Potential and limitations of multi-temporal earth observation data to improve model results of tree species distribution in Mexico

A. Cord [a,b]*, M. Schmidt [b], S. Dech [a,b]

[a] Department of Remote Sensing, Institute of Geography, University of Würzburg, Mineralogiegebäude, 97074 Würzburg, Germany
[b] German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Germany

Abstract –The study explored the potential of multi-temporal remote sensing data for distribution modeling of selected tree species belonging to the genera *Pinus spp.* (pine) and *Quercus spp.* (white oak) in Mexico. Several environmental predictor data sets at 1 km^2 spatial resolution were used in combination with the Maxent algorithm (Phillips et al., 2004), namely (1) phenological metrics derived from the Terra-MODIS 16-day vegetation indices product MOD13A2 averaged over the seven years of the study period from 2001 to 2007, (2) topographic data (elevation, slope, and aspect) of the SRTM mission, and (3) a series of bioclimatic variables (WorldClim, Hijmans et al., 2005) derived from monthly temperature and rainfall values. Different model scenarios were compared and showed that remote sensing data contributed significantly to discover habitat characteristics even within similar climatic conditions. Moreover, a sharper delineation of the predicted areas and better exclusion of regions that had suffered land cover change was possible. The improved distribution maps can contribute to long-term and sustainable conservation planning and management of biodiversity hotspots.

Keywords: MODIS, vegetation phenology, time series, species distribution modeling, maximum entropy, Mexico.

1. INTRODUCTION

Spatially-explicit knowledge of species distributions is of high relevance for decision-makers in the fields of conservation biology and land management planning. Species distribution models (SDMs) have become a key element in documenting biodiversity (Ferrier et al., 2004; Saatchi, 2008) and for applied and theoretical ecological research (Guisan and Thuiller, 2000; Austin, 2002). The conceptual background for SDMs is based on the ecological niche model in which a species can be quantitatively represented by a multidimensional combination ("hypervolume", Hutchinson, 1957) of abiotic and biotic variables required for a viable population to persist. So far, the majority of SDM studies applied only climate and topography data as environmental predictors of a species' distribution. Even though these parameters broadly determine the species' ecological niche, models may produce inaccurate predictions when important local or regional factors are missing. Remote sensing data open up the possibility to enlarge this spectrum of causal or driving forces for species' distribution and abundance beyond topographic and climatic conditions. They provide measurements and surrogates directly related to vegetation type and structure, biomass, and other ecosystem variables that collectively improve our understanding of habitat characteristics.

In general, remote sensing data can contribute to biodiversity research and specifically SDM by (1) improving both spatial and temporal resolution and (2) adding new information sources and dimensions of environmental variables to the input data (Saatchi et al., 2008). Recent studies (e.g. Thuiller et al., 2004; Pearson et al., 2004) started to incorporate categorical land cover data derived from remote sensing imagery. However, these usually yield a fairly small number of nominal variables meaning the thematic land cover classes and are therefore often not detailed enough to improve predictions of species' distributions (Bradley and Fleishman, 2008).

Novel analytical techniques have recently been developed that more fully exploit the temporal information of remotely sensed imagery (beyond spectral signatures) in order to quantify a broader range of ecosystem characteristics. This multi-temporal data allows for the extraction of phenological, seasonal, and latitudinal variations in vegetation cover over space and time and can thus contribute to improve SDMs. The characteristics of the phenological cycles appear to be directly related to both vegetation type and species diversity and thus indirectly to small-scale heterogeneity of climatic and topographic conditions in the corresponding study region. However, only recently published studies directly apply remote sensing data as model input parameters (primarily products provided by the MODIS science team such as the *Vegetation Continuous Fields* (VCF), *Vegetation Indices* (VI) or *Leaf Area Index* (LAI) products). Among these, only a fistful of very recent publications (Prates-Clark et al., 2008; Reed et al., 2008; Saatchi et al., 2008; Viña et al., 2008) take advantage of the high temporal resolution of the data and use selected characteristics of the time series (related to vegetation phenology) as model input.

The question to be addressed is how continuously available multi-temporal remote sensing data can provide information on vegetation and landscape characteristics that affect and mirror the spatial distributions of species. In this context, this study aimed to investigate the potential of multi-temporal MODIS data for modeling the spatial distribution of selected tree species belonging to the genera *Pinus spp.* (pine) and *Quercus spp.* (white oak) in Mexico.

