

Published as:

Moisen, G. G., R. D. Cutler, and T. C. Edwards, Jr. 1996. Generalized linear mixed models for analyzing error in a satellite-based vegetation map of Utah. Pages 459-466 in H. T. Mowrer, R. L. Czaplewski, and R. H. Hamre, editors. Spatial accuracy assessment in natural resources and environmental sciences, USDA Forest Service General Technical Report RM-GTR-277.

**Generalized Linear Mixed Models for Analyzing Error in a
Satellite-based Vegetation Map of Utah**

Gretchen G. Moisen
Research Forester
USDA Forest Service Rocky Mountain Research Station
507 25th Street, Ogden, UT 84401
801.625.5384 (PH)
801.625.5723 (FAX)
gretchen@nr.usu.edu (E-MAIL)

D. Richard Cutler
Associate Professor
Department of Mathematics and Statistics
Utah State University, Logan, UT 84322-3900
801.797.2811 (PH)
801.797.1822 (FAX)
richard@sunfs.math.usu.edu (E-MAIL)

and

Thomas C. Edwards, Jr.
Research Ecologist and Associate Professor
USGS Utah Cooperative Fish and Wildlife Research Unit
Department of Fisheries and Wildlife
Utah State University, Logan, UT 84322-5210
801.797.2529 (PH)
801.797.4025 (FAX)
tce@nr.usu.edu (E-MAIL)

ABSTRACT. With the increasing demand for broad-scale vegetation maps for ecosystem management and conservation planning comes the need for flexible tools to assess thematic accuracy of these maps. In this paper, we use a generalized linear mixed model (GLMM) to explore the relationship between thematic accuracy in the blackbrush cover-type of a satellite-based vegetation map of Utah and various topographical and heterogeneity components of that map. Because of the difficulty in accessing many rugged areas of this State, two strata were defined based on proximity to roads. Vegetation type was recorded on heterogeneous linear clusters of sample points within the "off-road" strata, and on randomly distributed sample points within the "road" strata on selected USGS quadrangle maps. A binary response (correctly classified / incorrectly classified) was modeled as a function of both fixed and random effects accounting for spatially autocorrelated observations and different covariance structures for the random effects. The modeling exercise suggested a strong relationship between map error in the blackbrush cover-type of the Colorado Plateau of Utah, and stratum, slope and local heterogeneity.

INTRODUCTION

Thematic accuracy of vegetation cover maps derived from satellite imagery may be related to many factors, including elevation, aspect, slope, local heterogeneity and distance to vegetation boundaries. Exploring the relationship between the components of the vegetation classification model and its uncertainty is a logical step in an analysis of map error that is sensitive to both map use and subsequent improvements of the map.

Although many new techniques are being explored to address map uncertainty, generalized linear mixed models (GLMM's) have yet to be applied. Through a GLMM, data from any one of a variety of continuous and discrete distributions can be linked to a linear structure that may contain both fixed and random effects. GLMMs can also account for correlation among observations as well as among random effects terms in the linear structure (Wolfinger and O'Connell 1993). This flexibility may prove valuable in addressing map uncertainty as well as have numerous other broad-scale applications.

In this study, we use a GLMM to explore the relationship between the error in the blackbrush cover-type of a vegetation map of Utah and various topographical and heterogeneity components of that map.

DATA

A cover-map of Utah, ~219,000 km² in size, was developed from a state-wide Landsat Thematic Mapper (TM) mosaic created from 24 scenes at 30 m resolution (Homer et al. 1997). A total of 38 cover-types were modeled. Modeling was accomplished using a four step modeling approach. Steps included: (1) the creation of a statewide seamless mosaic of TM images; (2) the subsetting of the mosaic into 3 ecoregions, the Basin and Range, Wasatch-Uinta and Colorado Plateau (after Omernik 1987); (3) the association of 1,758 state-wide field training sites to

spectral classes; and (4) the use of ecological parameters based on elevation, slope, aspect and location to further refine spectral classes representing multiple cover-types.

Following development of this cover-map, field data were collected to assess its thematic accuracy (see Moisen et al. 1994 for design considerations, Edwards et al. 1997 for analysis). Of primary interest were estimates of by-class and by-ecoregion accuracy of the map at the base model of one ha. A total of 100 7.5-min quadrangles were randomly selected roughly proportional to the area of the three ecoregions. Two strata were identified on each quadrangle based on proximity to roads. The "road" stratum consisted of all land within 1 km of a secondary or better road. All other lands fell within the "off-road" stratum. On each quadrangle, ten points were randomly selected within the road stratum, while ten were collected in a randomly oriented heterogeneous linear cluster within the off-road stratum. Data were then used to assess map accuracy based on procedures outlined in Edwards et al. (1997).

