Ecosphere

Predicting aboveground forest biomass with topographic variables in human-impacted tropical dry forest landscapes

M. Salinas-Melgoza, M. Skutsch and J.C. Lovett

Appendix S2

Topographic variables

Here the calculation of independent variables (Table S1) (topographic variables and distances to roads and distance to settlements) is described. All the topographic variables were derived from a Digital Elevation Model (DEM), which is a digital representation in raster format of the relief of a surface between points of known elevation, which was obtained from the national statistical institute of Mexico (INEGI 2017). The spatial resolution of this spatial data base is of 15 × 15 meters. This DEM was the main input used to obtain the other topographic variables. Previous to calculation of the topographic variables used as dependent variables, preprocessing procedure was performed in order to remove from DEM local depressions (sinks) that would affect posterior calculations using the depression filling algorithm of Planchon and Darboux (2001).

The topographic variables were derived from the DEM, using RSAGA (Brenning 2008) that provides access to System for Automated Geoscientific Analyses, SAGA v 2.1.4 (Conrad et al. 2015) software within R v 3.4.0 environment (R Core Team 2017). We also used QGIS Desktop 2.18.7 with GRASS 7.2.0 (QGIS Development Team 2016) for data manipulation. R packages used for calculations, data processing and analysis were: raster (Hijmans et al. 2016), spatialEco (Evans 2017), rgdal (Bivand et al. 2017).

Table S1. Topographic variables and distance to roads calculation.

Variable (units)	**Calculation**
Elevation above sea level (m)	Obtained directly form DEM (INEGI 2017)
Aspect (N,S,E,W)	Obtained using R, RSAGA, and SAGA. Method used was poly2zevenbergen. Input: DEM
Profile curvature (Degrees/m)	Obtained using R, RSAGA, and SAGA. Method used was poly2zevenbergen. Input: DEM
Planar curvature (Degrees/m)	Obtained using R, RSAGA, and SAGA. Method used was poly2zevenbergen. Input: DEM
Tangential curvature	Obtained using R, RSAGA, and SAGA. Method used was poly2zevenbergen. Input: DEM
Total insolation (kWh/m^2)	Annual averaged obtained using R, RSAGA, and SAGA: integrated module “Incoming Solar Radiation” (Conrad et al. 2015) using the Lumped Atmospheric transmittance atmospheric effect. Input: DEM
Diffuse insolation (kWh/m^2)	Annual averaged obtained using R, RSAGA, and SAGA: integrated module “Incoming Solar Radiation” (Conrad et al. 2015) using the Lumped Atmospheric transmittance atmospheric effect. Input: DEM
Direct insolation (kWh/m^2)	Annual averaged obtained using R, RSAGA, and SAGA: integrated module “Incoming Solar Radiation” (Conrad et al. 2015) using the Lumped Atmospheric transmittance atmospheric effect. Input: DEM
Slope (%)	Obtained using R, RSAGA, and SAGA. Method used was poly2zevenbergen. Input: DEM

Topographic position indices using different scales (number of pixels in the immediately surrounding area included in calculation: 1) 5 pixels 2) 11 pixels 3) 15 pixels 4) 19 pixels 5) 25 pixels 6) 35 pixels 7) 45 pixels 8) 61 pixels	Implement the calculation of tpi from De Reu et al. (2013) using spatialEco (Evans 2017)
Topographic wetness index	This variables was obtained using R, RSAGA, and SAGA. Type of area used was square root of catchment area and catchment slope as type of slope. Input: DEM
Distance to road (m)	This variable was calculated with QGIS Desktop NNJoin plugin ver. 1.2.2 (Tveite 2015). Spatial data bases of roads were obtained from IIEGJ (2011).
Distance to human settlement (m)	This variable was calculated with QGIS Desktop NNJoin plugin ver. 1.2.2 (Tveite 2015). Location of human settlements was obtained from INEGI (2010).

