Post-Modeling Histogram Matching of Maps Produced Using Regression Trees

Andrew J. Lister[1] **and Tonya W. Lister**[2]

Abstract.—Spatial predictive models often use statistical techniques that in some way rely on averaging of values. Estimates from linear modeling are known to be susceptible to truncation of variance when the independent (predictor) variables are measured with error. A straightforward post-processing technique (histogram matching) for attempting to mitigate this effect is presented, and a comparison with untransformed model estimates is made. Histogram matching enhanced the contrast visible in the final map and produced estimates that mimicked the range of estimates in the original data set but performed worse overall with respect to absolute error of prediction. Examples of cases where histogram matching might be an effective post-processing method are given.

Introduction

Advances in computer software and hardware have increased the prevalence of spatial predictive modeling of ecological data. Modeling methods range from simple spatial interpolation to sophisticated linear and nonlinear multivariate techniques. Most of the techniques that are commonly applied rely on averaging procedures. For example, simple linear interpolation generally involves defining a search radius around areas to be estimated and applying the average value of the attribute of known data within that radius to the area. Similarly, linear modeling approaches such as regression rely on averaging of deviations from a "best fit" line to arrive at parameter estimates. In nearly all cases, either the estimate itself is an average or part of the parameter estimation process is based on averaging. This trait of linear modeling can lead to a compression of the variance of the set of estimates if the independent (predictor) variables are measured with error (Curran and Hay 1986).

In the linear regression context, the predictor variables are assumed to be measured without error (Montgomery and Peck 1982). Curran and Hay (1986), however, provide a concise review of reasons why this assumption is not true in remote sensing studies. The effects of errors in the predictor variables in multiple regression have been well documented (Curran and Hay 1986, Whitemore and Keller 1988, Elston *et al.* 1997). Generally, the parameter estimates are underestimated, leading to an underestimation of large values and an overestimation of small values (Curran and Hay 1986, Cohen *et al.* 2003), resulting in a compression of the variance of the set of estimates relative to that of the training data.

In this article, satellite and other geospatial data are used to predict biomass (megagrams aboveground dry biomass [not including foliage] per hectare) in Maine, using U.S. Department of Agriculture (USDA) Forest Service Forest Inventory and Analysis (FIA) plot information as training data. FIA has a network of inventory plots across the country and collects and reports information on the status of and trends in the Nation's forest resources (Gillespie 1999). Due to the nature of the data sets (very small plot area to pixel area ratios, smoothed predictor data, lack of strong functional relationships between the dependent and predictor data), the predictions obtained by a regression tree approach (Quinlan 1993) had truncated variance. This article describes a post-processing technique—histogram matching—that attempts to deal with this systematic overprediction in the lower tail and underprediction in the upper tail of the distribution of training data.

[1] Research Forester, U.S. Department of Agriculture (USDA), Forest Service, Northeastern Research Station, Forest Inventory and Analysis, Newtown Square, PA 19073. Phone: 610–557–4038; e-mail: alister@fs.fed.us.
[2] Research Forester, USDA Forest Service, Northeastern Research Station, Forest Inventory and Analysis, Newtown Square, PA 19073. Phone: 610–557–4034; e-mail: tlister01@fs.fed.us.

Methods

Data from 2,210 FIA plots collected in Maine between 1999 and 2003 were used in the study.[3] The distribution of plots in the study area is based on a hexagonal tessellation with one FIA plot randomly located in each 2,428.2-ha hexagon. Each FIA plot consists of four circular 14.6-m (48-ft)-diameter subplots, with one subplot located in the center and three equidistant subplots distributed symmetrically around and located 31.6 m (120 ft) from the center subplot. The subplots occupy 0.07 ha (0.17 ac), and the subplot array can be subtended by a circle of 0.4 ha (1.0 ac) in area. The value of the total aboveground dry biomass was calculated from live tree data collected on each plot using equations found in Wharton *et al.* (1997).