2. DATA AND METHODS

2.1 Species occurrence data

Species presence data was provided by the *National Commission for the Knowledge and Use of Biodiversity* (CONABIO) of Mexico. The country is a major hot spot of diversity for both genera *Pinus* and *Quercus* which are of high importance for forestry and conservation biology. Available occurrence data including field observations and herbarium samples were scanned for double entries and randomly split into model training (80 %)

* Corresponding author

and model validation (20 %) samples for each species. As shown in Figure 1, study species were selected to adequately cover a variety of sample sizes and different geographic species' ranges over the study region.

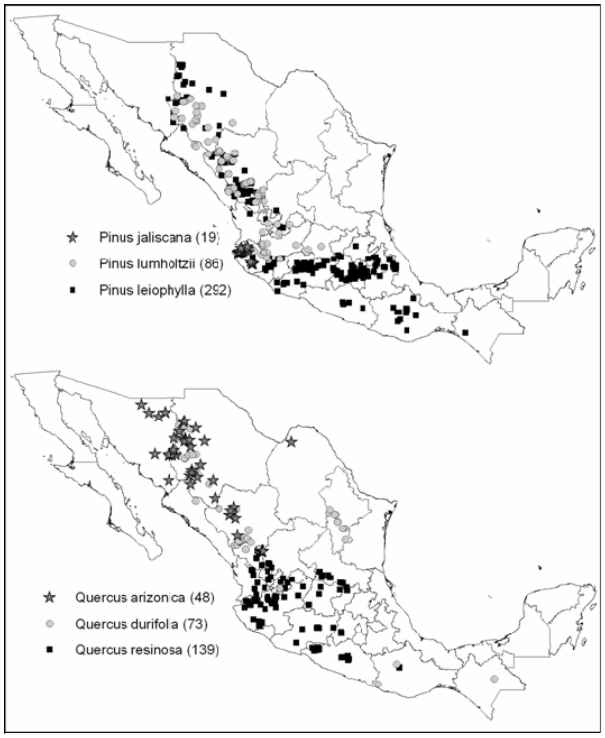

Figure 1. Distribution patterns and geographical ranges of study species. Numbers of samples per species are indicated in brackets.

2.2 Remote sensing variables

We used the MODIS vegetation indices product designed for vegetation studies and the extraction of canopy biophysical parameters (Justice et al., 2002), namely the MODIS-Terra 16-day L3 global standard product with 1 km² spatial resolution (MOD13A2, Version 5). The MODIS product consists of twelve two-dimensional *Science Data Sets* (SDS) including vegetation indices, quality estimates, critical ancillary data such as view and sun zenith angles, and selected surface reflectance bands. In this study, out of the vegetation indices the *Enhanced Vegetation Index* (EVI) was used since it is less susceptible to background soil effects and atmospheric disturbances and does not saturate in high biomass regions (Huete et al., 2002).

For the study period (January 2001 to December 2007), the MODIS tiles h07v05, h07v06, h07v07, h08v05, h08v06, h08v07, h09v05, h09v06, and h09v07 covering the entire Mexican territory were obtained from the *NASA Earth Observing System Data Gateway* (https://wist.echo.nasa.gov/api/). All tiles were mosaicked and reprojected from sinusoidal to geographic coordinates (WGS 1984) using the freely available MRT software (*MODIS Reprojection Tool*, Version 4). For the generation of enhanced EVI time series, the *Time Series Generator* (TiSeG) software developed by Colditz et al. (2008) was applied. The *Quality Assessment Science Data Sets* (QA-SDS) were used in TiSeG to interprete pixel-level quality information and compute two critical indices per pixel - the number of invalid observations and the maximum gap length between two valid observations. Annual EVI time series for the years 2001 to 2007 were produced using linear temporal interpolation between valid observations. Annual phenological metrics (e.g. mean, maximum, minimum, range, standard deviation, date of maximum and minimum) were calculated from the interpolated time series and subsequently averaged over the seven years of the study period. These phenological features account for vegetation seasonality and net primary production as important dimensions for characterizing vegetation type and plant species composition.

2.3 Topographic and climatic predictors

Topographic data acquired during the SRTM mission were aggregated at 1 km² spatial resolution. Additional information layers (slope and aspect) were calculated using ArcMap 9.2 software and integrated as environmental predictors. Moreover, a series of bioclimatic variables was obtained from the WorldClim data base (Hijmans et al., 2005, WorldClim version 1.4, http://www.worldclim.org/bioclim.htm). These climate parameters express spatial variations in seasonality and limiting climatic factors and represent biologically meaningful variables for characterizing species' distributions. Altogether, the WorldClim data layers include eleven temperature and eight precipitation metrics which were developed using long-term time series from 1950 to 2000 of a global network of more than 4,000 weather stations.

