Published as:

Moisen, G. G., and T. C. Edwards, Jr. 1999. Use of generalized linear models and digital data in a forest inventory of northern Utah. Journal of Agricultural, Biological, and Environemntal Statistics 4:372-390.

USE OF GENERALIZED LINEAR MODELS AND DIGITAL DATA IN A FOREST INVENTORY OF NORTHERN UTAH

Gretchen G. Moisen And Thomas C. Edwards, Jr.

Gretchen G. Moisen is Research Forester, USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401 (E-mail: gmoisen/rmrs_ogdenfsl@fs.fed.us)

Thomas C. Edwards, Jr. is Research Ecologist and Associate Professor, USGS Biological Resources Division, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, UT 84322-5210 (E-mail: tce@nr.usu.edu).

ABSTRACT.--*Forest inventories, like those conducted by the Forest Service, Forest Inventory and Analysis Program (FIA) in the Rocky Mountain Region are under increased pressure to produce better information at reduced costs. Here, we describe our efforts in Utah to merge satellite-based information with forest inventory data for the purposes of reducing the costs of estimates of forest population totals and providing spatial depiction of forest resources. We illustrate how generalized linear models can be used to construct approximately unbiased and efficient estimates of population totals while providing a mechanism for prediction in space for mapping of forest structure. We model forest type and timber volume of five tree species groups as functions of a variety of predictor variables in the northern Utah mountains. Predictor variables include elevation, aspect, slope, geographic coordinates, as well as vegetation cover-types based on satellite data from both the Advanced Very High Resolution Radiometer (AVHRR) and Thematic Mapper (TM) platforms. We examine the relative precision of estimates of area by forest type and mean cubic-foot volumes under six different models including the traditional double sampling for stratification strategy. Only very small gains in precision were realized through the use of expensive photo-interpreted or TM-based data for stratification, while models based on topography and spatial coordinates alone were competitive. We also compare the predictive capability of the models through various map accuracy measures. The models including the TM- based vegetation performed best overall, while topography and spatial coordinates alone provided substantial information at very low cost.*

Key words: *AVHRR, generalized additive model, GIS, map accuracy, sample survey, stratification, satellite imagery, TM*

1. INTRODUCTION

Forest inventory programs, like those conducted by the USDA Forest Service, Forest Inventory and Analysis Program (FIA) in the Rocky Mountain Region, are under increased pressure to produce more information at reduced costs and increased frequency. The traditional objective of FIA has been to estimate statewide forest population totals (e.g., forest area, volume, growth and mortality) on a periodic basis. Historically, this has been accomplished through a two-phase sampling procedure where phase one consists of aerial photo-based information collected on an intensive sample grid, and phase two consists of a subset of that grid visited in the field (Chojnacky 1998). Photo-interpreted cover-type and ownership are typically used for stratification of phase two field points, resulting in improved precision of estimates of forest population totals. This strategy of combining aerial photo and field data through double sampling for stratification has been used successfully by FIA in the Rocky Mountain States for many years.

While the current two-phase sampling procedure employed by FIA provides unbiased and precise estimates of forest resources at regional scales, some problems exist. First, aerial photography available in any given state will vary in quality, scale and date. Second, some inconsistency naturally exists between different photo-interpreters in terms of correct location of sample points on photos and correct vegetation classification. Third, the process is extremely expensive and slow. Consequently, there is a need to develop methods that use satellite-based data in lieu of photo-interpretation (PI) that maintain the required precision in FIA estimates of population totals.

Moreover, the most valuable management tool to many land managers is a map depicting the spatial arrangement of forest attributes at resolutions finer than those obtainable from current FIA sampling grids. These can be difficult to generate. While vegetation cover-type maps produced by programs like the USDI Gap Analysis program (see Homer, Ramsey, Edwards, and Falconer 1997; Scott et al. 1993) have been useful to some extent in meeting the need for fine-scaled information, most, if not all, of these maps lack any spatial depiction of attributes (e.g., basal area, canopy closure, stand density) for their forest types. This reduces their utility for applications like the identification of suitable wildlife habitat (Edwards, Deshler, Foster, and Moisen, 1996), or for estimating attributes of understory components (Stenbeck and Congalton 1990), stand density and volume (Franklin 1986), or similar characteristics necessary for sound forest management.

Consequently, maps of forest inventory variables are in high demand. By linking extant inventory information with other digital, remotely-sensed biophysical information, there exists opportunity to predict and map extensive tracts of forest type and attributes at fine resolutions. Recent studies have found that integrating ancillary data, such as elevation, aspect, and slope, with spectral information from satellites (e.g., Landsat Thematic Mapper) can enhance precision for classification of forest vegetation types and attributes (e.g., Strahler and Logan 1978; Woodcock, Strahler, and Logan 1980; Frank 1988). Often this ancillary data is at fine resolution, such as the 90 m resolution Defense Mapping Agency's digital elevation models. Once

relationships between variables collected at sample points (responses) and other ancillary biophysical data (predictors) are modelled, spatially explicit maps of the response variables can be built, validated and used in management situations (Frescino, Edwards, and Moisen, in review).

While several papers in this issue address the problem of producing cost-effective estimates more frequently, in our paper we examine methods for efficient estimation and mapping at one point in time. We illustrate how generalized linear models (GLMs) can be used to construct approximately unbiased and efficient estimates of population totals while providing a mechanism for prediction in space. We model forest type and volume of several major tree species as functions of a variety of predictor variables in the northern Utah mountains. Predictor variables include elevation, aspect, slope, vegetation cover types based on two different satellite platforms, as well as spatial coordinates. We examine the relative precision of estimates of percent area by forest type and mean timber volume under six different models, including the traditional double sampling for stratification strategy. We also compare the predictive capability of the models through various map accuracy measures.

2. DATA

2.1 Study Region

Our study region was the Northern Utah Mountain Ecoregion (Omernik 1987) (Figure 1). The Ecoregion is composed of two mountain ranges: the Wasatch running north-south, and the Uinta running east-west. Elevation ranges from 1300 to 3900 m, and nearly 50% of the land is administered by National Forests. Here, topography and geographic position serve as strong surrogates for the ecological processes governing the distribution of vegetation (Cronquist, Holmgren, Holmgren, and Reveal, 1972). Precipitation is mainly a function of elevation, latitude, and storm patterns from the west and the Gulf of Mexico, with local effects from slope exposure and/or aspect (Mauk and Henderson 1984). Principal forest types are Aspen (*Populus tremuloides*), Douglas-fir (*Psudotsuga menziesii*), lodgepole pine (*Pinus contorta*), spruce-fir (*Picea engelmanni - Abies lasiocarpa*), and pinyon-juniper woodlands (*Pinus edulis - Juniperus scopulorum*).