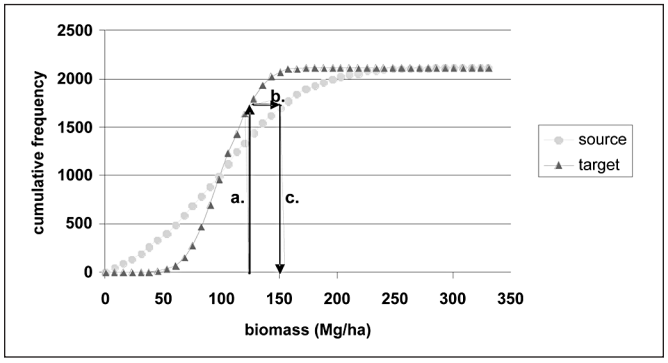
The predictor data used were contained in a multilayered ERDAS IMAGINE image and consisted of 271 250-m resolution layers, including multidate and monthly composites and derived indexes of imagery from the Moderate Resolution Imaging Spectroradiometer (MODIS)-satellite-borne sensor (Justice and Townsend 2002), several rasterized summaries of the State Soil Geographic soils database compiled by the USDA Natural Resources Conservation Service (1994), summaries of the land cover classes found in the National Land Cover Data (NLCD) database (Vogelmann *et al.* 2001), mean monthly and annual temperature and precipitation from the PRISM climate database (Daly *et al.* 2004), rasterized Bailey's Ecoregions (Bailey 1996) and U.S. Geological Survey NLCD 2001 mapping zone (which is similar to an ecoregion map) (Homer and Gallant 2001), a rasterized grid representing distance to streams (U.S. Geological Survey 1999), and various derivatives of the National Elevation Dataset (Gesch *et al.* 2002).[4]

Leica Geosystems' ERDAS IMAGINE image processing software was used to extract values for each of the predictor layers at the locations where the 2,210 FIA plots used in the analysis were located. Cubist regression tree software was used to derive regression tree models of forest biomass. These models then were applied to the stack of predictor layers in ERDAS IMAGINE to create a set of spatially referenced model predictions. To create a validation data set, 220 plots were randomly withheld from the Cubist modeling.

Histogram matching was applied to the output map generated by Cubist and ERDAS IMAGINE so that its frequency distribution of pixel values matched that of the training data. In the histogram-matching technique, a lookup table that specifies the relationship between the cumulative distribution function of a source histogram and a target histogram is generated. Using that lookup table, the target data are transformed so that their distribution matches that of the source data. (For a description of histogram matching. see fig. 1.)

Histogram matching is typically used to standardize raw satellite images that were acquired at different times and under different conditions to facilitate mosaicking of the images or classifications

Figure 1.—*Example of histogram matching applied to the biomass estimates produced from Cubist modeling in Maine. To generate a lookup table, the cumulative frequency histograms of the actual data (source) and the predictions (target) were generated. A lookup table was generated as follows: for a given biomass value in the actual data (e.g., 125 Mg/ha below), the cumulative frequency was determined (a) and related to the corresponding cumulative frequency (b) and value (e.g., 150 Mg/ha below) (c) of the predicted data. The lookup table built in this manner was used to reclassify the map of predictions so that its frequency distribution matched that of the original data.*



[3] U.S. Department of Agriculture (USDA), Forest Service. 2000. Field data collection procedures for phase 2 plots. In: Forest inventory and analysis national core field guide, version 1.4. Internal report on file at: USDA Forest Service, Washington Office, FIA, 1601 North Kent Street, Suite 400, Arlington, VA 2220. Vol. 1.
[4] Complete details of the steps used to prepare the data and data derivatives are on file at the USDA Forest Service, Northeastern Research Station, 11 Campus Blvd, Ste. 200, Newtown Square, PA 19073.

conducted on the images (e.g., Homer *et al.* 1997). Cohen *et al.* (2001) used histogram matching as a post-processing approach to facilitate the post-classification juxtaposition of images acquired under different conditions. In the current study, however, we resorted to a post-processing approach for pragmatic reasons. FIA has currently chosen to operationally use a modeling protocol that relies on the Cubist regression tree approaches, which can produce outputs with truncated variance (Curran and Hay 1986).

