A scalable approach for evaluating ecological system condition and success of management interventions

Jeffrey S. Evans, Senior Landscape Ecologist & Biometrician
The Nature Conservancy, Global Lands Science
Visiting Professor | University of Wyoming | Ecosystem Sciences
Office, Agriculture - C6, Laramie, WY 82070
jeffrey_evans@tnc.org | (970) 672-6766

INTRODUCTION

To meet a variety of goals at local, regional, continental and global scales, in assessing the status of the ecological systems that we work in, we propose a scalable and hierarchical approach that leverages state-of-the-art machine learning methodology, a suite of remote sensing products and various other earth observation data. By specifically designing a modeling framework that is both scalable and hierarchical we can systematically start "focusing in" on ecological impacts across relevant geographical extents that inform decision making across various levels of our organization. This is a critical consideration because, information that is critical to local land managers may not be relevant to policy and decision makers responsible for very large geographies. A hierarchical framework allows for generating results that are relevant across large geographies but remain relevant in informing results as the analytical scale (eg., process, spatial) is increased and more locally relevant data provided.

STUDY AREA

In providing detailed evaluations of our modeling approach, we are planning on implementing case studies in the Crown of the Continent (encompassing the northern portion of the Rocky Mountains, US/Canada); northern Australia indigenous lands (Australia); and East Kalimatan (Borneo, Indonesia). The first geography is an example of TNC's work on land protection and habitat management, whereas the latter two are represent areas of TNC work with indigenous peoples.

METHODS

SPATIAL SAMPLING AND RESPONSE (DEPENDENT) VARIABLE

As a first step, we must consider how a sampling schema will be defined to address our specific problem. In this case two critical assumptions should be met: 1) the sample captures the spatial variation across the region of interest and; 2) the sample captures the variation in biophysical condition. This is to ensure that latent spatial patterns are captured, nonstationarity does not bias the sample and ecological responses are appropriately represented given how characteristics, such as topography, effect occupancy, abundance and productivity.

The dependent variable, defining the *y* side of a statistical equation, represents the desired outcome of a modelling effort. First, for our sampling approach we will generate a spatially balanced (Stevens & Olsen 2012) stratified random spatial point sample based on spatial variables that represent a temperature and moisture gradient (i.e., solar insolation, slope and aspect interaction, temperature, precipitation). As an alternative to creating nominal variables for the stratification we will use gradient boosting (Friedman 1999) to perform a nonlinear regression that fully captures the distributional qualities of each stratifying covariate. The output will be probabilistic, and sample selection will be based on quantiles of the probabilities. This

will allow us to automate the creation of large spatial samples, for use in creating a response variable, that capture the biophysical gradient. Given that this is a hierarchical framework, we will define the first level of our hierarchy, suitable for large scale assessment of system status, using a 15-year trend slope of Leaf Area Index (LAI) coupled with a weighted Kappa (Cohen 1968) of landcover change, temporal variation of LAI (variance within a 1KM window) and finally, fractional cover of vegetation.

The LAI is a dimensionless ratio of leaf area to per unit ground surface area. This ratio has been related to processes such as photosynthesis, evapotranspiration and carbon flux (Leuning et al., 2005). Long-term monitoring of LAI can provide an understanding of dynamic changes in productivity and impacts on vegetated systems. Furthermore, LAI can serve as an indicator of stress in forests, thus, it can be used to examine relationships between environmental stress and change vectors. The temporal-slope in LAI will be derived using the nonparametric Kendal tau with the Theil-Sen slope repeated-medians approach (Siegel 1982). The weighted Kappa statistic, weighted proportional change corrected for potential random agreement, is a well-accepted and easy to interpret statistic utilized in change detection (Van Vliet 2011). The results represent the proportional change in nominal classes at the pixel-level, given a surrounding (focal) area. The advantage of using a weighted Kappa is that not all change is equal. If a landcover transitions from forest to grassland it should not be treated the same as if it is converted to urban. By weighting the transitions, we can account for (penalize) certain transitions and make the resulting kappa metric much more relevant. Our dependent variable will then be defined using scaled versions of temporal LAI trend (with a fixed intercept) and kappa's resulting proportional change. These two variables, used in concert, will allow us to evaluate both vegetation condition, as a function of above-ground biomass, and functional changes in landscape pattern.

To this end, we will utilize Moderate Resolution Imaging Spectroradiometer (MODIS) TERRA sensor data (Friedl et al., 2010) for evaluating 8-day LAI composites and annual landcover. This multi-temporal global sensor data covers a broad timespan (16-17 years) and can be readily utilized for global analysis. Specific MODIS products will include the MODIS 500m annual land cover (Land Cover Type 1: IGBP global vegetation classification scheme) to track changes in land cover over time and MODIS 500m 8 Day LAI/FPAR for temporal LAI composites.

