cell by cell) with the respective baseline data sets. The percentage disagreements were computed from contingency tables. Given  $m_{ij}$ , the number of pixels (or cells) assigned to class  $i$  in the baseline and class  $j$  in a perturbed dataset for  $i, j=1, 2$ , the percentage disagreement  $p$  is given by

$$p = 100(m_{12} + m_{21})/M$$

Where  $M$  is the total number of pixels (or cells).

Table 8 shows the values of  $p$  for data sets numbered 1 through 9 in Table 7.

It can be seen from this table that:

1. Biases in GCP location have a significantly greater impact on the agreements between the baseline and perturbed data sets than the number of GCP's as evidenced for data sets 4 through 7.
2. The joint effects of classification and registration errors are less than the sum of the two as seen by comparing rows (1,6,8) and (2,7,9). Evidently, this is due to overlap in the sets of erroneous pixels (cells) caused by the two types of perturbation.
3. No significant change is seen between the disagreement values before and after SAS aggregation. This is due to the fact that SAS aggregation is merely a sampling of the pixels.
4. Aggregation by the DLU method reduces the disagreements considerably. The differences are more significant in the case of registration errors than for classification errors.
5. After the modelling step, the disagreements are further reduced. It can be seen that the last two columns in the table are roughly proportional.

Their ratios are approximately equal to the ratio of the number of cells with high agricultural potential to the total number of cells, as is to be expected. (They would be exactly equal if the RSD disagreements were uniformly distributed throughout the image).

6. Comparing the preaggregation and post-DLU-aggregation values in rows 1,2,3 of Table 8 with the probabilities in Table 1, it can be noted that the predicted misaggregation probabilities in Table 1 are smaller. This is due to the several simplifying assumptions made in deriving Table 1. An examination of the difference image between the baseline and the perturbed classifications indicates that classification errors occur in groups of several pixels rather than being randomly distributed as was implied in the derivation of Table 1.

7. The results of the Monte Carlo simulation of misregistration shown in Tables 2,3 and 4 yield larger misaggregation error estimates than in rows 5,6,7 in Table 8 (after DLU aggregation). This is likely due to the fact that the simulation regards each cell as containing a boundary while about 40% of the cells in the image are homogeneous. Also, the distributions of the simulated and actual boundaries may be different.

#### IV CONCLUSION

This paper has attempted to characterize the behavior of a specific type of model used for decision making with Geographic Information Systems. The outputs of such a "suitability" model will have varying amounts of error depending on the errors in input data. The process of preparing remotely sensed data as input to a GIS has been analyzed. The errors associated with classification and registration, the two major steps in the process, have been examined. Attention has been focussed on models requiring resolutions less than that of the remotely sensed data. In such cases, the errors caused during classification and registration are partially compensated for by aggregation of pixels. This compensation is quantified through an analytical model, a Monte Carlo Simulation and experiments with a typical geographic data base. It is found that error reductions of the order of 50% occur due to aggregation of 5x5 pixel areas.

Further work in this area should include:

- (i) Sensitivity analysis for outputs from other types of models, especially those using multi-cell decision rules (as opposed to "per cell" decisions considered here);
- (ii) A more general characterization of classification errors where correlations among neighboring pixels are taken into account;
- (iii) A general means of describing boundaries and their statistical properties to facilitate prediction of effects of registration errors on a given class of data sets.

#### V REFERENCES

1. W. G. Schneider, Jr., "Integrated use of Landsat Data for State Resource management," Lexington, Kentucky, The Council of State Governments, 1979.
2. R. K. Boyd, P. Jones, and R. Dasgupta, "Survey of State-Based Geographic Information System Applications," Greenbelt, Maryland, Computer Sciences Corporation, Technical Memo 80/6330, 1980.

3. M. E. Wehde, "Impact of Cell Size on Inventory and Mapping Errors in a Cellular Geographic Information System," Brookings, South Dakota, South Dakota State University, 1979.
4. Staff, "ELUDS Fact Sheet," Allentown Pennsylvania, Pennsylvania Power and Light Company, 1976.
5. Community Soil Information Project, "Soil and Water Data For Environmental Planning, City of Medina," Wayzata, Minnesota, Hennepin County Soil and Water Conservation District, 1978.

Table 1

Probability of Misaggregation as a function of M (pixels per cell) and  $p_{12}$  (probability of misclassification)

M \ $p_{12}$	.05	.10	.15	.20	.25	.30	.35	.40	.45
4	.049	.097	.143	.189	.235	.281	.326	.372	.417
6	.047	.091	.133	.177	.222	.270	.319	.371	.424
8	.045	.085	.123	.163	.207	.256	.308	.365	.424
9	.044	.081	.117	.156	.200	.249	.305	.366	.432
12	.042	.075	.106	.141	.182	.230	.286	.350	.419
15	.039	.069	.097	.128	.166	.213	.271	.339	.417
16	.039	.067	.095	.125	.163	.209	.266	.335	.412
20	.036	.062	.086	.114	.148	.192	.250	.321	.405
24	.034	.057	.079	.105	.136	.178	.235	.309	.398
25	.034	.056	.077	.102	.133	.174	.231	.305	.397

