Tree Height Estimations Using Remote Sensing and Data Mining Techniques

Mario Guevara[1]*, Brendan Malone[2], Alma Vázquez[1], Claudia Aguilar[1], Julián Equihua[3], Isabel Trejo[1], Michael Schmidt[1] and René R. Colditz[1].

1 National Commission for the Knowledge and Use of Biodiversity, Av Insurgentes Sur 4903 Parques del Pedregal, Tlalpan, 14010, DF, Mexico
2 Faculty of Agriculture and Environment, University of Sydney, NSW, Australia, CRICOS 00026A.
3 Instituto Tecnológico Autónomo de México, DF, México Rio Hondo No. 1, Tizapán Progreso, 01080

* *mguevara@conabio.gob.mx*

Abstract: We explored the use of remote sensing products, climate information sources, and field data observations to generate predictive tree heights maps at a national level by means of data mining. We analyzed a 19,367 field tree height database, leaving out 20% (3,873) for validation purposes. Predictors used here are nationally available in form of raster layers that were resampled to 1km (standing wood volume, canopy cover, water balance, and biomass). Prediction methods were multiple linear regressions, simple regression trees, conditional inference trees and Cubist. The prediction of the cubist method had the best correlation with validation data (r=0.68) and the lowest value of mean square error (0.9). All methods except conditional inference tree overestimates minimum values and all four methods underestimates maximum values. Stratification criteria, more predictors and combinations (such as digital elevation models and derived attributes), the use of finer resolutions, and the test of more methods will be a part in future works aiming to improve the accuracy of predictions. This framework can be useful to reporting requirements with regard to assessment of forest dynamics, deforestation rates, and other ecological studies.

key words: data mining, tree height, remote sensing.

1 Introduction

The most feasible way to collect and update periodically information about ecosystem processes is satellite remote sensing (Hall et al., 1999). Acquisition of remote sensing data is rapid, non-destructive, and cost effective (Viscarra et al. 2011). Remote sensing methods for quantitative mapping have proven useful inputs for a variety of modeling applications, in soils, (McBrantney 2003), topography and terrain analysis (Bertoldi et al. 2011, Mulder et al. 2011), water erosion (Vrieling 2006), hydrological processes (Su 2011, Mizlow et al. 2009), climate (Verstraete et al. 2008) and land use (Rozenstein and Karnieli 2011). There is also widespread interest for using remote sensing products for vegetation mapping and monitoring (Simard et al. 2011, Walker et al. 2007). De Sy (et al. 2012) presents an extensive review about remote sensing data availability with a focus on forest information products.

Vegetation parameters, like tree height, basal area, cover or density, are needed to produce estimates of forest carbon stocks, structural diversity, and biomass production. Spatial knowledge about these vegetation attributes is important for successful

implementations of climate change mitigation policies and decision making related to land use evaluation and planning (Saatchi et al. 2011). Vegetation structure is linked to soils, climate and topography (White et al. 2005). Landform and geological features are dominant factors producing high levels of spatial variability (Jenny, 1941, McKenzie and Ryan 1998, Boettinger 2010). This is the reason for field data collection in practical applications, requiring calibration and validation to avoid misleading interpretations of remote sensing products. At large areas, spatial variability makes field information difficult, expensive and time consuming. While remote sensing provides spatially-explicit data for large areas at considerable low costs, the quality of attributes that can be extracted has to be controlled; thus, field assessments continue to be an essential component of forest inventories (Kohl et al. 2006). Remote sensing is useful to detect and classify spatial characteristics of natural or transformed landscapes and its associated land covers (such as urban, water, soils or vegetation). However, environmental complexity makes remote sensing measures susceptible to under or overestimations.

At the national level, we combine field data of tree height with remote sensing products related to vegetation productivity (stand wood volume, canopy cover and biomass), and an empirical water balance layer to investigate significant correlations between them. The objective was to know if these auxiliary sources of information were good enough to predict tree height of Mexican forests for a continuous raster based on available field data observations (points), assuming that they are related in theory. Tree height data for modeling was chosen, because it is an input needed for above-ground biomass estimations and it is also a good indicator of forest structure. Prediction factors used are expected to be related with vegetation structure.