The topographic and bioclimatic layers were gridded to pixel location, extent and cell size of the MODIS data in order to maintain spatial consistency with the remote sensing data. All environmental data (climate, topography and time series metrics) were clipped for the extent of the study area (UL: 123.92° W; 39.91° N and LR: 82.61° W; 10.24° N) using DIVA-GIS software (Version 5.4, http://www.diva-gis.org/) and converted to ASCII files. Pixels classified as water according to the MODIS water mask were excluded from the data sets and marked with "no data" flags. Five different sets of explanatory environmental variables were prepared and separately used as input parameters for the SDM: (1) climate; (2) climate and topography; (3) climate, topography, and time series; (4) time series and topography; and (5) time series.

2.4 Maximum Entropy model

We applied the Maxent (Maximum Entropy) algorithm which was recently introduced for species distribution modeling (Phillips et al., 2004). Maxent has been proven to be very useful in comparative studies (Elith et al., 2006) and tested under diverse modeling scenarios, e.g. for different taxonomic groups, different species sample sizes, and a wide range of study region extents. The high computing efficiency of Maxent enables the use of large numbers of input layers covering wide areas as necessary for this study and thus allows for modeling complex responses to environmental variables. The software is freely available and well documented (http://www.cs.princeton.edu/~schapire/maxent/). Maxent models were run in batch mode using the following settings: Auto features, create response curves, make pictures of predictions, do jackknife tests, logistic output format (ASCII), random test percentage = 0, regularization multiplier = 1, maximum iterations = 500, convergence threshold = 0.0001, and maximum number of background points = 10,000. The five different sets of environmental variables (see Section 2.3) were independently used as input explanatory variables for the species distribution models.

3. RESULTS AND DISCUSSION

3.1 Model performance

ROC (*Receiver Operating Characteristic*; Deleo, 1993) plots were obtained by plotting all sensitivity values (true positive fraction) on the y-axis against their equivalent values (1-specificity, false positive fraction) on the x-axis for all logistic thresholds (Figure 2). The plots provide a quantitative representation of trade-offs between omission (1-sensitivity) and commission (1-specificity) errors.

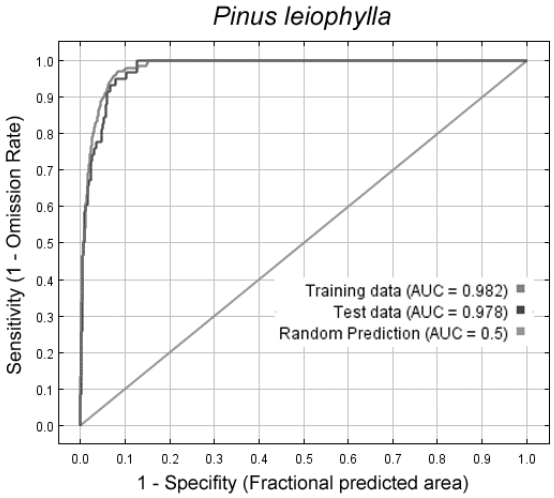

Figure 2. ROC (*Receiver Operating Characteristic*) plot for *Pinus leiophylla* obtained from the Maxent model. Note: Specificity is defined using predicted area rather than true commission which implies that the maximum achievable AUC is less than 1.

Based on the ROC results, AUC (*Area Under ROC Curve*) values were calculated for training and independent test data for the five different sets of environmental predictors (Figure 3). Measured by AUC, the individual model predictions did not differ significantly ($p<0.05$, two-tailed Wilcoxon signed-rank test, pairing by species) from each other for the first three environmental data sets (*climate*; *climate and topography; climate, topography, and time series*) but only between these and the data sets *time series and topography* and *time series*.

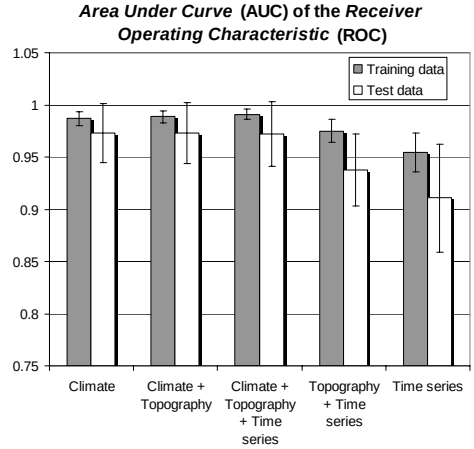

Figure 3. Mean training and test AUC values and standard deviation for all six study species for the five different sets of environmental predictors.