For this study, a subset of the state-wide data comprised of the blackbrush cover-type within the Colorado Plateau was modeled under a GLMM. Data consisted of 96 sample points collected on two strata (road, off-road) on 15 quadrangles. Anywhere from one and ten blackbrush points were available for each quad/stratum combination. Clustered blackbrush data in the off-road stratum were not necessarily adjacent sample points because blackbrush polygons were often intermixed with other cover-types not considered in this analysis.

MODEL

Using a logit link function, the binary response (correctly classified / incorrectly classified) was modeled as a function of both fixed and random effects while accounting for several covariance structures for random effects and for spatially autocorrelated errors. For this analysis, the observations on the 96 sample points were coded as 1 when the mapped cover-type agreed with the ground cover-type, and as 0 when they did not agree. Define \mathbf{y} to be our data vector of 96 0s and 1s satisfying

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \quad (1)$$

We used a logit link function

$$g(\mu) = \log\{\mu/(1-\mu)\} \quad (2)$$

and modeled

$$g(\mu) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}. \quad (3)$$

Here, $\boldsymbol{\beta}$ is a vector of unknown fixed effects with known model matrix \mathbf{X} , and \mathbf{v} is a vector of unknown random effects with known model matrix \mathbf{Z} . Assume $E(\mathbf{v}) = \mathbf{0}$ and $\text{cov}(\mathbf{v}) = \mathbf{G}$, where \mathbf{G} is unknown. An effect may be considered fixed if the inference space is limited to the observed levels of that effect. An effect may be considered random if the inference space is applied to a population of levels, not all of which are observed. Fixed effects considered in this

application included both discrete and continuous variables. Because quadrangle maps were randomly selected for subsampling from a population of quadrangles, quadrangles were modeled as random effects. Also, ε is a vector of unobserved errors with $E(\varepsilon|\mu) = \mathbf{0}$ and

$$\text{cov}(\varepsilon|\mu) = \mathbf{R}_\mu^{1/2} \mathbf{R} \mathbf{R}_\mu^{1/2}. \quad (4)$$

Here \mathbf{R}_μ is a diagonal matrix containing evaluations at μ of the variance function

$$V(\mu) = \mu(1 - \mu). \quad (5)$$

\mathbf{R} and \mathbf{G} were modeled using covariance structures detailed below.

Fixed Effects

Nine fixed effects variables were considered. Three were topographical variables used in the classification model itself. These were extracted from a 90 m Digital Elevation Model and include elevation in meters (ELEV), slope in degrees (SLOPE), and aspect. A transformation of aspect (TRASP), used by Roberts and Cooper (1989), takes the form

$$\text{TRASP} = \frac{1 - \cos(\text{aspect} - 30)}{2}. \quad (6)$$

This transformation assigns the highest values to land oriented in a north-northeast direction, the coolest and wettest orientation in Utah.

In addition to the three topographical variables, we considered four different measures of heterogeneity surrounding the sample point. Richness (RICH) is defined as the number of cover-types found in the surrounding 8 pixels. The other three heterogeneity variables, evenness (EVEN) and two measures of diversity (D_1 and D_2), are defined in Table 1. Higher values for all indices indicate increasing heterogeneity.

Two other fixed effects considered were the minimum distance in meters to a different map cover-type (DIST) and a variable indicating membership in the road or off-road stratum (STRATA). Strata was the only categorical fixed effect variable. All others were continuous.

Covariance Structures

Because quadrangle maps (QUAD) were randomly selected for subsampling from a the statewide population of quadrangles, these were included as random effects in the GLMM. Three covariance models for \mathbf{G} were considered (Table 2).

A spherical spatial covariance structure, illustrated in the last row of Table 2, was considered for \mathbf{R} . Here covariance between sample points is modeled as a function of distance between those points, accounting for both correlation between clustered locations and potential correlation

between sample points in relatively close quadrangles. Although numerous spatial structures could have been tried, the spherical structure is very flexible and converged more readily in preliminary trials.

Table 1. Measures of heterogeneity. Here S equals richness, n equals 8 pixels, and n_i is the number of pixels belonging to the i th of S cover-types.

Variable	Source	Formula
D_1	Hill (1973)	$D_1 = \exp \left(- \sum_{i=1}^S \left[\left(\frac{n_i}{n} \right) \ln \left(\frac{n_i}{n} \right) \right] \right) \quad (7)$
D_2	Simpson (1949)	$D_2 = \left[\sum_{i=1}^S \frac{n_i(n_i-1)}{n(n-1)} \right]^{-1} \quad (8)$
EVEN	Ludwig and Reynolds (1988)	$\text{EVEN} = \left(D_2 - 1 \right) / \left(D_1 - 1 \right) \quad (9)$

Table 2. Covariance structures for \mathbf{G} and \mathbf{R} .