Table 2 Probability (estimated) of Misaggregation Versus Registration Errors B=Identity Matrix,  $u_0=v_0=0.0$

$\Delta u_0 \backslash \Delta v_0$	-1.90	-.95	.00	.95	1.90
-1.90	.399	.315	.253	.219	.247
-.95	.292	.202	.163	.169	.202
.00	.264	.129	.000	.096	.169
.95	.202	.118	.079	.090	.135
1.90	.275	.225	.174	.152	.157

Table 3 Probability (estimated) of Misaggregation Versus Registration Errors B=Identity Matrix,  $u_0=v_0=0.5$

$\Delta u_0 \backslash \Delta v_0$	-1.90	-.95	.00	.95	1.90
-1.90	.194	.153	.159	.159	.194
-.95	.118	.024	.082	.082	.159
.00	.118	.082	.094	.094	.135
.95	.118	.082	.094	.094	.135
1.90	.229	.182	.159	.159	.165

Table 4 Probability (estimated) of Misaggregation Versus Registration Errors B=Identity Matrix,  $u_0=v_0=-.5$

$\Delta u_0 \backslash \Delta v_0$	-1.90	-.95	.00	.95	1.90
-1.90	.402	.318	.246	.257	.218
-.95	.296	.207	.162	.162	.168
.00	.257	.134	.050	.056	.106
.95	.263	.128	.034	.034	.095
1.90	.201	.117	.078	.078	.089

Table 5 Strata (Layer) Categories and Labels for the  
PP&L ELUDS Stratified Data Base (1 of 2)

Category	Layer	Variable	
LOCATION	1	Row	
	2	Column	
	3	Map Module	
SERVICE AREAS	4	Service Areas	
PP&L FACILITIES	5	PP&L Facilities - Point Data	
INFRASTRUCTURE	6	Highways	
	7	Railroads	
	8	Transmission Lines	
	9	General (Pipelines, Vortac Stations, etc.)	
	10	Scenic Roads/Canals/Trails	
PUBLIC LANDS (POINT DATA)	11	Historic Sites/Natural Areas (WPC) County Prefix	
	12	Historic Sites	
	13	Natural Areas (WPC)	
	14	Other	
PUBLIC LANDS (POLYGON DATA)	15	Public Lands (Polygon Data)	
COURSE LINES	16	Course Lines	
FUTURE LAND USE	17	Future Land Use Trends	
LAND USE DATA	18	Land Use and Land Cover	
ANALYSIS	19	Political Units	
	20	Hydrologic Units	
	21	Census County Subdivisions	
	22	Federal Land Ownership (to be added)	
	23	State Land Ownership (to be added)	
TERRAIN UNIT	24	Terrain Unit Polygon Number	
	25	Vegetation/Land Cover	
	26	Landform	
	27	Slope	
	28	Soils	
	29	Agricultural Potential	
	30	Soil Depth	
	31	Soil Permeability	
	32	Seasonally High Water Table	
	TERRAIN UNIT (Cont'd)	33	Geologic Code Number
		34	Rock Type
		35	Bedding
36		Surface Drainage	
37		Groundwater	
38		Porosity	
39		Ease of Excavation	
40		Cut Slope Stability	
41		Foundation Stability	
42		Mineral Resources	
43		Flood Prone	

Table 6 Residual Error (in Number of Pixels) Associated With GCPs Used in Generating the Various Registered Data Sets

	<u>Baseline</u>	<u>1/2 of GCPs</u>	<u>1/4 of GCPs</u>	<u>34 Worst GCPs</u>	<u>14 Worst GCPs</u>
Average Residual	0.6	0.6	0.6	3.0	5.9
Minimum	0.1	0.1	0.0	1.1	1.6
Maximum	1.8	1.8	1.7	31.6	29.3

Table 7 Matrix of Experimental Data Sets

		Classification Experiment			
Registration Experiment		<u>Baseline</u>	<u>Thresh old = .2</u>	<u>Thresh old = .5</u>	<u>Thresh old = .8</u>
	Baseline	0	1	2	3
	1/2 of GCPs	4			
	1/4 of GCPs	5			
	34 Worst GCPs	6	8		
	14 Worst GCPs	7		9	

Table 8 Percentage Disagreements from Baseline

<u>Data Set #</u>	<u>Before Aggregation</u>	<u>After Aggregation</u>		<u>After Modelling</u>
		<u>SAS</u>	<u>DLU</u>	
1	3.3	3.6	2.8	1.8
2	8.3	8.5	5.6	3.4
3	12.4	12.4	7.8	-
4	5.0	-	-	-
5	6.4	6.2	2.8	1.8
6	16.6	16.2	6.5	4.3
7	23.6	23.6	11.1	7.4
8	18.1	17.5	7.2	4.8
9	25.5	25.1	12.7	8.1