This adaptable method provides a generic framework for automating the generation of large stratified samples, using globally commonly available data, and for generating a response variable based on well accepted temporal landscape characteristics that, depending on satellite sensor, can be captured at various spatial scales. We will also explore using phenological characteristics (eg., growing season length, intra-annual variability, or green-up date) as additional metrics for LAI to improve interpretation.

PREDICTOR (INDEPENDENT) VARIABLES

For our independent (predictor) variables, in addition to contemporary satellite imagery, we will include covariates that account for environmental variability so that we do not miss latent (hidden) processes that may influence estimates of condition and change. These covariates will include climatic variability and deterministic processes influenced by topography. However, one should note that the modeling framework is adaptable across spatial and process scales making it flexible for a variety of sensors or data resolutions. A common resolution will be identified for the analysis, based on sensor data, and robust nonlinear local polynomial regression upscaling approaches will be implemented for resampling.

IMPUTATION MODEL

For estimating ecological condition and change we will implement a non-parametric version of *kNN* multiple imputation using random forests (Crookston and Finley 2008, Hudak et al. 2008). This method is

a very powerful, adaptive machine learning algorithm that will allow us to not only produce continuous landscape estimates of trend and change but also evaluate a composite of condition and change based on the imputed multivariate distances. Here we leverage the characteristics of the proximity matrix, produced by the random forests algorithm, to derive multivariate distances in hyper-dimensional statistical space. Since temporally-stochastic characteristics are accounted for in the training samples, we can use single-date satellite and deterministic data to model current landscape condition. Provided estimates will include: imputed proportional change, condition based on trend slope, and multivariate distances the incorporate both proportional change and trend slope. This *kNN* model allows multiple response variables (*y*) and a large set of independent (*x*) variables with no distributional assumptions or need for data transformations. One notable advantage is that high dimensional interactions are accounted for on both the *x* and *y* sides of the equation. For model specifications we will then apply a model selection procedure following Evans *et al.*, (2011) to identify key parameters. Estimates can be made to the continuous raster data used in specifying the model or extrapolated to a different time-point, of imagery, of a comparable geographic extent. This allows one to explore hypothesis representing different environmental (eg., climatic), landcover or trend conditions (eg., past-date imagery, simulated trends, simulated landcover transitions).

Model validation will be specific to each model (study area) but, also be implemented as an automated process, using a Bootstrap approach to generate an error distribution in evaluating model performance. We will also evaluate model fit utilizing common validation metrics such as sensitivity, specificity, AUC and log loss. In the case of continuous data, we will evaluate log-likelihood loss and root mean square error. For our trend analysis, a confidence region, p-value and z-value will be generated, at the pixel-level, for the temporal series. This will provide a means of evaluating the significance of the trend results and to filter out pixels that do not fall within a specified +/- statistical confidence

CAUSAL IMPACT ANALYSIS

Many methods have been proposed in evaluating the success of interventions. Some common approaches "Difference in Differences" (Ashenfelter & Card 1985) and "*Synthetic Control Matching*" (Abadie *et al*., 2010) are widely used in econometrics and social sciences but have some noted limitations. The most relevant is detailed in Kaul *et al.,* (2017) where the use of pre-intervention outcomes used together with covariates is statistically invalidated. This negates using covariates that indicate land tenure structures that would influence both evaluation and control units. The Synthetic Control method relies on the linearity of the model in relation to treated and untreated outcomes. It is likely that this assumption will be violated thus invalidating both the global and local model(s). As an alternative we will utilize as nonparametric structural model Causal Impact Analysis (Brodersen *et al.,* 2015) that allows for a causal influence through a MCMC framework. In essence, we build a Bayesian structural time series model based on multiple comparable control groups and then use the model to project (or forecast) a series of the baseline values for the time period after an intervention, representing a synthetic time series baseline of all possible state spaces of the timeseries thus, making it causal in nature. In social sciences parlance, this functionally establishes our counterfactual as a testable causal state-space. Now, with the Causal Impact algorithm, we can build a model based on other locations to project a series of the expected values for the same time period assuming there was no such 'Impacting Event' that occurred in these control units. This acts as a baseline, which indicates the numbers we would have expected in controls without the intervention(s). Once we establish this baseline, then we can calculate the differences between the two timeseries and evaluate the differences as the real impacts of the intervention. This can be specified using a mixed effects model where relevant land tenure, expected to influence outcomes, can be accounted for as a random effect. However, will implement a Dynamic Time Warping (DTW) algorithm to evaluate the temporal correlation structures (agreements) between controls and interventions, which negates the need for treating potential land-tenure effects as these "biased" controls would never be selected in the first place.

DISCUSSION

One advantage to this modeling framework is that we do not need to represent temporal characteristics in the "contemporary" design matrix (*x*), representing current condition, as it is inferred that current landscape condition/pattern is a function of the spatial-temporal process that formed them so, does not need to be directly accounted for. The independent data (*y*) however, does represent this spatial-temporal variability (which is what is being estimated) so, is accounted for within the training set. This provides support for basing model estimates and multivariate distances on the contemporary imagery and deterministic data alone.