Graphs of biomass on the 221 test plots versus the predicted values were produced, and the relationship between the set of actual and both sets of predicted values (uncorrected and histogram-matched) was described using a simple, first-order linear regression equation and associated coefficient of determination (R^2). In addition, mean absolute error (MAE) was calculated. To assess the efficacy of histogram matching with respect to removing the overestimation and underestimation that occurred in the tails of the distribution of original data, the slope of the regression, the R^2 and the MAE were compared.

In addition to global comparisons, a multiscale analysis of the mean absolute difference between the average biomass of the plots found in sets of grid cells superimposed over the area and the pixel-based estimates in those cells was calculated for each of a number of spatial scales (fig. 2). The goal of this analysis was not only to characterize the spatial agreement between the pixel-based estimates and the actual values, but also to reveal the scale at which biomass varied across the landscape.

Results and Discussion

Figure 3 shows a map of the uncorrected and histogram-matched biomass estimates. The estimates depicted in figure 3a (the uncorrected data) show less variability than those shown in figure 3b (the histogram-matched data). The histograms of estimates support this (fig. 4)—the range of values found in the uncorrected estimates is much narrower than that of the histogram-matched estimates and the actual plot data. In many situations, duplication of the variability found in the actual data is a desirable trait of a map of modeled estimates because it makes a map of the estimates more useful.

Table 1 gives a comparison of the simple linear regression parameters, the MAE and the R^2. Figure 5 depicts scatterplots of the actual versus uncorrected prediction and actual versus histogram-matched prediction. The parameters (slope and y intercept) and diagnostic information (MAE) from the simple linear regression analyses indicate that the histogram-matching procedure performed better with respect to producing estimates

Figure 2.—*Square grid cells of varying sizes (15 x 15, 19 x 19, 22 x 22, 25 x 25, 35 x 35, 50 x 50, 71 x 71, 87 x 87, and 100 x 100 km) were superimposed on the study area. In each grid cell, the mean value for biomass on the plots and that for the pixel-based estimates was determined, and differences were calculated and compared across different window sizes.*

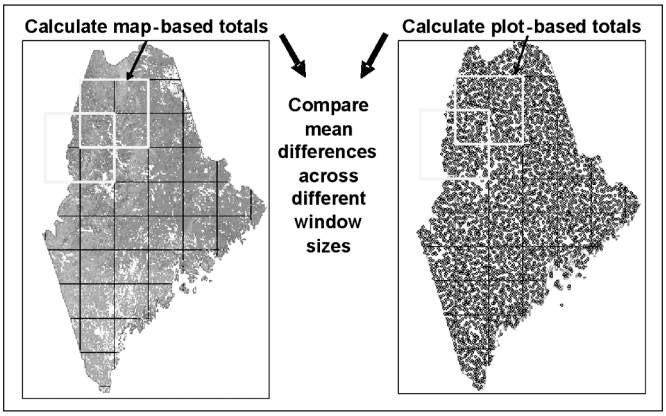


Figure 3.—*(a) Biomass map produce by Cubist modeling without post-processing. (b) Map produced by histogram matching of the map in (a) so the frequency distribution of pixel values matched that of the training data set.*

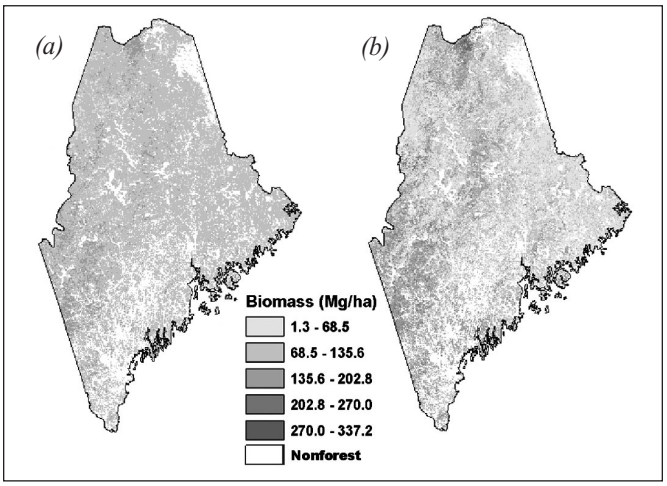


Figure 4.—*Frequency histograms of the training (actual) data set, the original prediction, and the histogram-matched (post-processed) prediction. The variance of the original predictions is truncated, whereas that of the histogram-matched data matches that of the actual data.*