However, all AUC values succeeded the threshold of 0.75 commonly stated as useful for discrimination (Elith et al., 2006) and differed significantly from random predictions ($p<0.01$, one-tailed Wilcoxon signed-rank test). Results evidenced that all sets of environmental predictors selected in this study represented in general adequate multidimensional variables to determine the distributions of the study species. Differences between high quality and close-to-reality against less reliable models can be obtained by direct comparisons of test samples' omission and commission errors. In general, the similarities observed between training and test data AUC support the robust performance of the Maxent algorithm to capture variations in environmental variables over different sets of point localities (Saatchi et al., 2008).

3.2 Remote sensing predictor importance

For all species, remote sensing data contributed less to the Maxent model than climate-topography predictors when both data sets were applied in the same model run. In this case, the relative explanatory contributions by remote sensing predictors alone - which represented 14 out of the overall 36 input variables - were between 0.3 % and 8.5 % (mean 4.3 %). For models with only remote sensing input variables, a clear trend towards major explanatory value of certain indicators was observed (Table 1). Especially annual minimum, mean, and maximum value of the EVI time series contributed to explaining the species distribution. On the other hand, annual range and standard deviation were less unique between species and thus of lower explanatory power. These findings were consistent with characteristics of the species-specific phenology curves derived directly from the time series data and averaged for all available occurrence sites per species.

Tab. 1. Relative explanatory contributions of EVI time series variables. The category "others" includes seven features derived from the first derivation with minor explanatory power.

Species	mean	min	max	range	stddev.	date of min	date of max	others
P. jaliscana	17.2	43.4	16.6	2.5	4.9	3.5	1.9	10.0
P. lumholtzii	19.8	34.4	0.0	0.1	4.4	1.7	12.6	27.0
P. leiophylla	25.8	7.6	30.5	1.3	1.0	7.7	14.7	11.4
Q. arizonica	18.3	12.2	5.3	0.1	1.5	11.5	21.3	29.8
Q. durifolia	3.1	21.2	18.0	0.1	0.2	8.0	14.4	35.0
Q. resinosa	21.3	8.3	19.4	2.0	0.7	18.5	8.4	21.4
Mean value	**17.6**	**21.2**	**15.0**	**1.0**	**2.1**	**8.5**	**12.2**	**22.4**

3.3 Distribution maps

For each species, probability of occurrence maps were derived and converted into binary presence / absence predictions for the logistic thresholds at multiples of 0.1 in the interval from 0.1 to 1.0 and at the maximum sensitivity and specifity threshold (Philipps et al., 2004). Predicted distributions were compared between the five different model scenarios (climate; climate and topography; climate, topography, and time series; time series and topography; time series) with respect to omission and commission errors. We thus propose running two independent models for each species, one with climate and topography data and one with remote sensing variables. Areas predicted as presence by both models accounted for both abiotic (climate and topography) and biotic (time series data) habitat suitability and were superior to models run with climate, topography, and time series data together. The synergistic combination of both information sources allowed for the sharper delineation of the predicted areas and

better exclusion of regions that had suffered human impact and land cover changes, e.g. due to urbanization, agriculture, or degradation (Figure 4).

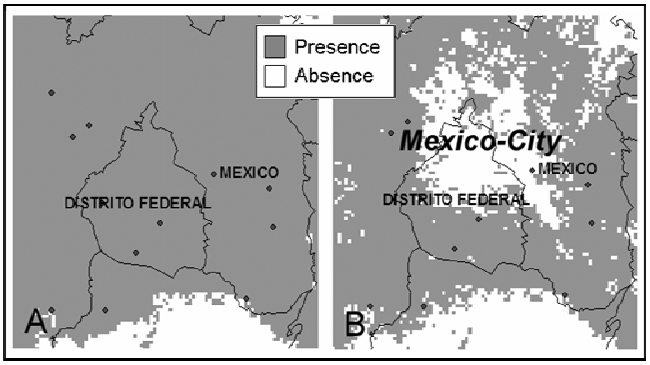

Figure 4. Predicted distribution for *Pinus leiophylla* with (A) climate and topography data and (B) climate and topography data in combination with remote sensing data. The outline of the Mexico-City metropolitan region is clearly depicted. Circles indicate species occurrence point localities.