Selection of topographic variables

Prior to the modeling analysis, variance inflation factors (VIF) were used to detect multi-collinearity among the topographic predictors, sequentially dropping the variables with the highest VIF scores, recalculating the VIFs and repeating this process until all VIFs were lower than 7 (Montgomery et al. 2012). In the literature, acceptable VIF thresholds vary from 10 to 3 (Zuur et al. 2010, Montgomery et al. 2012). This means that that the independent variables selected have a coefficient which is inflated by a maximum of seven as a result of linear dependence with other independent variables. Variables were selected out of the 18 for three different sets: firstly for all communities together, and then for the two subgroups (A and B) of communities, as defined by CART analysis, explained in methods section). At the overall level, eleven variables were selected: slope, elevation, cprof, cplan, ctan, dirinsol, difinsol, tpi25, tpi61, TWI and aspect. Ten variables were selected for group A of communities; slope, elevation, cprof, cplan, ctan, dirinsol, tpi19, tpi6, TWI, and aspect. Nine variables were selected for group B of communities; slope, elevation, cprof, cplan, dirinsol, difinsol, tpi25, tpi61 and aspect, which were selected in base of VIF procedure.

CART, MARS and PGLM procedure

The classification and regression tree analysis (CART) performed is extremely resistant to outliers and can be applied to data sets having a large number of independent variables (Steinberg and Colla 1995) to evaluate whether the dependent variables (AGB) can be explained by complex interactions between the independent variables. This procedure makes no distributional assumptions and does not seek cause-and-effect relationships between variables, but rather looks for statistical associations (Steinberg and Colla 1995).

CART performs well in capturing interaction effects among the independent variables and provides the relative importance of each explanatory variable within one fitted tree structure (Clark and Pregibon 1992). CART uses binary recursive partitioning on the dataset along the entire range of variation of the explanatory variables. This is in order to split the data set into homogeneous subsets based on their relationship to sets of several predictor variables. The primary split aims to maximize the reduction in deviance. Every split made finds the predictor variable that results in the greatest change in explained deviance by splitting the dataset (Clark and Pregibon 1992). When changes in explained deviance are not found, the tree is considered fully grown with terminal nodes, which have a conditional mean. This is the predicted value for all sites included in that terminal node. Lower branches were removed from this fully growth tree to avoid over-fitting the data and to ensure that the remaining branches are robust enough through a pruning process using the lowest estimated error rule (Clark and Pregibon 1992). Model goodness of fit is estimated by R^2, which is a normalized form of the residual sum of squares (RSS) that indicates the percentage of variance explained by each split. CART models were build using a cost-complexity parameter (CP), relative error in the predictions (rel error), mean value of the errors of the cross-validations (xerror) and cross validation standard deviation (xstd) and number of splits (nsplit) obtained from ten cross-validation procedures repeated three times. Each fitted tree provides variables and thresholds for those variables. Also surrogate splits variables and thresholds are provided to identify variables that may be masked for the primary split. CART analysis was conducted using the rpart package v. 4.1-10 (Therneau et al. 2015).

PGLM evaluates breakpoints iteratively along the extent of variation of the independent variables. Piecewise-Generalized Linear Models were fitted using the Segmented Package (Muggeo 2008). The Segmented Package iteratively fitted two or more segments across a range of breakpoints, using GLM models and then returned the model with the lowest residual sum of squares (Muggeo 2003). Changes in the slope of the GLM were assessed using the Davies' test, which chooses a number of fixed breakpoints along the X-axis and looks for statistically significant differences in regression slopes on each side of the breakpoints. Breakpoints estimates were included in the model only when the 95% confidence intervals do not overlap. Because piecewise-regression cannot deal with interaction between segmented variables, no piecewise-model including interactions between these explanatory variables was attempted.

Multivariate Adaptive Regression Splines (MARS, Friedman 1991) is a nonparametric regression procedure with no functional relationship assumption and is used in situations where there are many dimensions to be considered. In this study this analysis was used to evaluate whether forest biomass follows a continuous non-linear function with predictive variables, by fitting multi-variate splines. This analysis allows for the use of multiple variables that may not have common effects across the sample (Osei-Bryson 2014). It partitions or segments the multi-dimensional space into regions using CP, giving each region its own regression equation. MARS is a two steps recursive procedure that progressively increases the model complexity. A backward procedure removes the least significant function form of the model. Each time the procedure is performed, all possible knots for each variable are evaluated and the ones that minimize prediction error are selected. Eventually, in order to deal with overfitting, the CP of the model is reduced by a “pruning procedure”, which removes the functions that contribute least to the overall goodness of fit. Model fit is evaluated at each iteration through a “generalized cross-validation” (GCV) measure of mean square error. A Generalized Linear Model (GLM) argument was specified in the MARS model procedure, with a gamma error and log link, which was repeated independently. This procedure construct the MARS model with several piecewise linear basis functions, which use a threshold value called a knot (Friedman 1991, Faraway 2016). For each iteration the process builds a MARS model with the in-fold data, then measures the relationship (R^2) between predictions from MARS model and the ones made on the out-of-fold data (one tenth of the complete dataset). Then the mean R^2 (CVRsq) of these out-of-iteration R^2s is reported. We obtained the nonlinear model from the MARS procedure describing the basis function (BF), which are those functions that contribute significantly to the fit of the model (Friedman 1991).