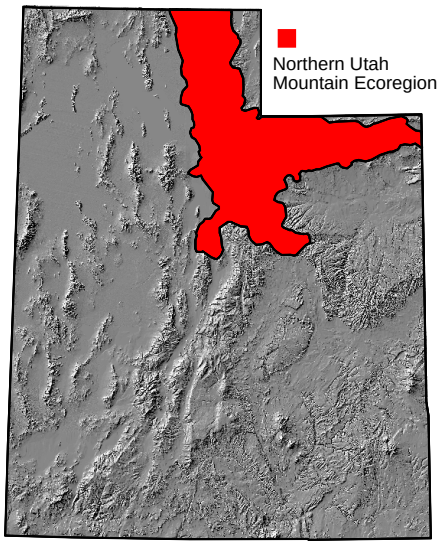

Figure 1. Northern Utah Mountain Ecoregion occupying the mountainous northeastern portion of the state.

2.2 Field and Photo-interpreted Data

The Rocky Mountain FIA program, responsible for inventorying lands in the Interior West, collected field and aerial photo data in Utah between 1991 and 1994. Under a two-phase sampling design, a Phase I sample of photo-interpreted points was collected on a 1 km grid. These points were plotted on 1:24,000 quadrangle maps and transferred to aerial photographs by

photointerpreters. An area of 1 acre (.4 ha) was examined around these points to determine the vegetative condition, or PI type. Traditionally, PI type and ownership criteria from Phase I have been used for stratification of secondary sample field plots. This results in improved precision of estimates of forest population totals, even though PI type is subject to variable misclassification rates. The Phase II was a field-visited subset of the Phase I. These 1 acre field plots were established on a 2.5 km offset grid for National Forest Lands and on a 5 km grid for other ownerships. In the proceeding analyses, only field plots on the 5 km grid were used, giving 988 plots for modelling in our study region. At each sample plot, forest variables such as site attributes, vegetation structure, tree volume, growth and mortality were measured. Details of the sample design, initial inventory estimates and analyses are reported in O'Brien (1996).

2.3 Digital Data

Predictor variables used in our analyses included elevation in meters (Elev), aspect in degrees (Asp), slope in Degrees (Slope), a vegetation class based on the .81 ha resolution (90 x 90 m grid cells) cover-map produced by Homer , Ramsey, Edwards, and Falconer (1997) through classification of Thematic Mapper (TM) imagery, a vegetation class from the 1993 Resources Planning Act forest type group map (Powell, Faulkner, Darr, Zhu, and MacCleery, 1993) based on the Advanced Very High Resolution Radiometer (AVHRR) platform, and Universal Transverse Mercator coordinates, (UTMe, UTMn). Elevation, aspect and slope were obtained on FIA field locations by intersecting the spatial location of each plot with the same location on 90 m digital elevation models produced by the Defense Mapping Agency (DMA). To avoid problems modelling with a circular predictor variable, a transformation used by Roberts and Cooper (1989) was applied to aspect, where

$$Aspect = .5\left(cos\left(\frac{\pi}{180}(Asp\text{-}30)\right) + 1\right) \quad . \qquad (1)$$

The transformed variable ranges from 0 to 1 and yields the highest values in the north-northeast direction (30°).

All digital data were coregistered and resampled at a 90 m grid to provide a constant resolution for modelling. Although resampling frequently adds error to derived maps, our selection of a 90 m resampling grid maintained the integrity of many of our predicator variables for two reasons. First, the cover-map of Utah was modelled at and is maintained at 90 m (.81 ha) resolution (see Homer, Ramsey, Edwards, and Falconer, 1997). Estimates of uncertainty in the cover-map were also obtained at this resolution (Edwards, Moisen, and Cutler, 1998). Second, Utah does not yet have complete coverage of 30 m digital elevation models, necessitating our use of 90 m DMA data for determining elevation, aspect, and slope. We recognize, like others before (Chrisman 1989, Goodchild 1989, Openshaw 1989), the unique problems posed by combining data of a diverse nature collected at different scales and at various levels of error and uncertainty, particularly when combined for predictive modelling, but acknowledge that estimating the impacts of these kinds of errors on our predictive modelling is beyond the scope of this paper.

We refer readers to works by Veregin (1988a, b) and Lunetta et al. (1991) for discussions on the impact of various sources of error on modelling.

3. ESTIMATION AND PREDICTION STRATEGIES

3.1 Double sampling for stratification

The six classes of models compared in this paper fall under one of three different estimation and prediction strategies. The first strategy (that currently used by FIA in the Interior West) relies on double sampling for stratification to produce estimates of population totals. Formulation for the estimates and their variances can be found in most sampling texts (see Cochran 1977, chap. 12) and procedures unique to forest inventory are documented in Chojnacky (1998). Predictions, or maps of the response variable of interest, are made by constructing Thiessan polygons about each 1 km PI point and assigning that polygon the stratum mean for that PI class. While useful for giving a crude depiction of the resource at the regional scale, it provides none of the detail necessary for smaller area maps and management questions because of the lack of wall-to-wall photo-interpreted data. Consequently, only estimates of population totals were made under Model-P, not predictions.

3.2 Stratified random sampling

The second strategy makes use of classified satellite data. Here, estimates of population means and totals are obtained through stratification of field plots based on AVHRR- or TM- based cover-type maps. Estimates and their variances through this stratified approach are available in most sampling texts (see Cochran 1977, chap. 5). Mapping forest variables under this strategy amounts to assigning the stratum mean to each classified pixel.

3.3 Generalized Linear Models

The third strategy relies on generalized linear models (GLMs) developed by Nelder and Wedderburn (1972) and described thoroughly in McCullough and Nelder (1989). They are an extension of the classical linear model. Suppose we have a vector ***y*** of length *n* that is assumed to be a realization of some random process **Y.** Following the notation of McCullogh and Nelder (1989), the classical linear model can be broken into 3 parts:

1. The *random component*: components of **Y** are independent and normally distributed with E(**Y**)=μ and constant variance σ^2;

2. The *systematic component*: covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ produce a linear predictor η given by

$$\eta = \sum_{1}^{p} x_j \beta_j; \tag{2}$$

3. The *link* between the random and systematic components, $g(\mu) = \eta$: $g(\mu) = \mu$.

A GLM extends the classical linear model in two ways. First, **Y** may come from any exponential family, not just the normal, and second, the link function $g(\cdot)$ may be any monotonic differentiable function, not just the identity link. As an illustration, define **Y** to be the response variable of interest (e.g., tree density, stand size class, canopy closure) with components of **Y** being independent, coming from an exponential family, and E(**Y**)=μ. Regardless of whether **Y** is continuous or discrete**,** μ can be linked to a model structure through the function $g(\mu) = \sum_{1}^{p} x_j\beta_j$.

