An Automated, Objective Procedure for Selecting Representative Field Sample Sites

Steven D. Warren, Mark O. Johnson, William D. Goran, and *Victor E. Diersing*
U.S. Army Construction Engineering Research Laboratory, Environmental Division, P. O. Box 4005, Champaign, IL 61824-4005

ABSTRACT: The selection of representative sample sites for land-condition analysis and ground truthing of digital satellite imagery can be a difficult and subjective task. A procedure has been developed which automates the site selection process and reduces subjectivity. Within a geographic information system, a data layer representing spectral land-cover types derived from an unsupervised classification of satellite imagery is overlaid with an ancillary data layer (in this case, a digital soil survey). Inventory sites are randomly assigned to areas represented by the unique land-cover/soil combinations in a stratified fashion based on their proportional contribution to the land area being surveyed. The process ensures proportional representation of the various land-cover types while providing a basis for post-classification sorting if field data reveal significant variability within the land-cover classes.

INTRODUCTION

EFFECTIVE NATURAL RESOURCE MANAGEMENT is dependent, in part, on accurate assessment of the kinds, amounts, and conditions of existing resources. When properly interpreted, satellite imagery can provide valuable information regarding various geologic, botanic, pedologic, and aquatic resources. The utility of remotely sensed imagery, however, is limited by the ability of the user to correlate spectral response patterns with conditions on the Earth's surface. The strength of such correlations is further limited by at least four factors: (1) the appropriateness and accuracy of image preprocessing techniques, (2) the adequacy of algorithms used to classify the remotely sensed images, (3) the suitability of ground truthing methods, and (4) the representativeness of field sampling locations.

Much has been written regarding preprocessing techniques and the relative merits of multispectral clustering algorithms (e.g., Jensen, 1986; Campbell, 1987). In addition, there is an abundance of information dealing with field inventory methods (*e.g.*, Mueller-Dombois and Ellenberg, 1974; Chapman, 1976). However, comparatively little has been written regarding procedures to ensure the representativeness of field sample sites as they relate to interpretation of remotely sensed imagery. Most site selection procedures reported in the literature are performed subjectively (*e.g.*, Fitzpatrick, 1988; Talbot and Markon, 1988) and are easily biased by desires to stay near roads and/or familiar landmarks (Joyce, 1978; Cibula and Nyquist 1987).

Random placement of sampling sites in the field is an important factor in the statistical reliability of the resulting data. However, because land-cover types (and their associated spectral categories) do not occur randomly in nature, a simple random allocation of sampling sites may not adequately represent all spectrally recognized land-cover categories. Stratification of sampling sites by spectral categories improves the likelihood that all categories are adequately represented. Due to the nature of remotely sensed digital data, however, considerable variability in cover type may exist within the spectral categories. To address this problem, some researchers have suggested the use of ancillary data, such as aspect or elevation, to assist the interpretation of spectral land-cover categories in a post-classification sorting process (Hutchinson, 1982).

The purpose of our research effort was to develop an automated, objective procedure to select sites for field sampling and verification of multispectral classification categories. The procedure utilizes geographic information system technology and *a priori* incorporation of ancillary data to maximize the representativeness of field sample sites.

METHODOLOGY

STUDY SITE

The site selection procedure described here has been implemented at 13 major U.S. Army training facilities across the United States and West Germany. It is illustrated with an example from Dugway Proving Ground, Utah. Dugway Proving Ground is approximately 340, 565 ha (840,900 ac) in size, and is located in the western desert of Tooele County, Utah. The facility has been protected from grazing since 1942 and comprises some of the largest remnants of the salt desert ecosystem in North America. Major landscape features include vegetatively depauperate salt flats, saltbush-greasewood (*Atriplex-Sarcobatus*) shrublands, dunelands, and granitic outcrops.

GEOGRAPHIC INFORMATION SYSTEM

Geographic information systems are designed to store, manipulate, analyze, and display spatial data that may be derived from a variety of cartographic and thematic sources. The speed, accuracy, and ease with which analysts can process spatial data are major benefits of geographic information systems in general. The system chosen for this project was the Geographical Resources Analysis Support System (GRASS)(Westervelt, 1988), a public domain system developed by the U.S. Army Construction Engineering Research Laboratory. GRASS includes functions to read, georeference, filter, and classify remote imagery, and to integrate images with other geographic data. Cartographic data used in the site selection process were resampled to 20-m grid-cell (raster) format prior to analysis to correspond with the resolution of the satellite imagery.