2 Materials and methods

Remote sensing products used in this exercise are nationally available in form of raster layers. These layers were resampled to 1km from its original resolution to facilitate computational requirements, and to avoid noise from local scale variation. The first two are radar based; standing wood volume in m^3/ha using data from ASAR Envisat at 1km spatial resolution (Santoro 2010) and biomass in tons/ha using data from Alos Palsar at approximately 93m spatial resolution (Kellndorfer, 2007). The third data set is canopy cover percentage using data from Landsat 4TM at 30m of spatial resolution (Hansen 2010). Additionally, precipitation and temperature data were extracted from the digital climate atlas of Mexico (UNIATMOS-UNAM, version 2.0, data of 1902 - 2011) at 926m spatial resolution. The empirical method of Thornwaite (1948) was applied to these data to produce a water balance layer assuming that vegetation can be partially explained by climate conditions. In theory, all of these products will have high correlations to vegetation structure parameters such as tree height.

The field data set used here corresponds to 19,367 samples sites, each one representing an average of tree height measurements in four circular plots of 900m^2 (one central plot and three replicates separated by equivalent distances of 45.5m) . This extensive sample dataset was obtained from the National Forest Commission (INFyS 2004 – 2009). For validation purposes, we leave out 20% of the samples (n=3,873). The sampling design follows a systematic grid, for forestry zones, samples are separated by 5km, for arid, semiarid, and other environments the distance between points is 10km.

Remote sensing products and climate layers were used as statistical predictors in response to the strength of correlation that they show with tree height. The objective of fitting data to statistical models is to find an adequate model in which is possible to map , vegetation properties across all areas where ground truth information does not exist but correlated environmental information is available. We implemented multiple data mining methods with R (Development Core Team 2011). Data mining refers to the process of (automatically) extracting models from large stores of data. Multiple linear regression (MLR) attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Assumptions like linearity, normality of residuals, and constant variance are important to consider for acceptable performance of MLR (GEOS 2011). Regression trees represent a recursive partition of data into increasingly homogeneous subsets. These models are non-parametric and are composed of clasification rules. We compared performance of two types of regression trees, simple trees (ST) and conditional inference trees (CT). In simple trees the number of nodes specifies the number of possible values that are predicted. This creates maps that are categorical in nature. Conditional inference trees (CT) produces more detailed products, in which embed tree-structured regression models into a well-defined theory of conditional inference procedures. This model estimate a regression relationship by binary recursive partitioning in a conditional inference framework (Hothorn et al. 2006) Regression trees models used here can accommodate both continuous and categorical predictor variables, have no statistical assumptions, and will perform an ad-hoc variable selection like determine what are the most important predictor variables (Malone 2012).Cubist (CU) is a rule-based method similar to regression trees but including a multivariate linear model for each terminal node. With the rule–based approach of Cubist each rule specifies the conditions under which an associated multivariate linear sub-model should be used.

While the outputs of the mentioned methods provide some useful diagnostics of the model fits, we require further validation of our model fitting which is done by comparing the observed values with the corresponding values that were predicted. We compared the performance of spatial prediction methods using the mean square error and the Pearson correlation coefficient (r) between validation samples and modeled values. It is important to consider that model quality and uncertainty of spatial inferences depends on two main factors: (1) the quality of data sets and its positional accuracy and (2) the strength of correlations between target variables and its prediction factors. Geostatistical methods were not used here, because no spatial autocorrelation was found with available data.

3 Results and discussion

Figure 1 shows how different are statistical distributions of data sets. The skewness indicates for each histogram the distribution of the majority of the values relative to the mean.