LITERATURE CITED

Bivand, R., T. Keitt, and B. Rowlingson. 2017. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.2–7. https://CRAN.R-project.org/package=rgdal

Brenning, A. 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. Pages 23–32 *in* Boehner, J., Blaschke, T. and Ontanarella, L. (Eds.). SAGA - Seconds Out (= Hamburger Beitraege zur Physischen Geographie und Landschaftsoekologie 19)

Clark, L.A., and D. Pregibon. 1992. Tree-Based Models. In Chambers, J.M. and Hastie, T.J. (Eds.). Statistical Models in S, Boca Raton, FL: Chapman & Hall/CRC. 377–420 Pp.

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann and J. Böhner. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development Discussions 8: 2271–2312.

De Reu, J., J. Bourgeois, M. Bats, A. Zwertvaegher, V. Gelorini, P. De Smedt, W. Chu, M. Antrop, P. De Maeyer, P. Finke, M.V. Meirvenne, J. Verniers and P. Crombe. 2013. Application of the topographic position index to heterogeneous landscapes. Geomorphology 186: 39–49.

Evans, J.S. 2017 spatialEco. R package version 0.0.1-7. https://CRAN.R-project.org/package=spatialEco>

Friedman, J.H. 1991. Multivariate Adaptive Regression Splines. Annals of Statistics 19: 1–141.

Hijmans, R.J. 2016. raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster

Instituto de Información Estadística y Geográfica de Jalisco (IIEGJ). 2011. Atlas de Caminos y Carreteras del Estado de Jalisco. http://iit.app.jalisco.gob.mx/sitios/caruca/index.html Accessed: Jun 30 2016.

Instituto Nacional de Estadística, Geografía e Informática (INEGI). 2010. Censo Nacional de Vivienda del Instituto Nacional de Estadística y Geografía (National Housing Census National Institute of Statistics and Geography). http://www.inegi.org.mx/est/contenidos/proyectos/graficas_temas/piramides/graf/2010.htm

Instituto Nacional de Estadística, Geografía e Informática (INEGI). 2017. Continuo de elevaciones Mexicano (Mexican Continuous Elevations) versión 3.0 (CEM 3.0) http://www.inegi.org.mx/geo/contenidos/datosrelieve/continental/continuoelevaciones.aspx Accessed: May 30 2017.

Montgomery, D.C., E.A. Peck, and G.G. Vining. 2012. Introduction to Linear Regression Analysis. Fifth edition. Wiley, New York.

Muggeo, V.M.R. 2008. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News 8(1): 20–25. URL http://cran.r-project.org/doc/Rnews/.

Muggeo, V.M.R. 2003. Estimating regression models with unknown break-points. Statistics in Medicine 22: 3055–3071.

Osei-Bryson, K.-M. 2014. Overview on Multivariate Adaptive Regression Splines. Pages 93–107 *in* Osei-Bryson, K-M and O. Ngwenyama. Advances in Research Methods for Information Systems Research: Data Mining, Data Envelopment Analysis, Value Focused Thinking. Springer Science+Business Media, New York.

Planchon, O. and F. Darboux. 2002. A fast, simple and versatile algorithm to fill the depressions of digital elevation models. Catena 46: 159–176.

QGIS Development Team. 2016. QGIS Geographic Information System Open Source Geospatial Foundation. URL http://qgis.osgeo.org

R Core Team. 2017. R: A Language and Environment for Statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.r-project.org

Steinberg, D., and P. Colla. 1995. CART: Tree-Structured Non-Parametric Data Analysis. San Diego, CA. Salford Systems. 336 Pp.

Therneau, T., B. Atkinson, and B. Ripley. 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–10. https://CRAN.R-project.org/package=rpart

Tveite, H. 2015. NNJoin Plugin: Nearest neighbour join. Join vector layers based on nearest neighbor relationships. https://plugins.qgis.org/plugins/NNJoin/version/1.2.2/

Zuur, A.F., E.N. Ieno., and C.S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1: 3–14.