The **x**'s may be discrete variables (e.g., remotely-sensed cover-type, precipitation zone), or continuous variables (e.g., elevation, aspect, slope, latitude, longitude). The resulting model can be used to predict the response variable over any surface

$$\hat{\mu}(x_i) = g^{-1}\left(\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right). \tag{3}$$

and the model structure itself provides a mechanism to explore ecological relationships.

An approximately unbiased estimate of the population mean (T) can be obtained by predicting the response over the entire population (N) and taking the mean,

$$\hat{T} = N^{-1}\sum_{i=1}^{N} g^{-1}\left(\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right). \tag{4}$$

While a linear model, like a regression estimator, requires only the mean value of each predictor variable in the population, the nonlinear form above does not simplify as nicely and prediction must be performed at each x_i. This can be computationally prohibitive. For example, in our study ecoregion, there are over three million pixels on which to make predictions. Our logistical solution was to subsample those pixels on a 1000 m grid and construct estimates of population means based on the ~25,000 pixel subsample. Given the approximate unbiasedness of (4), estimates of the mean square errors (variances) of these estimates of population means were computed through 10-fold crossvalidation (CV) (Ripley, 1996). In our applications, we partitioned our sample of 988 plots into ten approximately equal sized dependent random groups, fit models excluding each group in turn, and predicted over the corresponding $\mathbf{x}_i$'s of the excluded group. See Section 4.4 for further discussion of CV methods for mean squared error.

Going a step further from the traditional linear model, generalized additive models (GAMs) (Hastie and Tibshirani 1986) extend GLMs by allowing the systematic component discussed above (Equation 2) to assume the more general form

$$\eta = s_0 + \sum_{1}^{p} s_j(\mathbf{x}_j), \tag{5}$$

where $s_1, s_2, \ldots, s_p$ are smooth function estimated in a nonparametric fashion. A local scoring algorithm is used to estimate the $\hat{s}_j(\cdot)$'s . This algorithm uses scatterplot smoothers to generalize the usual Fisher scoring procedure for computing maximum likelihood estimates. A wide variety of scatterplot smoothers can be used, like running mean, running median, Loess, kernal estimate or spline (see Hastie and Tibshirani 1990 for a discussion of smoothing techniques). The use of GAMs in our work was only for exploratory data analyses and to suggest covariate transformations resulting in simplified or better GLMs as final models.

3.4 Concerns over spatial autocorrelation

When we first began our modelling exercise in Northern Utah, we constructed directional variograms on a suite of variables revealing what we would characterize as large scale spatial patterns driven largely by elevation. By fitting GLMs that included elevation, aspect, slope, and general geographic position as predictor variables, we succeeded in "detrending" our data and nothing but noise was left in directional variograms of the residuals. Perhaps the most important thing to remember is that we are modelling a response collected on a 5 km grid. In northern Utah, field plots collected at this distance are likely to change drastically in elevation, slope and aspect, and so we are primarily concerned with modelling large-scale spatial variation through a deterministic model. Our purpose in including UTM coordinates as predictor variables was to capture broad trend as opposed to trying to smooth the response in a small geographic area.

4. EXAMPLES

Six models of forest status and timber volume were constructed in the Northern Utah Mountain Ecoregion. Models and the model building processes are described below with model specifics summarized in Table 1. All analyses were performed in S-PLUS.

4.1 Double Sampling for Stratification

"Model P" mimics the traditional estimation strategy used in the Rocky Mountain FIA Unit. Recall that each point on a 1 km grid was assigned one of six PI classes (Table 1). Each field plot on the 5 km grid was stratified according to PI class and double sampling for stratification estimates produced for percent area by the following forest types: all timberland types (ALL), aspen (AS), Douglas-fir (DF), lodgepole pine (LP), and spruce-fir (SF). Mean cubic-foot volume was also estimated by the preceding species groups. Prediction, or mapping of forest resources, was not attempted under this model because of the lack of wall-to-wall coverage of PI'ed data.

4.2 Stratified Random Sampling

In "Model-A", each 90 m pixel was assigned one of 10 cover-types from the AVHRR-based cover-type map (Table 1). Field plots were stratified and estimates produced for percent area by forest type and mean cubic-foot volume by each of the 5 species groups above. Maps of forest type and volume were then produced by assigning each 1000 m pixel the mean of that AVHRR-based cover-type stratum. "Model-T" is the same as Model-A except 29 TM-based cover map classes were used for stratification (Table 1).

Table 1. Description of six models used in estimation and prediction of forest type and timber volume. The notation poly(x1,x2,3) indicates a bivariate orthogonal polynomial of order 3.

Label	Class	Predictors	Details
P	Double sampling for stratification	PI points	Strata: Conifer, Aspen, Cottonwood, Pinyon-Juniper, Misc. Hardwood, Nonforest
A	Stratified random sampling	AVHRR	Strata: Douglas-fir, Ponderosa Pine, Lodgepole Pine, Spruce/Fir, Chaparral, Pinyon-Juniper Woodland, Western Hardwood, Aspen, Nonforest
T	Stratified random sampling	TM	Strata: Water, Spruce-fir, Ponderosa Pine, Lodgepole, Mountain Fir, Juniper, Pinyon, Pinyon-Juniper, Aspen, Oak, Maple, Mountain Shrub, Sagebrush, Sagebrush/Perennial Gr, Grassland, Alpine, Dry Meadow, Wet Meadow, Barren, Lodgepole/Aspen, Ponderosa Pine/Mountain Shrub, Spruce-fir/Mountain Shrub, Mountain Fir/Mountain Shrub, Aspen/Conifer, Mountain Riparian, Lowland Riparian, Agriculture, Urban, Wetland
topo	Generalized linear model	Elev, Aspect, Slope, UTMe, UTMn	$g(\mu)=$ 1 + poly(Elev,Slope,3) + poly(Elev,Aspect,3) + poly(Slope,Aspect,3) + poly(UTMe,UTMn,3)
Atopo	Generalized linear model	AVHRR, Elev, Aspect, Slope, UTMe, UTMn	$g(\mu)=$ 1 + AVHRR poly(Elev,Slope,3) + poly(Elev,Aspect,3) + poly(Slope,Aspect,3) + poly(UTMe,UTMn,3)
Ttopo	Generalized linear model	TM, Elev, Aspect, Slope, UTMe, UTMn	$g(\mu)=$ 1 + TM poly(Elev,Slope,3) + poly(Elev,Aspect,3) + poly(Slope,Aspect,3) + poly(UTMe,UTMn,3)