Structure	Form
Simple	$\mathbf{G}_{ij} = \sigma^2$ for $i = j$, else 0 (10)
Compound symmetry	$\mathbf{G}_{ij} = \sigma_1^2 + \sigma^2$ for $(i = j)$, else σ_1^2 (11)
Varying coefficients	$\mathbf{G}_{ij} = \sigma_{ij}^2$ for $(i = j)$, else 0 (12)
Spherical spatial	$\mathbf{R}_{ij} = \sigma_e^2 \left[1 - \left(3d_{ij}/2\rho \right) + \left(d_{ij}^3/2\rho^3 \right) \right]$ for $d_{ij} \leq \rho$, else 0 (13)

Model Fitting Strategy

Parameters in the GLMM were estimated through pseudo-likelihood procedures as described in Wolfinger and O'Connell (1993) using a SAS macro supplied by Russ Wolfinger of SAS Institute Inc. This macro uses PROC MIXED and the Output Delivery System, requiring SAS/STAT and SAS/IML release 6.08 or later.

An iterative model fitting strategy was adopted. After identifying quadrangle maps as our random effects, all fixed effects were included in the model. We tried all combinations of covariance structures for \mathbf{G} as listed in Table 2. The best covariance structure was selected based on Akaike's Information Criterion and Schwarz's Bayesian Criterion (Wolfinger 1993). In subsequent iterations fixed effects were dropped based on significance of parameter estimates, likelihood ratio tests, and predictive capability of the model. Again, best covariance structure was selected for each iteration. Because of collinearity, the four measures of heterogeneity were considered in the model separately.

RESULTS

Likelihood ratio tests and parameter significance levels led us to favor a parsimonious model containing only STRATA, SLOPE and D_2 as fixed effects. Exclusion of other variables had little impact on the predictive capability of the model based on confusion matrices and plots of predicted values from different model trials. The covariance structures selected were simple and spherical for \mathbf{G} and \mathbf{R} , respectively. The signs of the fixed effects parameters indicate the relationship between error and the variables (Table 3). In this case, positive values for SLOPE and D_2 indicate that error increases as SLOPE and D_2 increase (Figures 1a -b). In contrast, the negative value for the off-road stratum indicates that probability of error is less in the off-road stratum and greater in the road stratum. The estimate of r , the spatial covariance parameter in \mathbf{R} , suggests that spatial dependence is negligible between sample units greater than 322 meters apart (Figure 1c).

Table 3. Parameter estimates and their standard errors for final GLMM.

Parameter	Estimate	SE	Pr > χ^2 or (Z^*)
Intercept	- 1.78	0.85	0.04
STRATA (off-road)	- 1.39	0.59	0.02
SLOPE	0.12	0.07	0.08
D_2	0.58	0.31	0.06
r	322.46	124.86	0.01*
s^2	1.63	1.37	0.23*