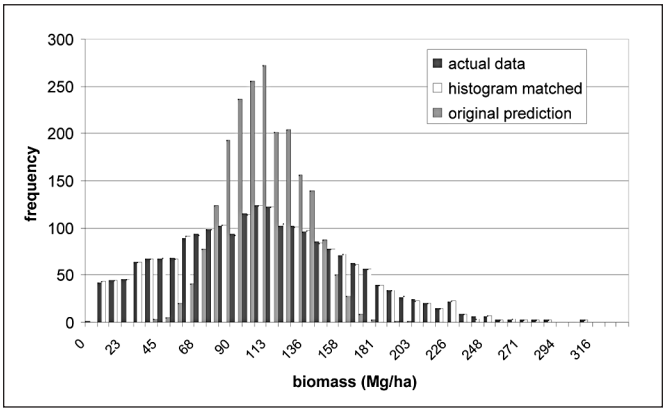
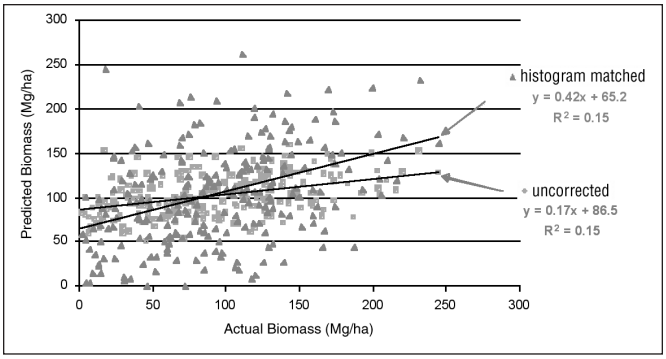


Figure 5.—*Scatterplots of the original (red circles) and histogram-matched (green triangles) values versus the validation data withheld from the Cubist modeling (n = 221). Simple linear regression lines describing the scatterplots indicate that the set of histogram-matched predictions more closely follow a 1:1 line than do the uncorrected estimates.*



that followed a 1:1 observed versus predicted line. The uncorrected estimates, however, performed better overall (the MAE of the histogram-matching data was 20 percent higher than that of the uncorrected data). An analysis of the MAE per decile of the actual data (table 2) indicates that the histogram-matching procedure only performed better in the lowest and highest deciles because the procedure stretched out the distribution of predictions and thus lowered the variance in the tails of the distribution. This stretching led to the poorer performance of histogram matching in deciles near the median, however, because a wide range of uncorrected values was transformed into output values close to the median (note the steep slope of the target histogram near the median [fig. 1]). If maintenance of the full range of variability of actual data is wanted, histogram matching appears to be a valid option. The tradeoff, however, is overall lower accuracy for the resulting set of estimates.

Table 1.—*Overall comparison of MAE of the uncorrected (original) and the histogram-matched (post-processed) predictions.*

	Uncorrected	Histogram-matched
MAE	39.70	47.70
slope	0.17	0.42
intercept	86.50	65.20
R^2	0.16	0.15

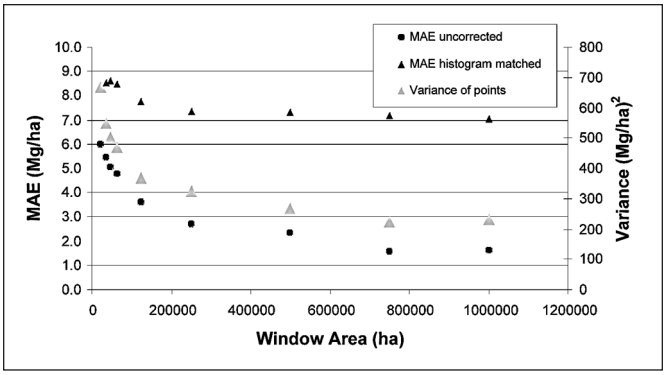
Table 2.—*Per decile comparison of the simple linear regression parameters and diagnostics of the uncorrected (original) and the histogram-matched (post-processed) predictions.*