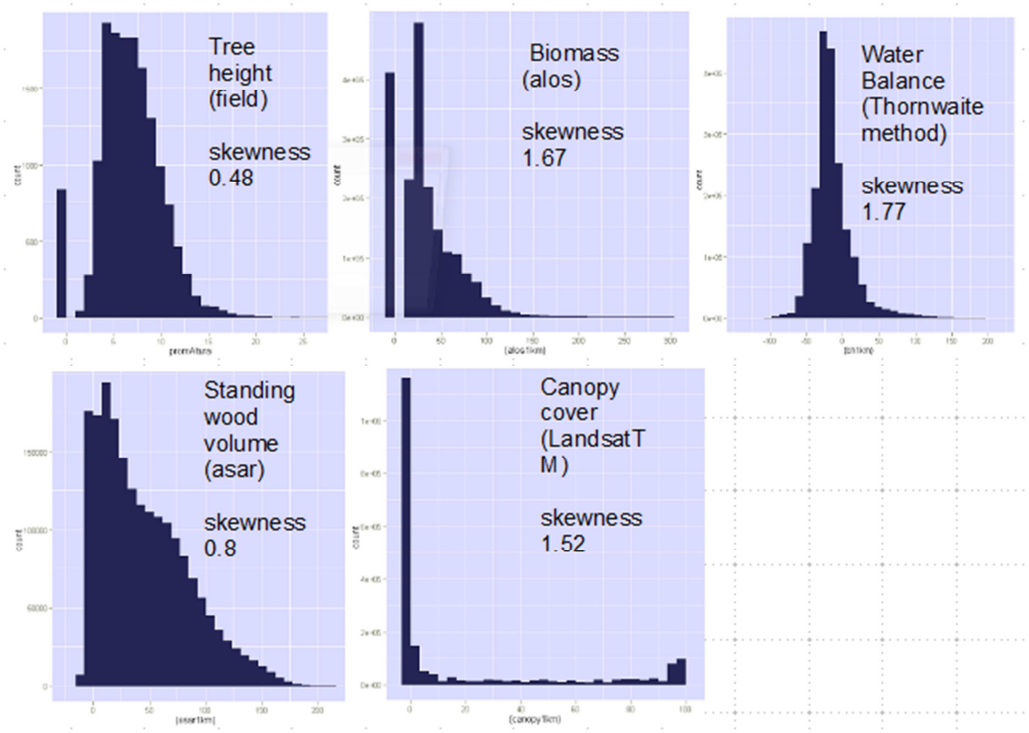

Figure 1 Frequency histograms and skewness value of all five data sets

Table 1 shows Pearson correlation coefficient between tree height and predictors. Correlation values between tree height and predictors are lower than expected. Correlations with other vegetation parameters in the data base, such as base area, crown cover, diameter at chest height or density were lower.

Table 1 Correlation between tree height and predictors significant at alpha 0.01. Degrees of freedom for all cases are 15,482.

Predictor	***r***
standing wood volume m3/ha	0.48
biomass in tons/ha	0.36
canopy cover percentage	0.47
Thornwaite water balance	0.25

Figure 2 shows spatial predictions of all four methods. The most important predictor for MLR, CT and CU methods was canopy cover percentage but for ST method it was standing wood volume, the additional contribution of a water balance layer was not important to any model.

