Forest biomass estimation over regional scales using multisource data

A. Baccini,[1] M. A. Friedl,[1] C. E. Woodcock,[1] and R. Warbington[2]

Received 21 February 2003; revised 21 February 2004; accepted 16 April 2004; published 19 May 2004.

[1] A combination of statistical models and multisource data were used to map above-ground forest biomass for National Forest lands in California. To do this, data from the Moderate Resolution Imaging Spectroradiometer were used in combination with precipitation, temperature, and elevation data. The results show that coarse resolution remotely sensed data in combination with relevant topographic and climate data can be used to map above-ground biomass with good accuracy over large areas. For the data sets considered, empirical models based on a 2 percent sample explained 73 percent of the variance in biomass in the remaining 98 percent of the data with a root mean square error of 44.4 tons/ha. These results suggest that it should be feasible to improve estimates of above-ground carbon stocks at regional to continental scales in the near future. *INDEX TERMS:* 0933 Exploration Geophysics: Remote sensing; 1640 Global Change: Remote sensing; 1615 Global Change: Biogeochemical processes (4805). **Citation:** Baccini, A., M. A. Friedl, C. E. Woodcock, and R. Warbington (2004), Forest biomass estimation over regional scales using multisource data, *Geophys. Res. Lett.*, *31*, L10501, doi:10.1029/2004GL019782.

1. Introduction

[2] Incomplete information regarding the spatial distribution of carbon stored in biomass introduces substantial uncertainty to current estimates of the global carbon budget [*Brown and Schroeder*, 1999]. Much of this uncertainty is attributable to poor knowledge of forest biomass [*Schroeder et al.*, 1997]. For example, differences in estimates of biomass stored in Brazil's Amazonian forest vary by a factor of 2 (from 39 PgC to 93 PgC) [*Houghton et al.*, 2001], and estimates of carbon emissions caused by tropical land use change in 1990 vary from 1.2 to 2.2 Pg C yr^{-1} [*Houghton*, 1992].

[3] In this paper, we describe research that uses a combination of data sources to map above-ground forest biomass for eighteen National Forests in California. Estimation and mapping of biomass at this scale presents considerable technical and logistical challenges using field-based methodologies. Our results show that by using a combination of remotely sensed data, topographic information, and climate variables it is possible to map forest biomass over regional scales with good accuracy.

2. Background

2.1. Forest Inventories

[4] Collection of field data to estimate biomass generally involves destructive sampling [*Brown*, 2002]. Such procedures are time consuming and very expensive. As a result, forest biomass data are relatively rare, and when available, tend to be representative of small areas and local conditions [*Schroeder et al.*, 1997]. More commonly, forest biomass is estimated using timber volume information collected through forest inventories. Such inventories employ statistical sampling using field plots, where forest parameters (i.e, tree species, height, diameter at breast height) are measured directly. Conversion of timber volume to above-ground biomass is then accomplished by applying a biomass expansion factor (BEF) to the timber volume data. In the United States, the USDA Forest Service Forest Inventory and Analysis (FIA) is the major source of timber volume and is used in this work to calibrate and test empirical models using remote sensing and other data sources to predict forest biomass.

2.2. Remote Sensing

[5] Remotely sensed data have been used to map land cover, land use change, and forest structural variables such as forest density and tree height [*Franklin et al.*, 2000; *Puhr and Donoghue*, 2000]. Remotely sensed data do not estimate the amount of biomass present in the forest directly, but rather, measure other characteristics such as crown size and forest density, which are correlated with biomass.

[6] In general, there has been very little success in mapping biomass over large areas from remote sensing. Most commonly, regression models have been used to relate biophysical parameters such as forest volume derived from field measurements with remotely sensed observations obtained from optical and active microwave instruments [*Gemmell*, 1995; *Myneni et al.*, 2001]. While radar image intensity shows good correlation with forest structure parameters in some situations, microwave reflectivity tends to saturate at low biomass levels (≈50–250 tons/ha, depending on the frequency [*Ranson et al.*, 1997]), which is a severe limitation in areas such as California where the above-ground biomass in forest stands ranges from about 50 to 400 tons/ha. In the future, LIDAR remote sensing will likely provide the best means of forest biomass mapping [*Lefsky et al.*, 1999]. However, LIDAR remote sensing is in its infancy and there is not a well-demonstrated way to scale from the inherently fine resolution of LIDAR measurements with limited geographic sampling to produce maps of large areas.