4. CONCLUSIONS

In general, the reliability of SDM depends on accuracy of species occurrence data, knowledge about the magnitude of sampling biases, selection of (environmental) predictor variables, choice of spatial scale in terms of resolution and extent, and the statistical algorithm employed. In the context of this study, especially species sample size and extent of the species potential range have to be considered as they impact the explanatory power attributed to the remote sensing predictors. The application of continuous remote sensing variables is superior to categorical land cover data since they provide direct measurements related to vegetation structure, species composition, and other ecosystem variables. Irrespective of the fact that reliable and high-quality land cover information is not available worldwide, phenological characteristics based on satellite observations can easily be adapted to the specific ecology of the study species. The suggested approach is thus much more flexible. Potential drawbacks of using MODIS data that have to be taken into consideration are short-term fire events and natural inter-annual variations. Further studies are intended to investigate whether the methodology can be transferred to other taxonomic / ecological groups and study areas. The synergetic combination of parameters derived from remote sensing data with SDMs is still in the fledgling stages and has promising potential for new approaches to be developed in the field of theoretical and applied ecological research. As demonstrated, the opportunity to more accurately map species distributions can significantly contribute to long-term and sustainable conservation management in biodiversity hotspots – especially against the background of accelerating anthropogenic impact and climate change.

REFERENCES

M.P. Austin, “Spatial prediction of species distribution: an interface between ecological theory and statistical modeling”, Ecological Modelling, Vol. 157, 101-118, 2002.

B.A. Bradley and E. Fleishman, “Can remote sensing of land cover improve species distribution modelling?”, Commentary, Journal of Biogeography, Vol. 35(7), pp. 1158-1159, 2008.

R.R. Colditz, C. Conrad, T. Wehrmann, M. Schmidt, and S. Dech, “TiSeG: A flexible Software Tool for Time-Series Generation of MODIS Data Utilizing the Quality Assessment Science Data Set”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 46(10), pp. 3296-3308, 2008.

J.M. Deleo, “Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty”, Proceedings of the first international symposium on uncertainty modelling and analysis, IEEE Computer Society Press, College Park, MD, pp. 318-325, 1993.

J. Elith, C. H. Graham, R.P. Anderson et al., “Novel methods improve prediction of species’ distributions from occurrence data”, Ecography, Vol. 29, pp. 129-151, 2006.

S. Ferrier, G.V.N. Powell, K.S. Richardson et al., “Mapping more of terrestrial biodiversity for global conservation assessment”, Bioscience, Vol. 54, pp. 1101-1109, 2004.

A. Guisan and W. Thuiller, “Predicting species distribution: offering more than simple habitat models”, Ecology Letters, Vol. 8, pp. 993–1009, 2005.

R.J. Hijmans, S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis, “Very high resolution interpolated climate surfaces for global land areas”, International Journal of Climatology, Vol. 25, pp. 1965-1978, 2005.

A. Huete, K. Didan, T. Miura, E.P. Rodriguez, X. Gao, and L.G. Ferreira, “Overview of the radiometric and biophysical performance of the MODIS vegetation indices”, Remote Sensing of Environment, Vol. 83, pp. 195-213, 2002.

R.E. Hutchinson, ”Concluding remarks”, Cold Spring Harbor Symposium on Quantitative Biology, 22, pp. 415-427, 1957.

C.O. Justice, J.R.G. Townshend, E.F. Vermote, E. Masuoka, R.E. Wolfe, N. Saleous, D.P. Roy, and J.T. Morisette, “An overview of MODIS Land data processing and product status”, Remote Sensing of Environment, Vol. 83(1-2), pp. 3-15, 2002.

R.G. Pearson, T.P. Dawson, and C. Liu, “Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data“, Ecography, Vol. 27, pp. 285-298, 2004.

S.J. Phillips, M. Dudik, and R.E. Schapire, “A maximum entropy approach to species distribution modelling”, Proceedings of the Twenty-First International Conference on Machine Learning, pp. 655-662, 2004.

C.D.C. Prates-Clark, S.S. Saatchi, and D. Agosti, “Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data”, Ecological Modelling, Vol. 211, pp. 309-323, 2008.

K.D. Reed, J.K. Meece, J.R. Archer, and A.T. Peterson, “Ecologic Niche Modeling of Blastomyces dermatitidis in Wisconsin”, PLoS ONE, Vol. 3(4), 2008.

S. Saatchi, W. Buermann, H. ter Steege, S. Mori, and T. Smith, “Modelling distribution of Amazonian tree species and diversity using remote sensing measurements”, Remote Sensing of the Environment, Vol. 112(5), pp. 2000-2017, 2008.

W. Thuiller, M.B. Araujo, and S. Lavorel, “Do we need land-cover data to model species distributions in Europe?“, Journal of Biogeography, Vol. 31, pp. 353-361, 2004.

A. Viña, S. Bearer, H. Zhang, Z. Ouyang, and J. Liu, “Evaluating MODIS data for mapping wildlife habitat distribution”, Remote Sensing of Environment, Vol. 112, pp. 2160-2169, 2008.