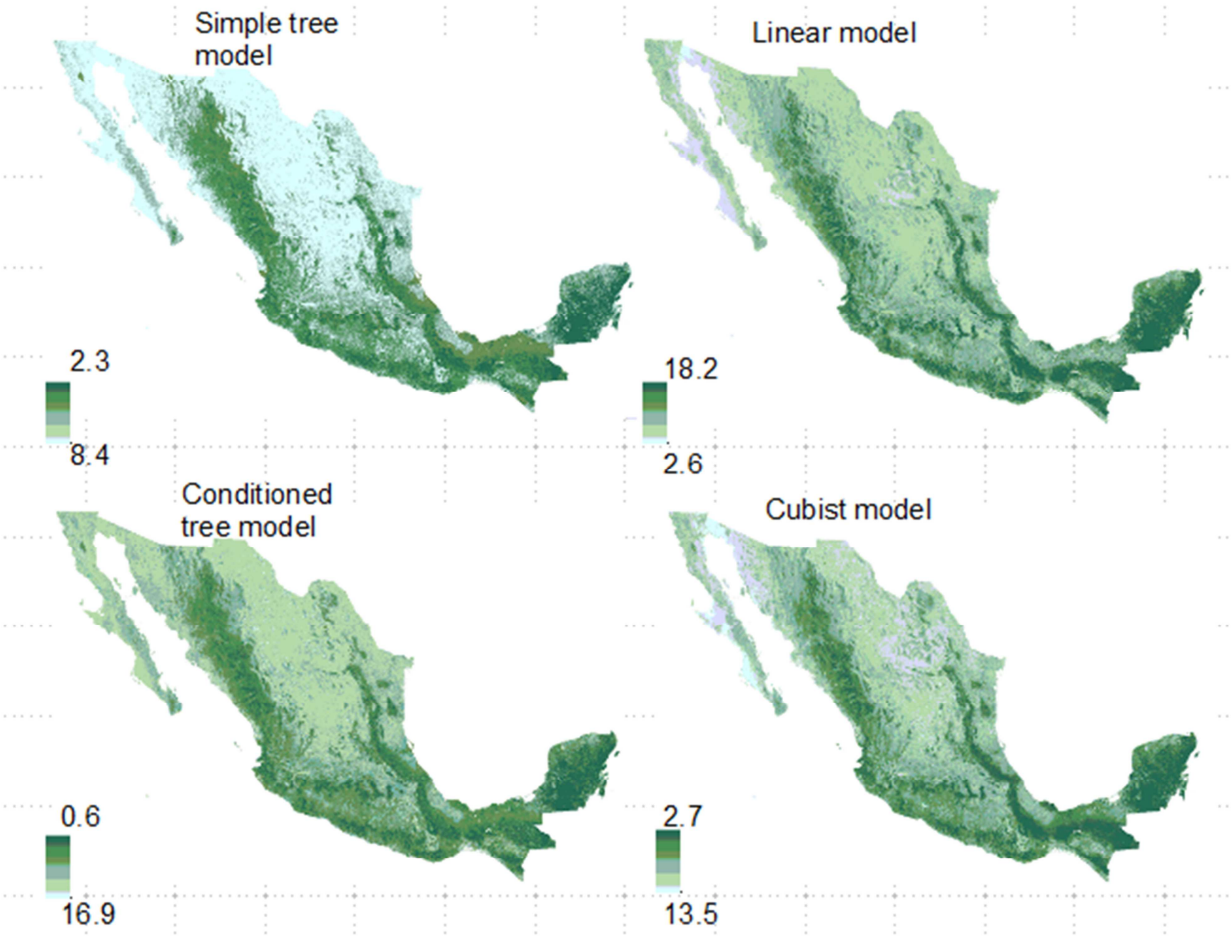

Figure 2 Spatial predictions of tree heigth across Mexico (m).

Table 2 shows descriptive statistics of modeled and field data. Model validation values are higher than correlation between field data and predictor's data. All methods except conditional inference trees overestimate minimum values and all four methods underestimate maximum values.

Table 2 Descriptive statistics of modeled data in comparison with field tree height data

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.
MLR	2.6	3.6	4.2	4.7	5.5	18.2	0.1
ST	2.3	2.3	3.9	4.4	6.3	8.4	0.2
CT	0.6	2.6	4.1	4.2	6.1	16.9	0.3
CU	2.7	3.6	4.5	4.9	6.2	13.5	0.2
Field data	0.6	4.6	6.6	6.8	8.7	28.4	3.2

Table 3 shows mean square errors for predictions and *r* values for the correlation between validation sample and modeled data. The mean square error is the sample estimate of the variance of the regression residuals and is useful to identify the contribution of each predictor to the model. We used original values of data for this analysis, but we found that,

as reported earlier by Gallardo-Cruz (et al. 2012) transformation to natural logarithm of data sets reduces mean square error values.

Table 3 Error metrics and validation value for all predictions made significant at alpha 0.01.