Percentile	MAE uncorrected	MAE histogram-matched
10	75.2	56.2
20	62.3	61.8
30	38.9	39.3
40	33.3	42.0
50	22.8	43.4
60	17.6	42.5
70	17.3	41.2
80	22.4	40.7
90	33.7	49.9
100	71.8	59.5

Figure 6 shows results of the multiscale analysis of agreement between plot-based and pixel-based estimates. The figure indicates that a similar pattern of decreasing MAE as the size of the analytical windows increased from 23,400 to 1,000,000 ha. Although the uncorrected data set agrees much better with the FIA plots at each scale, the patterns of decrease are the same, suggesting that the relative interpretability of the spatial distribution of biomass on both maps at different resolutions is similar.

The variance of the window-based means (calculated across all windows at each resolution) (fig. 6) shows a decrease in plot variance as the analytical window increases in size. Because the MAE follows a very similar pattern, the decline in MAE with

Figure 6.—*MAE (absolute value of average biomass/ha of FIA plots—average biomass/ha of pixels) computed for the uncorrected and the histogram-matched estimates at several scales. The variance of the window-based means of the plots is shown for each window size is also shown (yellow triangles, secondary y axis).*



increasing window size is probably driven by the convergence of the window-based means on the global mean as the analytical window size approaches the size of the entire study area. In other words, as the window size increases, the number of plots per window increases, the variance of the biomass decreases, and the mean biomass in the windows approaches the global mean. Because the maps of predictions reflect the variability of biomass across the landscape, they follow this same pattern.

A notable feature of figure 6 is the rate of decrease in MAE (and plot-based variance) as window size increases. At a certain spatial scale (approximately 500,000 ha), increasing window size leads to only a small corresponding dip in MAE or variance, suggesting that this is similar to the scale of spatial autocorrelation of biomass in this area (Isaaks and Srivastava 1989). This information could be useful when designing sampling protocols or studying regional scale trends in biomass-related processes.

Ideally, FIA would either adapt its modeling approach or use input data that are less prone to measurement error. For example, the resolution mismatch between the plot data and the predictor data could be addressed. Statistical methods other than regression trees, which rely in part on linear modeling, also could be used. Future research could thus involve using more robust techniques and predictor data with more of a functional relationship with the plot data. Practically speaking, however, FIA currently cannot produce regional- and national-scale maps without implementing the Cubist-ERDAS IMAGINE approach and without using coarse-scale data. The Cubist- ERDAS IMAGINE approach offers many benefits, including computational efficiency, ease of use, and batch processing options. Similarly, the coarse-scale data are inexpensive, easily acquired, and manageable in terms of storage space and processing requirements. On the other hand, FIA's methods do not leave much room for altering the modeling protocol, hence the appeal of a straightforward post-processing attempt to correct systematic errors.

This article illustrates the use of histogram matching as a post-processing method. The technique did not improve the overall accuracy of the map but could contribute to other potential map uses. It allowed for the transformation of biomass predictions so that their frequency distribution matched that of a target distribution (herein, that of the FIA data). It created a map that showed more contrast, revealing the visual spatial pattern in the data set more effectively than the uncorrected estimates. It performed worse overall in terms of MAE but performed better at making predictions in the tails of the distribution of training data. This predicting ability can be a desirable trait for a modeling data set. For example, if rare events are of interest, the depiction of predictions that fall into the rare category, such as areas with biomass greater than 300 Mg/ha, might aid the land manager seeking to identify areas with old growth forest or unique ecological characteristics. It also helps the map user understand the overall spatial pattern of biomass across the landscape. The disadvantage of the histogram-matching approach is the potential loss of reliability of the map as a whole. In the case of biomass maps, the specific interests of the user should thus be taken into account when selecting a modeling technique and any post-processing methods.