4.3 Generalized Linear Models

"Model-topo" was an attempt to capture the relationship between forest type and timber volume on FIA ground plots, and corresponding satellite-based elevation, aspect, and slope, as well as geographic position in the Northern Utah Mountain Ecoregion. To explore the relationship between response and predictor variables, we began the modelling process using GAMs. In the case of forest type, all types were modelled individually as a binary response; something was either Douglas-fir or not Douglas-fir. We specified a logit link and variance function of $\mu(1-\mu)$. Specifying the appropriate link and variance function for volume was a little more difficult. Frequently, individual tree volume is modeled as a function of individual tree measurements (like diameter and height) and log transformations on the response are practical variance-stabilizing tools. In our case, though, we are modelling stand volume as a function of geographic and other site predictor variables. Naturally, many zero values exist for volume on nonforested lands, and variability is very high in forested areas. We found though exploratory analyses that the variances appeared proportional to the means (after adjustment for predictor variables), with

proportionality constants substantially larger than one. The variance of volume by species (within bins defined by combinations of predictor variables) were plotted against mean volumes of those bins revealing linearly increasing patterns. Consequently, quasilikelihood estimation was used in a "Poisson-like" model with a log link and variance proportional to the mean. Although this type of model is typically applied to counted data, McCullagh and Nelder (1994, p. 200-204) discuss an example application to continuous data.

After defining link and variance functions, several considerations went into specifying how predictor variables would be brought into the models. First, interaction exists between topographic variables in describing site suitability for tree species in Utah. For example, sites at the same elevation but different aspects typically support different tree species because of gradients in radiation, soil moisture, and other ecological factors. For that reason it makes sense to include explanatory variables jointly in a predictive model. While higher order interactions are more suitable from an ecological perspective (e.g., modelling elevation, aspect, and slope in trivariate functions), two-way interactions can be examined and scrutinized graphically. Consequently, forest type and volume by species were modelled as nonparametric functions of the following 2-way interactions: elevation and slope, elevation and transformed aspect, and slope and transformed aspect. This was accomplished through the use of local regression within a GAM (see Hastie and Tibshirani 1990 for description). Figure 2 illustrates the bivariate smooth functions resulting from the model of timber volume for all species. Recall that these GAMs were built for the purpose of suggesting suitable parametric functions to be used in the next stage of modelling.

Our next step was to develop a model containing the topographic and position variables that would be suitable for both prediction in space and estimation of population totals. Nonparametric models are more flexible than their parametric counterparts but can yield unreasonable predictions in the case of extrapolation. Our preference was for a parametric models given logistical considerations. Based on our initial GAM analysis, the final Model-topo included 3 bivariate functions (elevation and slope, elevation and transformed aspect, slope and transformed aspect) in the form of third order orthogonal polynomials. "Model-Atopo" and "Model-Ttopo" also contained the topographic and spatial predictor variables in the same polynomial forms as Model-topo, but also added the AVHRR- or TM-based classes discussed under Section 4.2 as predictors.

4.4 Evaluation Criteria

To assess the predictive performance of the models, 10-fold cross-validation (CV) (Ripley 1996) was performed. To describe the general process, one-tenth of the plots were randomly selected and withheld as an independent test set, the models fit, and predictions made over the test data only. Measures of performance were calculated based on this independent test set. Next, another one-tenth of the total number of plots were randomly selected from amongst those not yet withheld, and the process continued until all plots had been withheld once, and performance measures produced. Means over these 10 sets were then reported as final results. The process provides a mechanism by which to assess model performance without sacrificing data useful in model construction. Performance measures obtained under CV may be more optimistic than

what some independent test sets would reveal. Their primary use in this study, however, is for comparison between models evaluated under the same process.

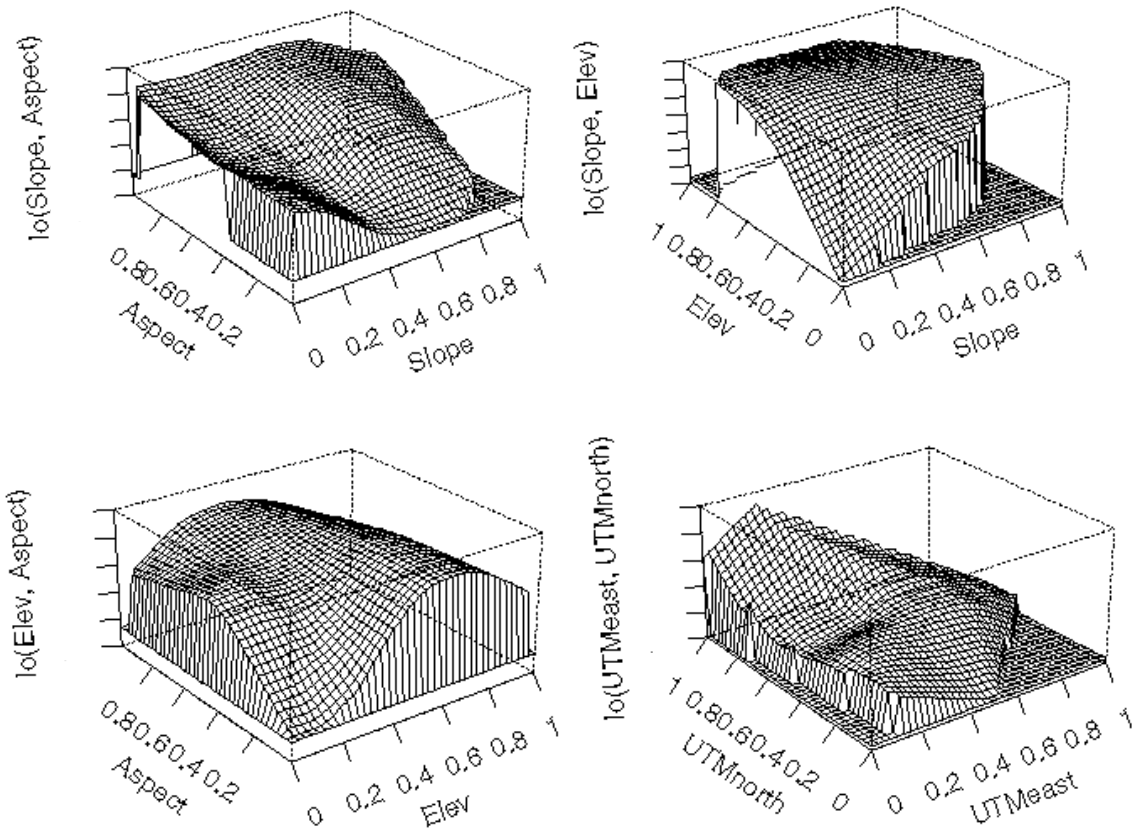

Figure 2. Bivariate smooth functions resulting from a model of timber volume as a function of elevation, slope, aspect and UTM coordinates.