Model	Mean square error	Validation *R*
MLR	0.1	0.64
ST	0.36	0.59
CT	0.64	0.65
CU	0.09	0.68

Multiple linear regression model was weak because model assumptions were not respected. The relation between tree height and predictors may not be linear and some dependence between them is expected, however mean square error was also lower than regression trees methods and sample validation and modeled data correlation value was higher compared with simple trees and similar to conditional inference prediction. An advantage of conditional inference trees compared to simple trees is that it uses a covariate selection scheme that is based on statistical theory, e.g. selection by permutation-based significance tests, and thereby avoids a potential bias in regression trees (Hothorn et al. 2006). Spatial dependence and statistical distribution of data are other two important considerations for choosing the most adequate model and to produce unbiased, accurate and consistent results.

The prediction of the cubist method had the best correlation with validation data and the lowest value of mean square error. Previous studies have shown stronger *r* validation values at regional and continental scales, for example r=0.7 in relation to tree height mapping using the random forest algorithm (Simmard et al. 2011) and r=0.8 for forest carbon stocks using a maximum entropy approach (Saachi et al, 2011). For regional studies using finer spatial resolution data , for example, Gallardo-Cruz et al. (2012) found *r* validation values of 0.93 for linear models in tropical dry forest with 2.6m Quickbird imagery and field tree height data. The *r* values of this study are lower than expected but statistically acceptable, considering that we map tree height throughout different vegetation types across the entire country (~ 1,960,000 km^2). Currently, we are addressing this issue using: a) stratification criteria, b) additional predictors with combinations and its original resolutions (such as digital elevation models and derived attributes or band values from new remote sensing technologies like lidar or hyperspectral data) and c) the test of more spatial prediction methods like machine learning for which good results were found by Baccini (et al. 2012) estimating the total net emission of carbon from worldwide tropical deforestation rates.

4 Conclusions

Vegetation attributes related to forest structure represented by remote sensing methods need field data for its adequate validation and interpretation. Correlation values between remote sensing layers and field data were lower than expected. The best prediction of tree height data was produced using the Cubist method. Sample validation and modeled data correlations of all four methods are higher than correlations between field data and

predictors. For practical applications in relation to forest structure mapping, remote sensing and climate products showed in this exercise mislead tree height representation. Decision making related to land planning based on remote sensing products used here as predictors is not recommended.

This framework the results of this study are useful for further national assessments of forest dynamics and monitoring of deforestation rates (e.g. CONAFOR´s INFyS data base) and the relation to remote sensing capacities.

Step by step, the use other possible significant predictors of tree height, like digital elevation models and derivatives, vegetation indexes and spectral reflectance values of several passive or active sensors, as well as finer resolutions or more different prediction methods will be subject of future work.

5 References

Baccini A. Goetz S., Walker W., Laporte N. Sun M., Sulla-Menashe D., Hackler J.,Beck P.,Dubayah R., Friedl M. , Samanta S. and Houghton R. 2012 Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps PNAS 2 DOI: 10.1038/NCLIMATE1354

Bertoldi L., Massironi M., Visonà D.,Carosi R., Montomoli Ch., Gubert f., Naletto, Pelizzo M. 2011, Mapping the Buraburi granite in the Himalaya of Western Nepal: remore sensing analysis in a collisional belt with vegetation cover and extreme variation of topography Remote sensing of Environment, 115, 1129-1144

Boettinger J.L. 2010 Cap. 2 Environmental Covariates for Digital Soil Mapping in the Western USA en J.L. Boettinger et al. (eds.), Digital Soil Mapping, Progress in Soil Science 2, Springer

De Sy V., Herold M., Achard F., Asner G., Held A., Kellndorfer J., Verbesselt J. 2012
Synergies of multiple remote sensing data sources for REDD+ monitoring Current Opinion in Environmental Sustainability, In Press, Corrected Proof.

