Su Zhang 711969212
STOR890-Enrionmental Statistics
Paper Summary
04/28/2009
Paper Presented: Andrew T. Hudak et. al. "Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height". Remote Sensing of Environment 82 (2002) 397–416

The paper mainly focused on the integration of Lidar (light detection and ranging) and ETM+ data to estimate the tree height the forest. In order to estimate the tree height in the forest, the empirical statistical methods are required. In this paper, five different statistical methods are used: regression (OLS), Krigging, Cokrigging, Krigging and Cokrigging on the residuals from Regression. The research have two objectives: the first is to to estimate canopy height at locations unsampled by lidar, based on the statistical and geostatistical relationships between the lidar and Landsat ETM+ data at the lidar sample locations. The five methods are tested respectively and the accuracy is estimated. The second objective was to determine which spatial sampling pattern (contiguous transect or discrete point) and frequency (2000, 1000, 500, or 250 m) would optimize the integration of lidar and Landsat ETM+ data for accurately estimating and mapping canopy height in intensively managed, coniferous forest landscapes. Because the range and scale of a landscape can be characterized by the semivariogram, the different sampling frequency may produce different results compared with the scale of the forests. The sampling strategy is also important because it is best to get the optimal result by using the least sampling points. There should be some tradeoff between the sampling points and estimation accuracy.

Methods used in this study

The ordinary least square (OLS) regression is an aspatial method because it assumes spatial independence in the data. Users of regression models for mapping vegetation attributes must be mindful of this aspect, since geographic data are often autocorrelated. Therefore, it is possible that the regression cannot preserve the spatial pattern of the measured height distribution. The OLS multiple regression model takes the general form:
$Z = \alpha + \beta_i (X_i) + \varepsilon$
Where, in this study, Z is the dependent variable (height), X_i is the *i* explanatory variable (Landsat Bands 1–7 and UTMX and UTMY locations), β_i is the linear slope coefficient corresponding to X_i, α is the intercept, and ε is the residual error. The independent variables include the 7 bands and the geographic location information.

Ordinary Kriging (OK) is also used in the study. Spatial models are appropriate if there is spatial dependence in the data, as in this study. Kriging interpolates the sample data to estimate values at unsampled locations, based solely on a linear model of regionalization. The linear model of regionalization essentially is a weighting function required to krig and can be graphically represented by a semivariogram. The semivariogram plots

semivariance γ as a function of the distance between samples, known as the lag distance h, according to:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [Z(u_\alpha) - Z(u_\alpha + h)]^2$$

where $\gamma(h)$ is semivariance as a function of lag distance h, N(h) is the number of pairs of data locations separated by h, and z is the data value at locations u_α and $u_\alpha + h$. The sample semivariogram describes the spatial autocorrelation unique to that sample dataset. Many functions have been developed to simulate the shape of the semivariogram. In this study, the exponential models are used, which take the form:

$$\gamma(h) = c[1 - \exp(\frac{-3h}{a})]$$

where *a* is the practical range of the semivariogram, defined as the lag distance at which the model value is at 95% of the sill *c*, as the exponential model approaches the sill asymptotically.

The OK model used in this study estimates a value Z* at each location u and takes the general form:

$$Z^*(u) = \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) Z(u_\alpha)$$

where Z is the primary variable and λ_α and u_α are the weights and locations, respectively, of n neighboring samples. The OK estimator allows for a locally varying mean by forcing the kriging weights to sum to one:

$$\sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) = 1$$

Cokriging is a multivariate extension of kriging used in this study. The cross-semivariogram can defined as:

$$\gamma_{ij}(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [Z_i(u_\alpha) - Z_i(u_\alpha + h)] \times [Z_j(u_\alpha) - Z_j(u_\alpha + h)]$$

where $\gamma_{ij}(h)$ is the cross-semivariance between variables i and j, N(h) is the number of pairs of data locations separated by lag distance h, Z_i is the data value of variable i at locations ua and (ua + h), and Z_j is the data value of variable j at the same locations. With the primary variable Z and a single secondary variable Y, the OCK estimator of Z* at location u takes the form:

$$Z^*(u) = \sum_{\alpha_1=1}^{n_1(u)} \lambda_{\alpha_1}(u) Z(u_{\alpha_1}) + \sum_{\alpha_2=1}^{n_2(u)} \lambda_{\alpha_2}(u) Z(u_{\alpha_2})$$

where λ_{α_1} and u_{α_1} are the weights and locations, respectively, of the n_1 primary data, and λ_{α_2} and u_{α_2} are the weights and locations, respectively, of the n_2 secondary data. In this study, we used the traditional OCK estimator, which operates under two nonbiased constraints:

$$\sum_{\alpha_1=1}^{n_1(u)} \lambda_{\alpha_1}(u) = 1; \sum_{\alpha_2=1}^{n_2(u)} \lambda_{\alpha_2}(u) = 0$$

In this study, the primary data is the Lidar data and the secondary data is the Landsat ETM+ panchromatic data.

The histogram of the maximum canopy height data exhibited a strong positive skew. Each of the eight height datasets has to be normalized with a square root transformation (SQRTHT) prior to applying any of the estimation methods. The SQRTHT sample data were normal score transformed prior to modeling. This nonlinear, ranked transformation normalizes the data to produce a standard Gaussian cumulative distribution function with mean equal to zero and variance equal to one.

The SQRTHT sample data were regressed on the raw ETM+ Bands 1–7, as well as the Universal Transverse Mercator (UTM) X and Y locations, using stepwise multiple linear regression.

Residuals from the OLS regression models were exported from IDL as ASCII files and imported into GSLIB for kriging/cokriging. The same rules and procedures were followed for modeling the residuals as for modeling the SQRTHT data.

Results

For empirical models, from Table 1, the ETM+ band 7 was the first variable selected. For the spatial and integrated models, unique semivariogram models of the height and height residual datasets were generated for all eight of the sampling strategies tested.

For the global accuracy, histograms of the full populations of estimated height values were used to evaluate global accuracy. Deviations in the estimated height histograms away from the measured height histogram were a good indicator of estimation biases at various heights. From Fig 5, it was clear that correlations between measured and estimated heights were always better using the integrated models than using either the regression or spatial models alone (r=0.94).

From the scatter plots of measured vs. estimated height values, we can see that the regression models suffered the most from underestimating the taller heights while overestimating the shorter heights. We can also find that the r is almost the same for the whole data set and for the scatter plots. Therefore, the 700 points in these scatter plots were highly representative of the full population of height estimates, and their errors.

For the local accuracy, accuracy decreased as the distance from sample locations increased (Fig. 7). The spatial models were more accurate than the regression models below distances of approximately 200 m from the sample locations.

From fig.8, the mapping of the whole forest is depicted, for Kriging and Cokriging method, the apparent artifact is shown. For regression method, the spatial pattern is well preserved no matter what sampling strategy and frequency is chosen. From the residual map in Fig.9, All regression models, and all models derived from the two sparser point

sample datasets (2000 and 1000 m), failed to remove the spatial dependence from the residuals (Table 3) because the Moran's I statistic is significant.

Compared with the 5 models, the integrated methods proved superior because they preserved the spatial pattern in canopy height, like the regression models (Fig. 8), while also improving global and local estimation accuracy, like the spatial models (Figs.5–7). They have no apparent disadvantage relative to aspatial or spatial methods alone (Table 3). Therefore, we can choose the integrated methods in this study as the best menthod.