3. Data and Methods

3.1. Study Area and Data

[7] The study area encompasses eighteen National Forests in California. This area includes roughly 89,000 km^2,

[1] Department of Geography, Boston University, Boston, Massachusetts, USA.

[2] Remote Sensing Laboratory, Region 5, USDA Forest Service, Sacramento, California, USA.

0094-8276/04/2004GL019782$05.00

with elevations ranging from sea level to more than 4000 m. Total annual average precipitation varies from 16 cm on the southeastern side of the Sierra Nevada Mountains to 275 cm in northwestern coastal areas. The main life forms consist of conifer forest/woodland, hardwood forest/woodland, chaparral, soft chaparral, sagebrush scrub and herbaceous formations [*Franklin et al.*, 2000].

[8] The remote sensing data used in this study consist of 1-km surface reflectances derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). MODIS possesses seven spectral bands designed for land applications with wavelengths from 459 to 2155 nm. Because MODIS data are acquired with variable viewing geometry, we used Nadir Bidirectional Reflectance Distribution Function (BRDF) adjusted reflectances (NBAR) for this work [*Schaaf et al.*, 2002]. Each NBAR image provides surface reflectance data that have been normalized to a consistent nadir view geometry, and that are atmospherically corrected, cloud-cleared, and representative of 16 day periods. To improve separation between background vegetation, broadleaf forests, and needleleaf forests we used MODIS imagery acquired in both the summer (Aug. 16, 2001) and the winter (Jan. 01, 2001).

[9] Biomass data were provided by the Remote Sensing Laboratory for Region 5 of the United States Department of Agriculture Forest Service (USFS), who compiled a high resolution forest biomass map for National Forests in California. This map was generated by intersecting FIA-derived timber volume estimates with a forest cover map depicting 190 strata with uniform forest structure attributes including tree density, tree size, and regional forest type [*Woodcock et al.*, 1994; *Franklin et al.*, 2000]. The timber volume data were converted to biomass values by applying an expansion factor coefficient.

[10] For this work, we considered forested and shrubland areas. Unfortunately, biomass information was reported only for forested land cover types in the USFS data sets. To account for areas dominated by shrubs, we used a value of 48.8 tons/ha derived from *Riggan et al.* [1988] and *Gray* [1982]. The biomass for other land cover types (barren, grassland, water) was set equal to zero. To aggregate the Forest Service map to 1-km resolution, we overlaid the MODIS 1-km grid on the study area and computed the area weighted average biomass for each 1-km cell.

[11] To supplement the MODIS data, we also included climate and topographic variables in the analysis. The relationship between climate and terrestrial vegetation has long been established [*Holdridge*, 1947; *Box*, 1981]. More recently, climate and topographic variables have been used to predict vegetation composition at local to regional scales [*Franklin*, 1995; *Davis and Goetz*, 1990]. Here we use these variables in combination with remotely sensed information to improve the accuracy of predictions regarding above-ground forest biomass. The climate data set used for this work was developed by *Thornton et al.* [1997] and is representative of the 18 year period from 1980 to 1997. This data set provides mean monthly temperature and precipitation data at a spatial resolution of 1-km^2. The elevation data were extracted from the GTOPO30 global digital elevation data set (http://edcdaac.usgs.gov/gtopo30/gtopo30.html) at 1-km^2 resolution.

3.2. Analysis

[12] We first performed an exploratory analysis using generalized additive models (GAMs) [*Hastie and Tibshirani*, 1990] to investigate relationships between forest biomass and remotely sensed information, topography, and precipitation data. This approach is a more flexible extension of generalized linear models and has been previously applied with good success in ecological and vegetation modeling [*Franklin*, 1998; *Guisan et al.*, 2002; *Frescino et al.*, 2001]. Unfortunately, while GAMs represent a useful exploratory tool, they generally provide relatively poor predictive power on independent samples [*Hastie and Tibshirani*, 1990; *Frescino et al.*, 2001]. To provide predictions for biomass, we therefore require an alternative methodology.

[13] Tree-based models have been previously used in many contexts to predict both categorical and continuous variables. The basic theory behind this approach is reported by *Breiman et al.* [1984]. Tree-based models perform recursive partitioning of data sets, make no assumptions regarding the distribution of the input data, are able to capture non-linear relationships between the response and predictor variables, and provide easily understandable output. For the work reported here, we used a novel extension to tree-based models called *Random Forests* [*Breiman*, 2001]. This algorithm estimates large number of trees, in which different bootstrap samples of the data are used to estimate each tree. At each node, splitting is performed using a randomly selected sub-set of the predictor variables. The resulting model is more accurate and less sensitive to noise in input data relative to conventional tree-based modeling algorithms. For complete details, see [*Breiman*, 2001].