To assess the precision on the estimates of population means under Models- topo, Atopo, and Ttopo, CV estimates of the mean squared error were used as estimates of the variances. These estimates were compared to those obtained through direct calculation of the variances under Models- P, A, and T. The ratio of the standard errors to the means, or coefficients of variation, were also computed under each model and species group combination. To assess the predictive performance of models of forest type, CV estimates of percent correctly classified (PCC) and the Kappa (KHAT) statistic (Cohen 1960) were calculated. The Kappa statistic measures the proportion of correctly classified units after the probability of chance agreement has been removed, and has been used extensively in map accuracy work (Congalton 1991). Predictive performance of models of the continuous volume variable was evaluated through CV estimates of root mean squared error (RMSE) and percent of plots within 500 cu ft of the predicted value. In addition, the correlation coefficient (ρ) between observed and predicted values was calculated for each model.

5. RESULTS

Estimates of the percent of land by forest type and mean tree volume by species group are illustrated in Figure 3 through approximate 95% confidence intervals under each model. Only very small gains in precision (barely detectable in these graphics) were realized through the use

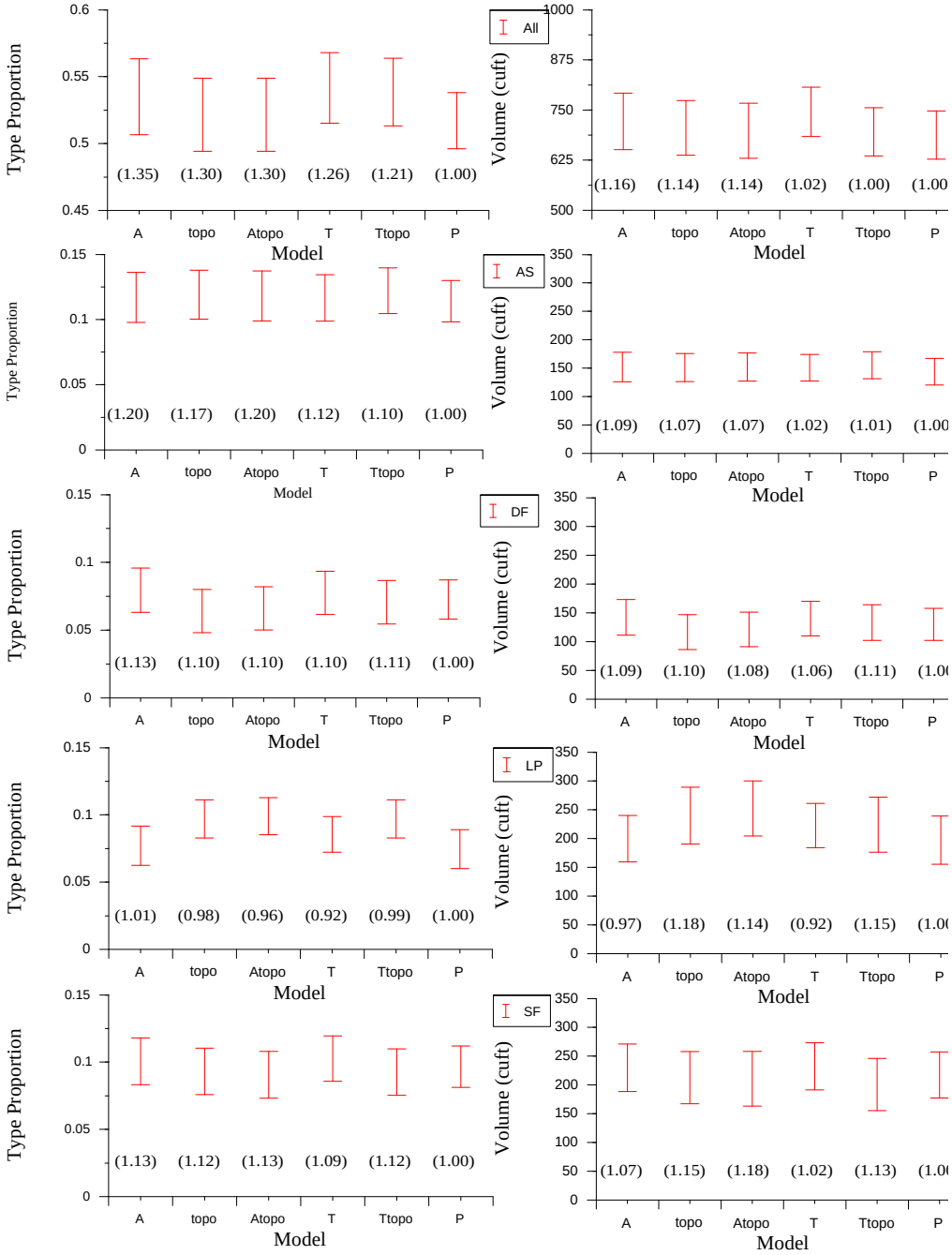

Figure 3. Ninety-five percent confidence intervals on estimates of the proportion of land by forest type and mean timber volume by species group under each of the six models. Numbers beneath each interval are the ratio of interval width to that obtained under Model P.

of expensive photo-interpreted or TM-based auxiliary data. The ratio of the interval width to that obtained using Model-P is printed below each interval.

Figure 4 illustrates predicted mean volume of all timber species for the northern Utah mountain ecoregion under Model-topo. Predictions were made based on the same 1 km grid subset of pixels used in constructing population mean estimates. Grey-scale is but one way to depict the resource in space. Other useful illustrations are color or contour maps or empirical distribution functions for the whole area or any specified subset like a 7.5 minute quadrangle or ranger district.