SATELLITE IMAGERY

Remote imagery recorded by the French SPOT (Systeme Probatoire pour l'Observation de la Terre) satellite was acquired for the Dugway Proving Ground area. Due to the large land area included, two images were required (#154826-98905041842572X, 04 May 1989, and #15472698905041842591X, 04 May 1989). The date of the imagery was selected to correspond to the approximate annual peak of plant growth. The images were geometrically rectified and then joined to form one composite image. The composite image was masked to exclude areas outside the military installation boundary. Areas determined by visual inspection to be clouds and shadows of clouds were also masked from the image. An unsupervised classification was performed on the unmasked portion of the image using all three spectral bands. The bands correspond to the green (0.50 to 0.59 μm), red (0.61 to 0.68 μm), and near infrared (0.79 to

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 56, No. 3, March 1990, pp. 333-335.

0099-1112/90/5603-333$02.25/0

0.89 μm) portions of the electromagnetic spectrum, respectively. The classification process used a two-pass clustering algorithm. In the first pass, the algorithm selected a subsample of pixels (*grid-cells*) and built clusters based on reflectances in the spectral bands. The second pass assigned all pixels to the clusters using the maximum likelihood classifier. The classification was limited to a maximum of 20 spectrally unique categories.

Site Selection Procedure

In order to automate the site selection process, an interactive program was written which prompted the user with regard to required inputs. The algorithm then selected and utilized the appropriate GRASS features to perform the necessary tasks (Figure 1).

The first step in the site selection procedure was to input the primary and ancillary data layers to be used. The spectral classification map layer served as the primary layer for the stratification of field inventory sites. From previous experience it was anticipated that considerable variability in botanical composition and plant density could exist within the computer-generated spectral catergories. This could be due to the limitation of 20 land-cover categories imposed during the unsupervised classification, or may result from the occurrence of more than one plant community with similar spectral reflectance properties. In the latter case, differences in vegetative cover and plant species composition are often related to characteristics of the underlying soils. Furthermore, in areas of sparse vegetation, soil properties themselves may dramatically influence reflectance patterns recorded by the satellite. In order to account for variability attributable to soils, a digital soil survey for Dugway Proving Ground was obtained from the U.S. Soil Conservation Service. The digital soil survey, containing a total of 23 soil series or complexes, served as the ancillary data layer for sample site stratification.

Within GRASS, the soil and spectral land-cover data layers were combined to create a cross-product map composed of all unique combinations of the categories in the two input layers. An internal report indicating the relative areal extent of each land-cover/soil combination served as the basis for the numerical allocation of field sample sites to the land-cover/soil categories. Time and monetary constraints at Dugway Proving Ground dictated the use of no more than 200 sites. These were allocated to the various land-cover/soil combinations based on their proportional contribution to the land area being surveyed.

Due to the complexity of the landscape, a large number of land-cover/soil combinations occurred which, when considered individually, were too small to merit sample sites. However, when considered as an aggregate, they comprised a significant proportion of the land area. This prevented the algorithm from allocating the full complement of 200 sites. The site selection program then began an iterative, stepwise process of searching the internal report for the largest category failing to receive a site in the first pass. A site was allocated and the process was continued until all 200 sites were distributed.

One at a time, the land-cover/soil categories were extracted from the cross-product data layer into a temporary file where spatially discrete clumps of cells were grouped into polygons recognized as unique entities. Some land-cover/soil combinations were represented by a single polygon while others were represented as a series of spatially disjunct polygons. The size of each polygon was calculated in GRASS. It was assumed that some degree of inherent spatial error would be present in both the imagery and soil source data (Walsh *et al.*, 1987), and that the probability of error would be greatest at the edges of polygons. Due to the likelihood of spatial error, as well as the land area needed to accommodate field methods used for ground truthing, polygons less than 2 ha (5 ac) in size were eliminated from further consideration in the site selection procedure.