[14] To assess the ability of Random Forests to produce meaningful predictions, we performed a cross-validation analysis in which subsets of the data set were randomly held out and used as testing data. In each case, the test data set was extracted using a random sample. In the results discussed below, we used multiple training sets, composed of 516, 1032, and 2581 1-km^2 cells, which correspond to 1, 2, and 5 percent of the area covered by the Forest Service maps.

4. Results

[15] Exploratory analysis using GAMs revealed complex relationships between several key predictors and above-ground forest biomass [*Gemmell*, 1995] (Figure 1). For example, above-ground forest biomass increased with elevation from about 800 m up to 2500 m. Above this elevation, biomass decreased with elevation. GAMs also revealed a strong negative relationship between biomass and MODIS shortwave infrared (Band 6) reflectance, but only at low reflectance levels. For reflectance values larger than 0.2, no discernible correlation was detected. Similarly, the relationship between total annual precipitation and biomass was positive up to about 1500 cm, above which the relationship saturates.

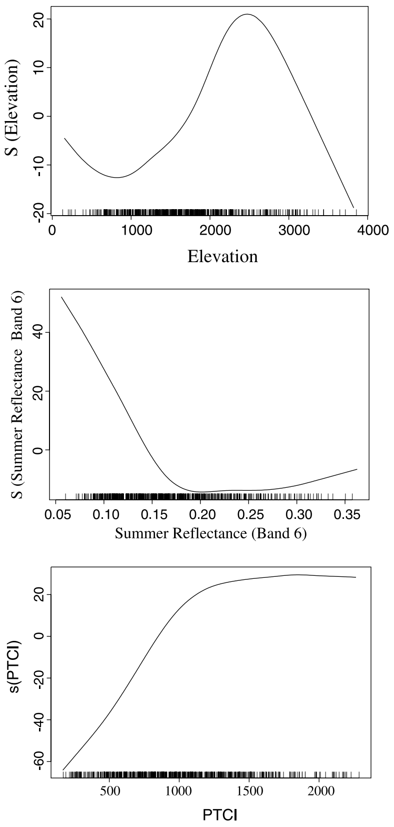

Figure 1. GAM results showing the relationship between the fitted function scaled to zero on the y axis and elevation, MODIS SWIR reflectance and total annual precipitation (PTCI). The bottom of each plot shows the frequency distribution of the observations.

[16] The Random Forest model estimated using MODIS, precipitation, and elevation data proved effective for predicting forest biomass (Figure 2). Depending on the size of the training data set, Random Forest estimated models with root mean square errors that ranged from 46.4 tons/ha to 41.2 tons/ha, with R^2's that varied from 0.68 to 0.75. At the same time, Random Forest tended to under-predict biomass above 250 tons/ha and over-predict biomass below 45 tons/ha. Despite this bias, 78% of the predicted values fell within ±50 tons/ha.

5. Discussion

[17] Data from MODIS Band 6 proved to be particularly important for this work. The physical basis behind this result is unclear. However, a number of previous studies have also identified the utility of SWIR data for estimating forest canopy structural variables [*Cohen and Spies*, 1992; *Puhr and Donoghue*, 2000; *Gemmell*, 1995]. These studies suggest that because the amount of diffuse radiation in the SWIR wavelength region is limited, shadows play an important role in controlling reflectance values in this band.

[18] In this context, the age structure of forests stands appears to be an important factor. The structure of young forests is usually characterized by a single canopy layer, high density, relatively few gaps, and trees of roughly the same size. Older forests, on the other hand, are characterized by a mix tree ages and sizes, and multiple canopy layers [*Cohen and Spies*, 1992]. This type of structure produces an increase in shadows, which decreases reflectance in the SWIR bands and helps to explain the negative relationship between above-ground biomass and data from MODIS Band 6.

[19] While the MODIS data provided a key source of information regarding above-ground biomass, climate and topographic variables were also important. Indeed, for areas such as California, which are characterized by a wide range of elevation and climate zones, these variables exert important control on the spatial distribution of above-ground biomass. For example, average annual precipitation was important for separating forests with large timber volume in northern coastal areas from lower volume forests in the Sierra Nevada. Similarly, precipitation proved to be useful for discriminating among Mediterranean vegetation types located in Southern California where oaks and shrubs dominate. The presence of broadleaf trees mixed with conifers created particular difficulties, and the model tended to underestimate biomass in areas characterized by broadleaf and conifer mixtures. In this context, the use of 1 km^2 spatial resolution was a key challenge for this work because virtually all grid cells included multiple forest stands and mixtures of forest and shrubs. Future efforts using somewhat finer (e.g., 500 m) resolution data should help to resolve this problem.