Literature Cited

Bailey, R.G. 1996. Ecosystem geography. New York: Springer-Verlag, Inc. 204 p.

Cohen, W.B.; Maiersperger, T.K.; Gower, S.T.; Turner, D.P. 2003. An improved strategy for regression of biophysical variables and Landsat ETM+ data. Remote Sensing of Environment. 84(4): 561–571.

Cohen, W.B.; Maiersperger, T.K.; Spies, T.A.; Oetter, D.R. 2001. Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. International Journal of Remote Sensing. 22(12): 2279–2310.

Curran, P.J.; Hay, A.M. 1986. The importance of measurement error for certain procedures in remote sensing at optical wavelengths. Photogrammetric Engineering & Remote Sensing. 52(2): 229–241.

Daly, C.; Gibson, W.P.; Doggett, M.; *et al*. 2004. Up-to-date monthly climate maps for the conterminous United States. In: Proceedings, 14th AMS conference on applied climatology; 2004 January 13–16; Seattle, WA: American Meteorological Society [CD-ROM].

Elston, D.A.; Jayasinghe, G.; Buckland, S.T.; *et al*. 1997. Adapting regression equations to minimize the mean squared error of predictions made using covariate data from a GIS. International Journal of Geographical Information Science. 11(3): 265–280.

Gesch, D.; Oimoen, M.; Greenlee, S.; *et al*. 2002. The National Elevation Dataset. Photogrammetric Engineering & Remote Sensing. 68(1): 5–32.

Gillespie, A.J. 1999. Rationale for a national annual forest inventory program. Journal of Forestry. 97(12): 16–20.

Homer, C.; Gallant, A. 2001. Partitioning the conterminous United States in mapping zones for Landsat TM land cover mapping. U.S. Geological Survey draft white paper. On file at: U.S. Geological Survey EROS Data Center, 47914 252nd Street, Sioux Falls, SD.

Homer, C.; Ramsey, R.; Edwards, T.C., Jr.; Falconer, A. 1997. Landscape cover type modeling using a multi-scene Thematic Mapper mosaic. Photogrammetric Engineering & Remote Sensing. 63: 59–67.

Isaaks, E.H.; Srivastava, R.M. 1989. Applied geostatistics. New York: Oxford University Press. 561 p.

Justice, C.O.; Townshend, J. 2002. Special issue on the moderate resolution imaging spectroradiometer (MODIS): a new generation of land surface monitoring. Remote Sensing of Environment. 83(1): 1–2.

Montgomery, D.C.; Peck, E.A. 1982. Introduction to linear regression analysis. New York: John Wiley & Sons. 504 p.

Quinlan, J.R. 1993. Combining instance-based and model-based learning. In: Utgoff, P., ed. Proceedings, 10th international conference on machine learning; 1993 June 27–29; Amherst, MA. San Francisco, CA: Morgan Kaufmann Publishers: 236–243.

U.S. Department of Agriculture (USDA), Natural Resources Conservation Service (NRCS). 1994. State soil geographic (STATSGO) database: data use information. Misc. Publication 1492. Washington, DC: NRCS, National Soil Survey Center. 88 p.

U.S. Geological Survey. 1999. The National Hydrography Dataset. United States Geological Survey fact sheet 106-99. Reston, VA: U.S. Geological Survey. 2 p.

Vogelmann, J.E.; Howard, S.M.; Yang, L.; *et al*. 2001. Completion of the 1990s National Land Cover Dataset for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering & Remote Sensing. 67(6): 650–662.

Wharton, E.H.; Alerich, C.L.; Drake, D.A. 1997. Estimating total forest biomass in New York, 1993. Resource Bulletin NE-139. Radnor, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 48 p.

Whitemore, A.S.; Keller, J.B. 1988. Approximation for regression with covariate measurement error. Journal of the American Statistical Association. 83(404): 1057–1066.