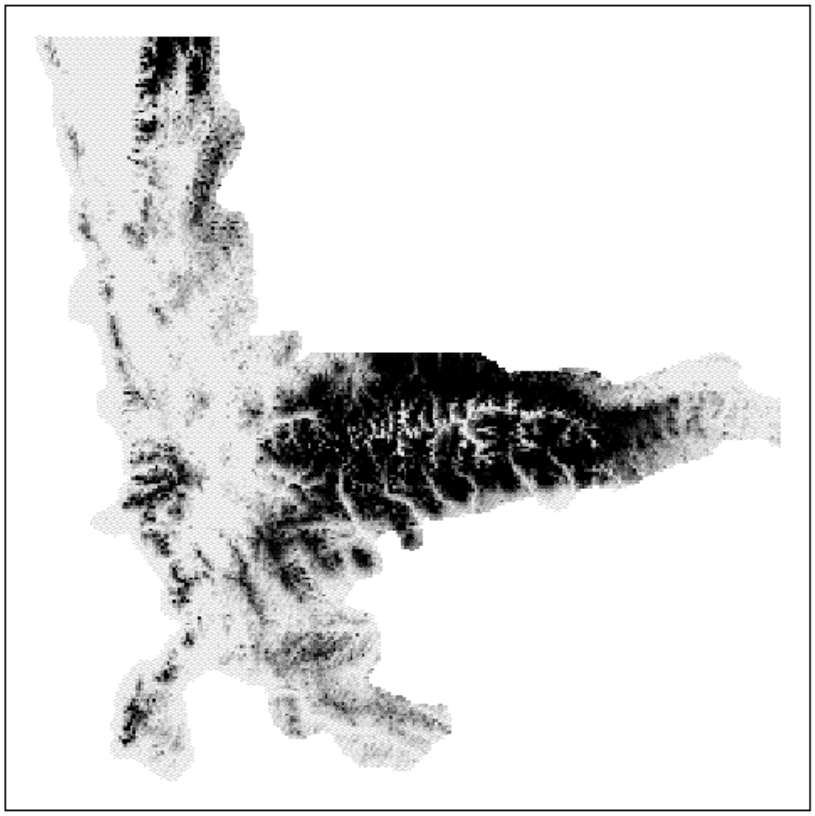

Figure 4. Predicted mean volume of all timber species for the Northern Utah Mountain Ecoregion under Model Topo. Predicted values ranged from 0 cubic ft per acre (light grey) to 7,525 cubic ft per acre (black).

Measures of how well each of the models predicted forest type and volume are illustrated in Figures 5 and 6 respectively. (Recall predictions were not made under Model-P because of the lack of wall-to-wall coverage of photo-interpreted points.) Performance in predicting forest type is measured by cross-validated PCC and KHAT, while ability to predict timber volume is evaluated through cross-validated RMSE, correlation coefficient, and percent of plots predicted within 500 cubic feet of the truth. In all cases except the RMSE plots, higher values indicate better performance. In most cases, the models including the TM- based vegetation (Models T or Ttopo) performed better than those including AVHRR-based vegetation (Models A or Atopo), though often not by amounts that would be considered large in practice. Models based solely on topography and geographic position (Model-topo) were surprisingly competitive.

6. DISCUSSION

6.1 Estimation of Population Means and Totals

A primary objective in forest inventories is to produce precise, unbiased estimates of population totals. Figure 3 illustrates how there was very little difference in the precision of estimates of both percent forested land and mean timber volume under the six different models. Stratification

using AVHRR and TM-based cover-type maps produced estimates of population totals that were only marginally less precise than those obtained using double sampling for stratification through photointerpretation. Given the availability, scale and low cost of AVHRR products, an investigation is underway to compare the relative efficiency of estimates obtained using photointerpreted versus AVHRR data throughout the diverse forest resources in the country. Another interesting result here was that Model-topo, based only on topography and geographic position, produced estimates with competitive precision levels. Again, this raises more cost savings issues.

6.2 Prediction

In terms of mapping, the relative performance of the models can be seen in Figures 5 and 6. For classification of forested versus nonforest lands, and for prediction of the grouped timber volumes, it is not surprising that models with the TM-based classifications performed the best. Models based on the AVHRR-based classes performed the worst of the suite. Again, Model-topo performed surprisingly well. This result underscores our belief that more information can be squeezed out of existing and inexpensive data sets through flexible model forms.

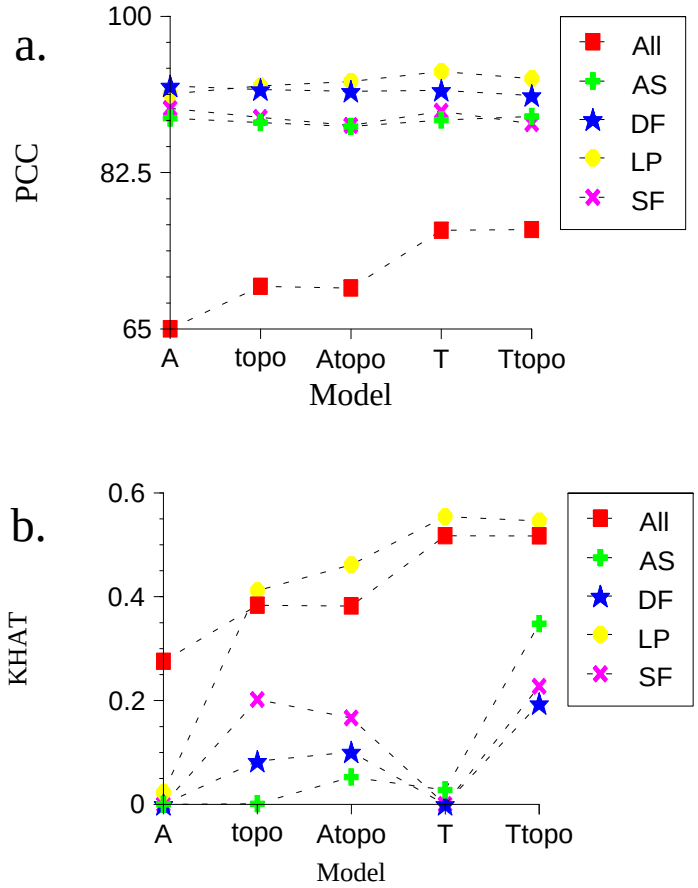

Figure 5. (a)-(b). Cross-validated PCC and KHAT values obtained for five forest type groups under each of six models.

Looking at the individual forest types in Figure 5a, it helps to remember that all types were modelled individually as a binary response; something was either Douglas-fir or not Douglas-fir. This is just the approach we chose for our simple comparisons. Modelling class as a polytomous response would make more sense in a production environment and insure that the proportion of all types summed to one. The point of this graph is that only very small gains in PCC were realized using expensive classified TM imagery over topography and geographic position alone. Certainly, other auxiliary data would be needed in less mountainous regions.