[20] While the cross-validated results from Random Forest were good, there remains room for improvement. Specifically, Random Forest tended to underestimate predictions for areas with high biomass and overestimate predictions for regions with low biomass. In the latter case, this effect is partly explained by the fact that a single value of biomass was used for the shrub class. Therefore, areas dominated by shrubs show little variance in biomass, which creates difficulties for empirical models such as Random Forest.

[21] At high levels of biomass (i.e., dense forests), uniformly low MODIS Band 6 reflectance explains part of the bias in model results. More generally, however, model

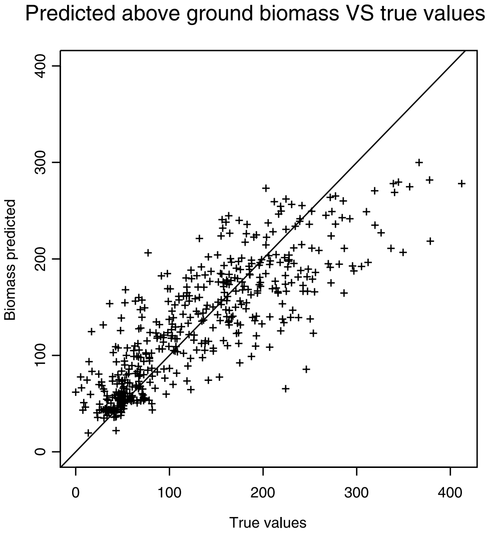

Figure 2. Predicted above-ground biomass vs Forest Service high resolution biomass data.

errors at low and high values of biomass values reflect a key weakness of tree-based models, which tend to penalize regions in the frequency distribution of the response variable (biomass) that are data sparse. In this case, the result is that values of biomass at the tails of the distribution (i.e., high and low biomass) tend to be poorly predicted. This problem can be mitigated by resampling the training data such that the distribution of biomass is more uniform across its entire range. However, this will tend to increase errors in biomass estimates that are closer to the median of the data. Thus, resolving this problem depends in part on the project objectives and priorities. Other methods of dealing with this problem have also been identified [*Cohen et al.*, 2003].

6. Conclusions

[22] Despite extensive research over the last decade on topics related to the global carbon cycle, knowledge regarding the stocks of carbon in standing biomass is limited. In this paper, we consider methods to estimate above-ground forest biomass over large areas using a relatively small training data set. Remotely sensed data, climate, and topographic variables all provided useful information in this regard. The methodology provides a simple, flexible, and powerful tool to combine and extract information from multivariate data in an environment characterized by complex non-linear relationships between forest biomass and remotely sensed, climate, and topographic variables.

[23] Using a sample of only 2 percent of the data, Random Forest was able to predict forest biomass for a wide range of vegetation formations with an RMSE of 44.4 tons/ha. However, limitations in the availability of data for shrub formations, in combination with the behavior of tree-based models, resulted in over-estimation (under-estimation) for low values (high values) of biomass. Despite these limitations, the results from this work suggest that there is good basis for pursuing biomass mapping at regional to continental scales using the current generation of remote sensing technology.

[24] **Acknowledgments.** This work was partially supported by NASA contract NAS5-31369.

References

Box, E. O. (1981), *Macroclimate and Plant Forms: An Introduction to Rredictive Modeling in Phytogeography*, 258 pp. and 25 maps, Dr. W. Junk, Norwell, Mass.

Breiman, L. (2001), Random forests, *Machine Learning*, *45*, 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*, Wadsworth, Belmont, Calif.

Brown, S. (2002), Measuring carbon in forests: Current status and future challenges, *Environ. Pollut.*, *116*, 363–372.

Brown, S. L., and P. Schroeder (1999), Spatial distribution of biomass in forests of the eastern USA, *For. Ecol. Manage.*, *123*, 81–90.

Cohen, W. B., and T. A. Spies (1992), Estimating structural attributes of Douglas-fir/western hemlock forest stands from Landsat and SPOT imagery, *Remote Sens. Environ.*, *41*, 1–17.