Gallardo-Cruz et al. 2012 Predicting Tropical Dry Forest Successional Attributes from Space: Is the Key Hidden in Image Texture? PLoS ONE 7(2): e30506

Hall F., Townshed J., Engman E., 1995. Status of Remote Sensing Algorithms for Estimation of Land Surface State Parameters REMOTE SENS. ENVIRON. 51:138-156

Hothorn T., Hornik K., and Zeileis A. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework.Journal of Computational and Graphical Statistics, **15**(3), 651–674

Jenny, H. 1941 Factors of soil formation: a system of quantitative pedology McGraw Hill Book Company Inc

Köhl M, Magnussen S., Marchetti M 2006. Sampling Methods, Remote Sensing and GIS Multiresource Forest Inventory Springer-Verlag Berlin Heidelberg 373 p.

Malone B. 2012, 5th Digital soil mapping workshop and intensive training. Faculty of Agriculture and Environment, University of Sydney.

McKenzie N. y P. Ryan. 1999 Spatial prediction of soil properties using environmental correlation Geoderma 89 67–94

Mulder V., De Bruin S., Schaepman M., Mayr R. 2011 The use of remote sensing in soil and terrain mapping — A review Journal of Environmental Management, 90, 225-2260

McBratney, A.B., Santos, M.L.M., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Milzow C., Kgotlhang L., Kinzelbach W., Meier P., Bauer-Gottwein P. 2009, The role of remote sensing in hydrological modelling of the Okavango Delta, Botswana

Rozenstein O., Karnieli A. 2011 Comparison of methods for land-use classification incorporating remote sensing and GIS inputs Applied Geography, 31, 533-544

Notes_11, GEOS 585A, Multiple Linear regression Spring 2011, available on http://www.ltrr.arizona.edu/~dmeko/notes_11.pdf

Saatchi S., Harrisc N., Brownc S., Lefskyd M., Mitcharde E., Salasf W., Zuttaa B.,Buermannb W., Lewisg S., Hagenf S., Petrovac S., Whiteh L., Silmani M.,and Morelj A 2011 PNAS 108 9899-9904

Simard, M., N. Pinto, J. B. Fisher, and A. Baccini 2011, Mapping forest canopy height globally with spaceborne lidar, J. Geophys. Res., 116, G04021, doi:10.1029/2011JG001708.

Su Z., Roebeling R., Schulz J.,Holleman I., Levizzani V., Timmermans W. Rott H., Mognard-Campbell N.,Jeu R., Wagner W., Rodell M.,Salama M., Parodi G., Wang L. 2011. Observation of Hydrological Processes Using Remote Sensing Treatise on Water Science, 2, 351-399 Journal of Environmental Management, 90, 2252-2260

Thornwaite, 1948, An Approach toward a Rational Classification of Climate, Geog. Rev. 38 55-94 AGS

Viscarra R. and Behrens T. 2010 Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (2010) 46–54

Vrieling A. 2006 Satellite remote sensing for water erosion assessment: A review CATENA, 65, 2-18

Verstraete M., Brink A., Scholes R., Beniston M., Stafford M. 2008 Climate change and desertification: Where do we stand, where should we go? Global and Planetary Change, 64, Pages 105-110

Walker W., Kellndorfer J., LaPoint E., Hoppus M., Westfall J. 2007 An empirical InSAR-optical fusion approach to mapping vegetation canopy height Remote Sensing of Environment, 109, 482-499

White A., Kumar P. and Tcheng D. 2005 A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States Remote Sensing of Environment 98 1 – 20

Data sets:

Remote sensing:

1- Standing wood volume m3/ha, by M.Santoro, GAMMA Remote Sensing AG, 2010, sensor ASAR Envisat, 1km.

2- Biomass tons/ha, by J. Kellndorfer, Woods Hole Research Centre, 2007, sensor Alos Palsar, ~93m.

3-Canopy cover percentage, by M. Hansen, SDSU-USGS, 2005-2010, Landsat 4TM 30m.

Climate:

1- Raster layers of precipitation and temperature from the Informatics Unit for Atmosferic and Environmental Sciences UNIATMOS-UNAM, 1902-2011, 926m.

Software used

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.