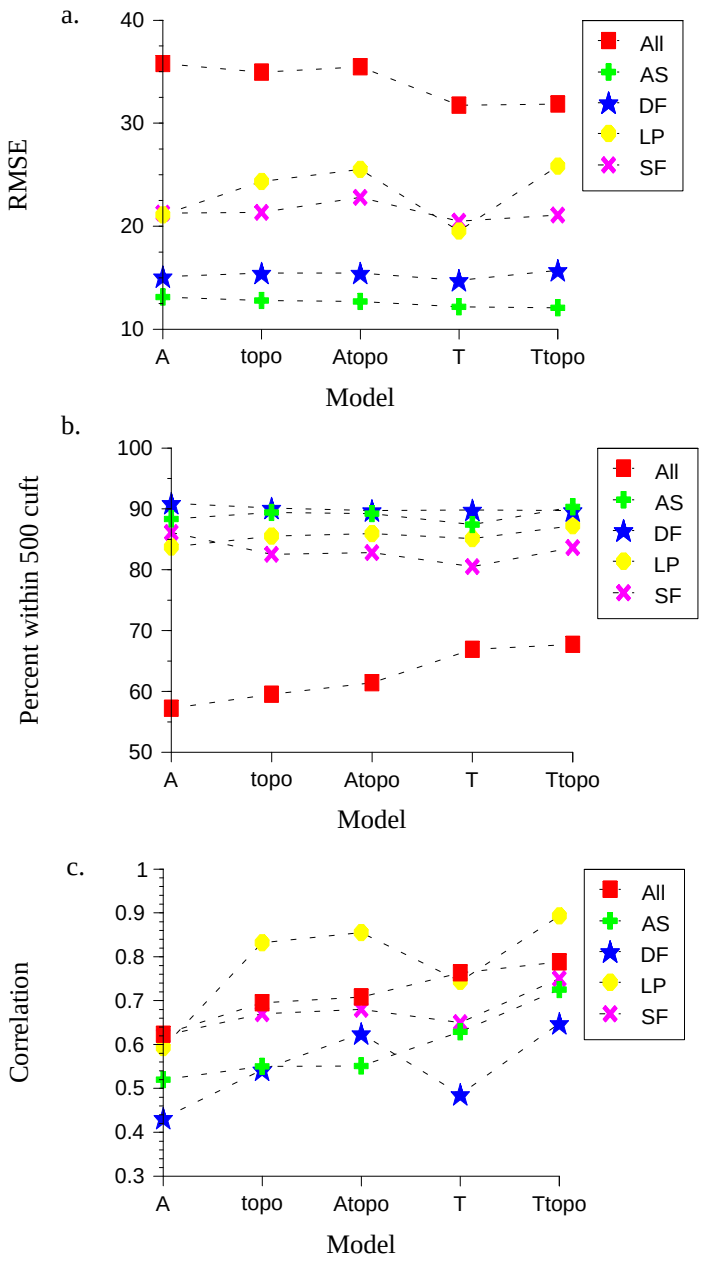

Figure 6. (a)-(c). Cross-validated RMSE, percent of plots within 500 cubic ft of predicted value, and correlation between predicted andtrue values for cubic foot volume of five species groups under each of the six models.

PCC is not the only measure of map accuracy available. In Figure 5b., the Kappa statistic provides a different picture of the predictive capability of the five models. It is interesting to note that KHAT equals zero for several individual forest types under Models A and T. Because these models simply assigned stratum means, and less than 50% of the field plots in that stratum had matching forest types, the strata means were *not* the forest type of interest. This illustrates how the addition of auxiliary variables like topography and geographic position through GLMs can improve or augment classification of existing cover-type maps.

In Figures 6b. and c., performance was measured in terms of percent of plots predicted within 500 cuft of the true value, and correlation between true and predicted values. Generally, Models Atopo and Ttopo performed better in predicting mean volumes than did the simple stratum means obtained in Models A and T, respectively. However, when performance was measured in terms of RMSE, little if any gain was realized through the addition of topography and geographic position. In the case of lodgepole pine, Atopo and Ttopo's performances were worse than their A and T counterparts because of a handful of extremely bad predictions that greatly affected the non-robust RMSE. Another point of interest in Figure 6 is the overall competitive performance of Model-topo, illustrating that topography and geography are reasonable surrogates for some vegetation driving processes and may be very useful when expensive TM-based classifications are not available.

6.3 Other Issues

Given the recent emphasis on increased data analysis capabilities in regional forest inventories, flexible models that lend themselves to exploratory work, like GAMs, may prove useful. An illustration of this is in Figure 2, the plot of fitted functions from preliminary work on Model-topo. The three dimensional plots allow for some exploration of the relationship between forest variables and topography or spatial position.

A potential application of forest structure maps derived from regional forest inventory data and satellite-based information is in the area of wildlife habitat mapping. In Frescino, Edwards and Moisen (In review) forest inventory data on seven ranger districts within this northern Utah mountain ecoregion were used to model forest class, presence of lodgepole pine, basal area of forest trees, percent cover of shrubs, and density of snags through GAMs. Special attention was paid to choosing from a wide variety of environmental gradient variables as predictors. Maps were generated over a single Ranger District and an independent set of field data collected for validating the models' performances over a small management area.

Another issue that needs to be raised is suitability of models to a production environment. In modelling the various forest response variables in this paper, exploratory work was done through GAMs and some subjective decisions made on "final" model forms. Yet, in a production environment, models must be objective and computationally reasonable. While GLMs and GAMs offer tremendous flexibility and interpretability, other modelling tools like classification and regression trees or artificial neural networks are better suited for objective prediction. A comparative discussion of potentially appropriate classes of models can be found in Moisen, Edwards, and Van Hooser (1997) and a more comprehensive investigation of a wide variety of modern regression methods for meeting multiple forest inventory objectives is underway.

7. CONCLUSIONS

In this work, we examined six different models for estimation and prediction of forest type and volume for all timber species together, aspen, Douglas-fir, lodgepole pine, and spruce-fir in the Northern Utah Mountain Ecoregion. Forest types and volumes were modelled as functions of elevation, aspect, slope, geographic position, and cover-types based on the AVHRR and TM platforms. Through this exercise we have learned several things. First, there are potentially cheaper alternatives to double sampling for stratification through photo-interpretation in the northern Utah Mountains. Precision on estimates of major forest types and major timber species volume declined very little though the use of cheaper auxiliary data like that from AVHRR or that based on topography and geographic position, under GLMs. Second, the inclusion of auxiliary variables, like those based on topography or geographic position, may substantially improve existing maps of both discrete and continuous forest response variables. Finally, the full value of digital data sets for improving forest inventory products may be better realized through more flexible statistical tools.