Cohen, W. B., T. Maierpserger, S. Gower, and D. Turner (2003), An improved strategy for regression of biophysical variables and Landsat ETM+ data, *Remote Sens. Environ.*, *84*, 561–571.

Davis, F. W., and S. Goetz (1990), Modeling vegetation pattern using digital terrain data, *Landscape Ecol.*, *4*, 69–80.

Franklin, J. (1995), Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients, *Prog. Phys. Geogr.*, *19*, 474–499.

Franklin, J. (1998), Predicting the distribution of shrub species in southern California from climate and terrain-derived variables, *J. Vegetation Sci.*, *9*, 733–748.

Franklin, J., C. E. Woodcock, and R. Warbington (2000), Digital vegetation maps of forest lands in California: Integrating satellite imagery, GIS modeling, and field data in support of resource management, *Photogramm. Eng. Remote Sens.*, *66*, 1209–1217.

Frescino, T. S., T. C. Edwards, and M. G. Gretchen (2001), Modeling spatially explicit forest structural attributes using generalized additive models, *J. Vegetation Sci.*, *12*, 15–26.

Gemmell, F. M. (1995), Effects of forest cover, terrain, and scale on timber volume estimation with thematic mapper data in a Rocky Mountain site, *Remote Sens. Environ.*, *51*, 291–305.

Gray, J. T. (1982), Community structure and productivity in Ceanothus chaparral and coastal sage scrub of southern California, *Ecol. Monogr.*, *52*, 415–435.

Guisan, A., T. C. Edwards, and T. Hastie (2002), Generalized linear and generalized additive models in studies of species distributions: Setting the scene, *Ecol. Modell.*, *157*, 89–100.

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

Holdridge, L. R. (1947), Determination of world plant formations from simple climatic data, *Science*, *105*, 367–368.

Houghton, R. (1992), Tropical forests and climate, in "Ecology, Conservation and Management of Southeast Asian Rainforests," edited by R. B. Primack and T. E. Lovejoy, pp. 263–290, Yale Univ. Press, New Haven, Conn.

Houghton, R. A., K. T. Lawrence, J. L. Hackler, and S. Brown (2001), The spatial distribution of forest biomass in the Brazilian Amazon: A comparison of estimates, *Global Change Biol.*, *7*, 731–746.

Lefsky, M. A., D. Harding, W. Cohen, G. Parker, and H. Shugart (1999), Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA, *Remote Sens. Environ.*, *67*, 83–98.

Myneni, R., J. Dong, C. Tucker, R. Kaufmann, P. Kauppi, J. Liski, L. Zhou, V. Alexeyev, and M. Hughes (2001), A large carbon sink in the woody biomass of northern forest, *Proc. Natl. Acad. Sci. U. S. A.*, *98*, 14,784–14,789.

Puhr, C. B., and D. N. M. Donoghue (2000), Remote sensing of upland conifer plantations using Landsat TM data: A case study from Galloway, south-west Scotland, *Int. J. Remote Sens.*, *21*, 633–646.

Ranson, K. J., G. Sun, R. Lang, N. Chauhan, R. Cacciola, and O. Kilic (1997), Mapping of boreal forest biomass from spaceborne synthetic aperture radar, *J. Geophys. Res.*, *102*, 29,599–29,610.

Riggan, P. J., S. Goode, P. M. Jacks, and R. N. Lockwood (1988), Interaction of fire and community development in Chaparral of southern California, *Ecol. Monogr.*, *58*, 155–176.

Schaaf, C. B., et al. (2002), First operational BRDF, albedo and nadir reflectance products from modis, *Remote Sens. Environ.*, *83*, 135–148.

Schroeder, P., S. Brown, J. Mo, R. Birdsey, and C. Cieszewski (1997), Biomass estimation for temperate broadleaf forest of the United States using inventory data, *Science*, *43*, 424–434.

Thornton, P. E., S. Running, and M. White (1997), Generating surfaces of daily meteorological variables over large regions of complex terrain, *J. Hydrol.*, *190*, 214–251.

Woodcock, C. E., et al. (1994), Mapping forest vegetation using Landsat TM imagery and a canopy reflectance model, *Remote Sens. Environ.*, *50*, 240–254.

A. Baccini, M. A. Friedl, and C. E. Woodcock, Department of Geography, Boston University, 675 Commonwealth Ave., Boston, MA 02215, USA. (abaccini@bu.edu)

R. Warbington, Remote Sensing Laboratory, Region 5, USDA Forest Service, Sacramento, CA 95814, USA.