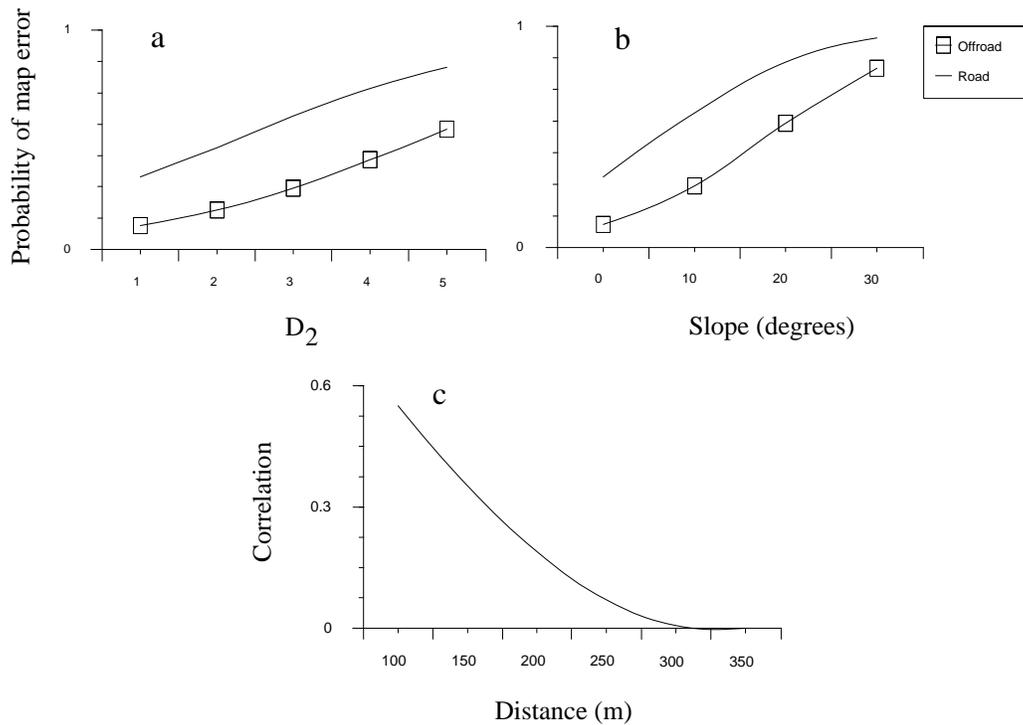


Figure 1a-b. Relationship between probability of map error and fixed effects.
 Figure 1c. Correlation modelled as a function of distance between sample points.

DISCUSSION

Frequently in broad-scale sample surveys clustered and subsampling sample designs are adopted for the sake of efficiency in estimation of population means and totals. However, such cost-effective designs can hamper efforts to further explore ecological relationships under a classical linear model framework by violating model assumptions like independence and normality of errors. In this study we illustrated how a GLMM affords the flexibility to analyze sample survey accuracy data. Through a GLMM we were able to include both fixed and random effects, account for spatially autocorrelated errors, and allow for a variety of covariance structures for the random effects.

The model we fit for blackbrush contained some information for map users beyond simple probability of misclassification by cover-type, and also provided information to the map-maker for model improvements. We learned that incorrect mapping of this cover-type tended to occur near roads in steep and heterogeneous areas. The fact that strata was a significant contributor to our model of map error could be an indication that vegetation away from roads differs from that near roads, and is governed by an environmental factors not accounted for in the other fixed effects. However, the better performance in the off-road strata could be an artifact of different quality in data collection efforts between the two strata. For example, georeferencing data in

linear clusters may have been easier or done with greater care, making off-road data less subject to positional error.

The inclusion of slope in our model might highlight the difficulty of classification in steep and often shadowy areas. Slope was not included in the initial classification model for blackbrush in the Colorado Plateau, and its inclusion might improve maps of that cover-type. The notion that classification in heterogeneous areas is tougher than in homogenous areas is not new, but our model helps determine the magnitude of heterogeneity's contribution to error. Also, numerous indices of heterogeneity are available in the literature and we illustrated that some indices may make a more significant contribution than others to models of map error. Map users might also find the model results helpful, making them more skeptical of mapped blackbrush on steeper slopes and in more heterogeneous areas.

ACKNOWLEDGMENTS

We would like to thank our reviewers for their thoughtful comments. We are also grateful to Ron Tymcio who tackled many GIS challenges, David Early for data collection, Collin Homer who served as our patient and ever-present practical conscience, and especially Scott Bassett, who can program anything, fast. This manuscript is a collaborative effort between the USDI National Biological Service's Gap Analysis Program and the USDA Forest Service Intermountain Research Station, Ogden, UT.

REFERENCES

- Edwards, T. C., Jr., G. G. Moisen, and D. R. Cutler. 1997. Assessing map accuracy in an ecoregion-scale, remotely-sensed cover map. *Remote Sensing of Environment* 60.
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427-432.
- Homer, C., R. D. Ramsey, T. C. Edwards, Jr., and A. Falconer. 1997. Landscape cover-type modelling using a multi-scene TM mosaic. *Photogrammetric Engineering and Remote Sensing* 63:59-67.
- Ludwig, J. A., and J. F. Reynolds. 1988. *Statistical Ecology*. New York: John Wiley and Sons. 337 p.
- Moisen, G. G., T. C. Edwards, Jr., and D. R. Cutler. 1994. Spatial sampling to assess classification accuracy of remotely sensed data. Pages 159-176 in W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Taylor and Francis, London.
- Omernik, J. M. 1987. Map supplement: ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77:118-125 (map).

Roberts, D. W., Cooper, S. V. 1989. Concepts and techniques of vegetation mapping. Pages 90-96 in D. Ferguson, P. Morgan, and F. D. Johnson, editors. *Land Classifications based on vegetation: applications for resource management*. USDA Forest Service General Technical Report INT-257, Ogden, Utah.

Simpson. E. H. 1949. Measurement of diversity. *Nature* 163:688.

Wolfinger, R. D. 1993. Covariance structure selection in general mixed models. *Communications in Statistics: Simulation and Computation* 22:1079-1106.

Wolfinger, R. D., O'Connell, M. 1993. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 48:233-243.