ACKNOWLEDGEMENTS

We would like to thank Ron Tymcio and Dennis Collins for their GIS support, Tracey Frescino for her help in developing the computing and modelling process, Richard Cutler for his statistical advice, Jim Alegria and other reviewers for their useful suggestions, and the ever present T.B. Murphy.

REFERENCES

Chojnacky, D. C., (1998), "Double Sampling for Stratification: a Forest Inventory Application in the Interior West", Research Paper RMRS-RP-7, Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, 15 p.

Chrisman, N. R., (1989), "Modelling error in overlaid categorical maps" Pages 21-34 In *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, Eds.), Philadelphia: Taylor and Francis, pp21-34.

Cochran, W. G., (1977), *Sampling Techniques*, New York: Wiley.

Cohen, J. (1960), "A Coefficient of Agreement of Nominal Scales," *Educational and Psychological Measurement*, 20(1), 37-46.

Congalton, R. G. (1991), "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data," *Remote Sensing of Environment*, 37, 35-46.

Cronquist, A., Holmgren, A. H., Holmgren, N. H., and Reveal, J. L., (1972), *Intermountain flora (Vol. 1): Vascular plants of the Intermountain West, U.S.A*, New York: Hafner Publishing Company, Inc.

Edwards, T. C., Jr., Deshler, E., Foster, D. and Moisen, G. G., (1996) "Adequacy of wildlife habitat relation models for estimating spatial distributions of terrestrial vertebrates", *Conservation Biology*, 10, 263-270.

Edwards, T. C., Jr., Moisen, G. G., and Cutler, D. R., (1998), "Assessing map uncertainty in remotely-sensed, ecoregion-scale cover-maps", *Remote Sensing and Environment*, 63, 73-83.

Frank, T., (1988), "Mapping dominant vegetation communities in the Colorado Rocky Mountain front range with Landsat Thematic Mapper and digital terrain data", *Photogrammetric Engineering and Remote Sensing*, 54, 1727-1734.

Franklin, J., (1986), "Thematic mapper analysis of coniferous forest structure and composition", *International Journal of Remote Sensing*, 7, 1287-1301.

Frescino, T. S., Edwards, T. C. , Jr., and Moisen, G. G. ,"Modelling spatially explicit forest structural attributes using generalized additive models", *Ecological Applications*, In review.

Goodchild, M. F. (1989), "Modelling error in objects and fields", In *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, Eds.), Philadelphia: Taylor and Francis, pp 107-113.

Hastie, T., and Tibshirani, R. J., (1986), "Generalized Additive Models", *Statistical Science* 1, 297-318.

Hastie, T. J., and Tibshirani, R., (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Homer, C. H., Ramsey, R. D., Edwards, T. C., Jr., and Falconer, A., (1997), "Landscape Cover-type Modelling Using a Multi-scene TM Mosaic", *Photogrammetric Engineering and Remote Sensing*, 63, 59-67.

Lunetta, R. S., Congalton, R. G., Fenstermaker, L. K., Jensen, J. R., McGwire, K. C. and Tinney, L. R., (1991), "Remote sensing and geographic information system data integration: error sources and research ideas", *Photogrammetric Engineering and Remote Sensing*, 57(6), 677-687.

Mauk, R. L., and Henderson, J. A. (1984), "Coniferous forest habitat types of northern Utah", General Technical Report INT-170, U. S. Department of Agriculture, Intermountain Forest and Range Experiment Station, Ogden, Utah.

Moisen, G. G., Edwards, T. C. Jr., and VanHooser, D., (1997), "Merging Regional Forest Inventory Data with Satellite-based Information Through Nonlinear Regression Methods", in *Proceedings of the Second International Conference on the Fusion of Earth Data*, eds. T. Ranchin and L. Wald, Sophia Antipolis, France, January 1998, p. 123-128.

McCullagh, P., and Nelder, J. A., (1989), *Generalized Linear Models*, London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M., (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society A*, 135, 370-84.

O'Brien, R.A., (1996), "Forest resources of Northern Utah Ecoregions", Resourse Bulletin INT-RB-87, Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station, 34 p.

Omernik, J. M., (1987), " Map supplement: Ecoregions of the Conterminous United States", *Annals of the Association of American Geographers*, 77, 118-125 (map).

Openshaw, S. (1989), "Learning to live with errors in spatial databases", In *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, Eds.), Philadelphia: Taylor and Francis, pp 21-34.

Powell, D. S.; Faulkner, J. L.; Darr, D. R.; Zhu, Z; and MacCleery, D. W., (1993)," Forest resourses of the United States, 1992", General Technical Report RM 234, Fort Collins, CO: U.S.

Department of Agriculture, Forest Service, Rocky Mountain Forest an Range Experiment Station. 132 p. + map.

Ripley, B. D., (1996), *Pattern Recognition and Neural Networks*, New York: Cambridge University Press.

Roberts, D. W., and Cooper, S. V. (1989), "Concepts and Techniques of Vegetation Mapping," in *Land Classifications Based on Vegetation: Applications for Resource Management*, eds. D. Ferguson, P. Morgan, and F. D. Johnson, General Technical Report INT-257, Ogden, UT: U. S. Department of Agriculture, Forest Service, Intermountain Research Station, pp. 90-96.

Scott, J. M., Davis, F., Csuti, B., Noss, R., Butterfield, B., Caicco, S., Groves, C., Edwards, T. C., Jr., Ulliman, J., Anderson, H., DErchia, F., and Wright, R. G., (1993), "Gap Analysis: A Geographic Approach to Protection of Biological Diversity", *Wildl. Monogr*. No. 123.

Stenback, J. M., and Congalton, R. G., (1990), "Using Thematic Mapper imagery to examine forest understory", *Photogrammetric Engineering and Remote Sensing*, 56, 1285-1290.

Strahler, A. H., and Logan, T. L., (1978), "Improving forest cover classification accuracy from Landsat by incorporating topographic information", in *Proceedings of the Twelfth International Symposium on Remote Sensing of Environment*, Ann Arbor, MI: Environmental Research Institute of Michigan, 927-942.

Veregin, H. (1989a), "Accuracy of spatial databases: annotated bibliography", Technical Paper 89-9, Nat. Center Geographic Infor. Analysis, Univ. Calif, Santa Barbara, Calif.

Veregin, H. (1989b), "A taxonomy of error in spatial databases", Technical Paper 89-12, Nat. Center Geographic Infor. Analysis, Univ. Calif, Santa Barbara, Calif.

Woodcock, C. E., Strahler, A. H., and Logan, T. L., (1980), "Stratification of forest vegetation for timber inventory using Landsat and collateral data", in *Fourteenth International Symposium on Remote Sensing of Environmen*t, San Jose, Costa Rica: 1769-1787.