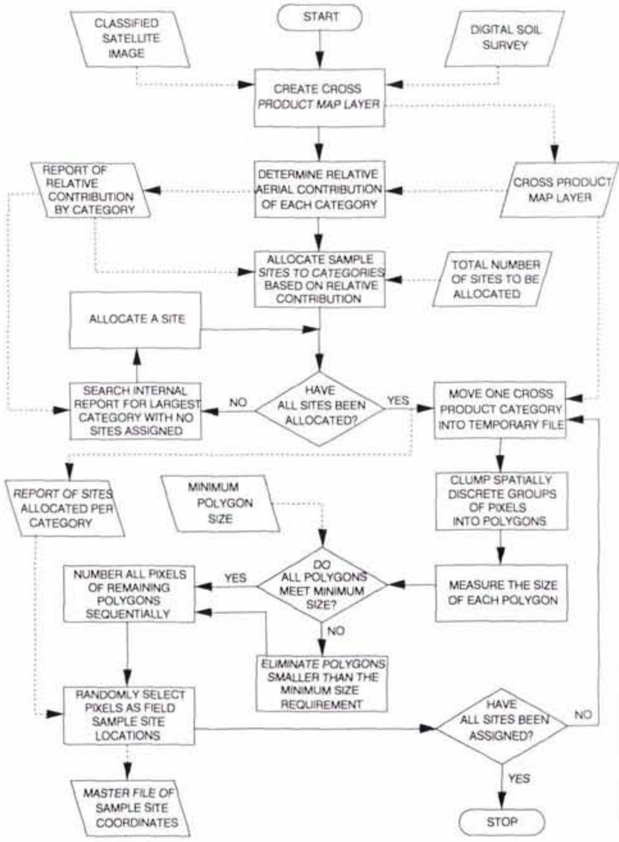

Fig. 1. Flow diagram for the site selection model.

Pixels within the remaining polygons were numbered sequentially. A random number generator was used to produce a set of numbers equivalent to the quota of sites allocated to the land-cover/soil category in the temporary file. Pixels corresponding to the randomly generated numbers identified field sample sites. The coordinates of each site were added to a master list. The next land-cover/soil category was then moved into the temporary file and the process was repeated until all categories of sufficient size to receive one or more sites were evaluated. When complete, the master list contained the coordinates of all 200 sample sites.

Ground Truthing

Field crews were provided with clear Mylar overlays registered to U.S. Geological Survey 7½-minute (1:24,000-scale) quadrangle maps (Plate 1). The overlays were printed with all polygons satisfying the minimum size requirement. The color scheme for the polygons was based on the spectral land-cover categories. Boundaries of soil mapping units were outlined in black. Pixels selected by the randomization process were labeled with conspicuous symbols. Field crews located and established permanent sample sites within the polygons containing selected pixels. Wherever possible, the field crews oriented sample sites away from the edges of the polygons where the potential for spatial error was the greatest. A global positioning system was used to locate sites which were difficult to identify by field observation alone. In the event that any given polygon was inaccessible, the crew selected the nearest polygon of the same land-cover color code and soil series.

DISCUSSION

The primary purpose for developing the model described here was to provide an objective procedure for selecting represent-