One important caveat is that these measures of ecological status are contextual, dependent on the system under evaluation and, to some degree, the agents of change. Certain change could be seen, locally as catastrophic, but not identified at broad spatial scales. One good example of this would be increased above-ground biomass that is attributed to invasive species. Whereas, this would locally be a very undesirable outcome, our first-level analysis would not identify this as a negative outcome. This is precisely why we are proposing a hierarchical approach where, by simply adding additional information that would put this invasive species process in context for the model, it would then be accounted for. We need tools that allow for rapid, large-extent evaluation of ecosystem status. Given this, it is not plausible to account for every global ecological caveat so, the first level of our hierarchy is intended to represent the broad-scale characteristics in landcover pattern and vegetation trend. The hierarchies are represented by how the independent and dependent data is structured and subsequent levels in the hierarchy can then build on this, functionally having the higher-levels act as a first-order trend in the model. In theory, adding more resolute information can be represent down to a site-specific level if desired. The end-result will be reflected in the multivariate distances.

Interpretation of model results could take many forms, from direct estimation of the dependent variable to interpretation of the imputed multivariate distance. In evaluating interventions one could leverage the imputed distances by taking the 95^{th} percentile of historic LAI, representing stable to highest increase in slope and the highest Kappa agreement ($k>=0.65$), and then evaluate the multivariate distances to these observations, which would functionally act as "reference conditions". These thresholds can easily be changed to evaluate different aspects of change and condition. For example, in evaluating imputed multivariate distances to "ideal" reference condition(s) we would look at pixels that exhibit the smallest multivariate distances to these specific observations. These would be the pixels that match these reference conditions thus, exhibiting similar spatial-temporal characteristics in the current landscape. This pixel subset could then be used in the context of land tenure to evaluate proportional success in both condition and change. This is the huge advantage of "multiple" imputation, the multiple variables used as the response variables interact in a way that the distances can be interpreted as a function of neighbor similarity to all the responses and not just one at a time.

In addition to conducting these analyses across an ecosystem type or a region, these analyses can be scaled to very localized areas of interest to be compared with random (but biophysically similar) areas. For example, to evaluate success of a project, results can be evaluated for a TNC project area and a suite of nearby control sites following certain assumptions. The values for the TNC project area could then be compared to the distribution of the values of control sites as a metric of whether our actions are having

detectable effects. The general workflow detailed above can also contribute to modeling of ecological condition based on new data, for more rapid, near-real-time analyses, even in new systems.

REFERENCES

Abadie, A., A. Diamond, J. Hainmueller. (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Ashenfelter, O., & D. Card. (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J., (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 213–220

Crookston, N. L. and A.O. Finley (2008). yaImpute: An R package for kNN imputation. *Journal of Statistical Software*. 23(10). 16 pp.

Brodersen, K.H., F. Gallusser, J. Koehler, N. Remy, S.L. Scott (2015) Inferring Causal Impact using Bayesian Structural Time-series Models. *The Annals of Applied Statistics* 9(1):247-274

Evans J.S., M.A. Murphy, Z.A. Holden, S.A. Cushman (2011). Modeling species distribution and change using Random Forests in Predictive species and habitat modeling in landscape ecology: concepts and applications. eds Drew CA, YF Wiersma, F Huettmann. Springer, NY

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327.

Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Machine. Technical report, Stanford University.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X. (2010). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment, 114, 168–182.

Garrigues, S., Allard, D., Baret, F., Weiss, M., (2006). Influence landscape spatial heterogeneity on the non-linear estimation of leaf area index from moderate spatial resolution remote sensing data. Remote Sensing of Environment 105:286-298.

Hudak, A.T., N.L. Crookston, J.S. Evans, D.E. Hall and M.J. Falkowski. (2008). Nearest neighbor imputation modeling of species-level, plot-scale structural attributes from LiDAR data. Remote Sensing of Environment 112:2232-2245.

Leuning, R., Cleugh, H.A., Zegelin, S.J., Hughes, D. (2005) Carbon and water fluxes over a temperate Eucalyptus forest and a tropical wet/dry savanna in Australia: measurements and comparison with MODIS remote sensing estimates. Agric. For. Meteorol. 129:151-173.

Stevens, D.L. & A.R Olsen (2004) Spatially Balanced Sampling of Natural Resources, Journal of the American Statistical Association, 99:465, 262-278, DOI: 10.1198/016214504000000250

Sen, P.K. (1968) Estimates of Regression Coefficient Based on Kendall's tau. Journal of the American Statistical Association. 63(324):1379-1389.

Siegel, A.F. (1982) Robust Regression Using Repeated Medians. Biometrika, 69(1):242-244

Van Vliet, J., Bregt, A., and Hagen-Zanker, A. (2011) Revisiting Kappa to account for change in the accuracy assessment of land-use change models. Ecological modeling 222, pp. 1367-1375.