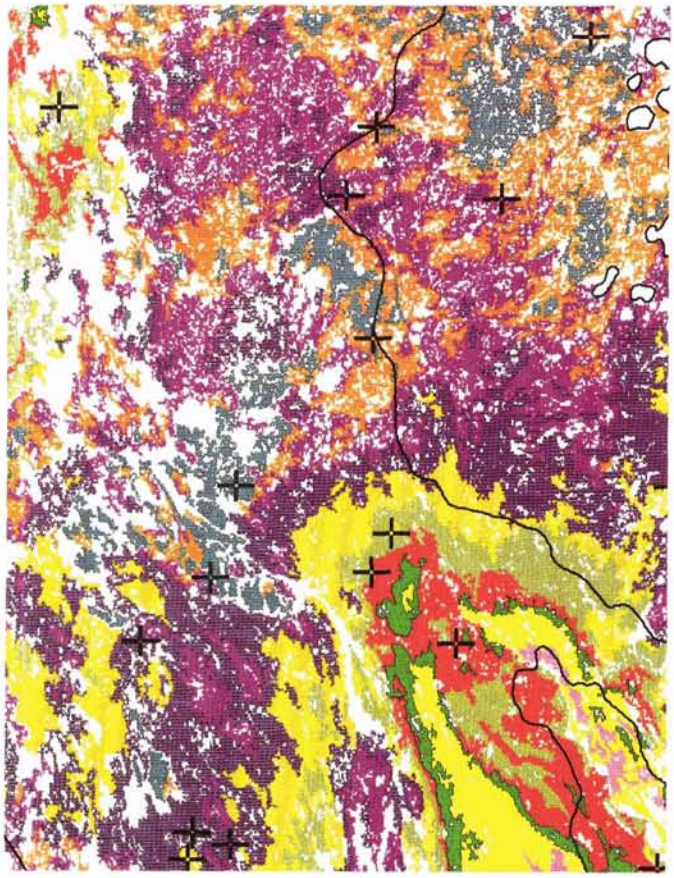

PLATE 1. An example of a site allocation overlay provided for the Granite Peak SW quadrangle map at Dugway Proving Ground. The color scheme for the polygons represents the various undefined spectral land-cover categories derived from an unsupervised classification of a SPOT image. Soil series delineations are outlined in black. Symbols identify polygons selected for inventory. Uncolored areas were dropped from the sampling scenario due to the disproportionately small contribution of the land-cover/soil combination to the installation as a whole or due to the small size of individual polygons (i.e., < 2 ha).

ative field samples for land-condition analysis and ground truthing of digital satellite imagery. Although it is not likely that subjectivity will ever be entirely removed from sampling procedures, the model greatly reduces subjectivity by limiting human interaction to a minimum number of required inputs. Representativeness of field sampling sites is improved by the stratification process and by the allocation of sites to the various spectral land-cover strata proportional to their contribution to the area being sampled. The *a priori* inclusion of an ancillary soil layer as a second level of stratification helps ensure a measure of variability within the land-cover strata. In addition, it provides a basis for post-classification sorting where ground truth data reveal significant variability within the spectral land-cover categories. Statistical reliability of data derived from field sampling is enhanced both by stratification and the random assignment of sites within the strata. Automation of the site selection model provides the additional benefit of time savings. Whereas our previous efforts to select sample sites without the aid of a geographic information system consumed up to several weeks and were replete with subjectivity, the model reduced the selection process to a matter of a few hours, assuming the availability of correctly formatted digital data.

The choice of digital soil survey as the ancillary data layer was based on the intended use of the resulting field data. Other researchers may find that alternative ancillary data layers are more pertinent to their particular applications. Indeed, the theory behind the site selection model can be extended to any number of applications in addition to ground truthing of remote imagery. Some researchers may wish to incorporate more than one ancillary data layer. However, potential users of the model should be aware that an increase in the number of ancillary data layers or the use of any map layer with a very large number of categories will likely result in a reduction in the size of individual polygons, thus affecting the number of polygons that meet the user-defined minimum size.

ACKNOWLEDGMENTS

We are grateful to Cynthia A. Abrahamson who provided invaluable assistance in producing attractive and useful geographic information system hardcopy products.

REFERENCES

Campbell, J.B., 1987. *Introduction to Remote Sensing*. Guilford Press, New York, 551p.

Cibula, W.G., and M.O. Nyquist, 1987. Use of topographic and climatological models in a geographical data base to improve Landsat MSS classification for Olympic National Park. *Photogrammetric Engineering and Remote Sensing* 53:67–75.

Chapman, S.B., 1976. *Methods in Plant Ecology*. Blackwell Scientific Publications, Oxford, England, 536p.

Fitzpatrick, E.B., 1988. Mapping grizzly bear habitat with Landsat MSS data. *Remote Sensing for Resource Inventory, Planning and Monitoring* (J.D. Greer, ed.). American Society for Photogrammetry and Remote Sensing, Falls Church, Virginia, pp. 129–144.

Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement. *Photogrammetric Engineering and Remote Sensing* 48:123–130.

Jensen, J.R., 1986. *Introductory Digital Image Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 379p.

Joyce, A.T., 1978. *Procedures for Gathering Truth Information for a Supervised Approach to a Computer-Implemented Land Cover Classification of Landsat-Acquired Multispectral Scanner Data*. NASA Reference Publication 1015, 43p.

Mueller-Dombois, D., and H. Ellenberg, 1974. *Aims and Methods of Vegetation Ecology*. John Wiley & Sons, New York, 547p.

Talbot, S.S., and C.J. Markon, 1988. Intermediate-scale vegetation mapping of Innoko National Wildlife Refuge, Alaska using Landsat MSS digital data. *Photogrammetric Engineering and Remote Sensing* 54:377–383.

Walsh, S.J., D.R. Lightfoot, and D.R. Butler, 1987. Recognition and assessment of error in geographic information systems. *Photogrammetric Engineering and Remote Sensing* 53:1423–1430.

Westervelt, J.D. (Editor), 1988. *Users and Programmers Manual for the Geographical Resources Analysis Support System Version 3.0*. U.S. Army Construction Engineering Research Laboratory ADP Report N-87/22, Champaign, Illinois.