

A big data and machine learning approach for monitoring the condition of ecosystems

Abstract—Ecosystems are highly valuable as a source of goods and services and as a heritage for future generations. Knowing their condition is extremely important for all management and conservation activities and public policies. Until now, the evaluation of ecosystem condition has been unsatisfactory and thus lacks practical implementation for most countries. We propose that ecosystem integrity is a useful concept that can be used to evaluate ecosystem condition through data science and machine learning. Based on a three tier (contextual, instrumental and hidden) model and a Bayesian network approach, we used field and remote sensing data to estimate the integrity of terrestrial ecosystems per 250 m in Mexico.

I. INTRODUCTION

CURRENTLY, there is a growing capacity worldwide to produce large amounts of environmental data (probably on the ranks of Big data), and the growing technological development to obtain and analyze them allow to tackle human and ecological problems under new paradigms[1]. Today it is possible to conceive spatially continuous maps that can be updated in almost “real time”, depending on the frequency of new available information[2]. One of such interests focus on Ecosystems because, despite being regarded as an externality to the economy, they are highly valuable as a source of goods and services and as a heritage for future generations. In particular, there is growing interest in monitoring their condition as a basis for all management and conservation activities and public policies. Mexico has produced big datasets, such as the [National Forest and Soil Inventory](#) (INFyS, for its Spanish acronym), the [National Information System on Biodiversity](#) (SNIB, for its Spanish acronym) the [National System of Biodiversity Monitoring](#) (SNMB, for its Spanish acronym). These, in combination with publicly available satellite imagery (e.g. MODIS, Landsat, Sentinel 1 and 2 collections) may progressively be used to assess the natural condition of the country[3].

In contrast to combining variables in *ad hoc* indices as has been done before[4][5], we capitalize on the large data-sets described above and machine learning technologies to produce an ecosystem integrity estimate. We trained a Bayesian network[6] to predict the value of an integrity index that relates to the condition an environmental unit might have at a point in time using a conceptual model and learning algorithms. The result is strongly data driven and science based on an explicit concept, which makes it both innovative and highly reproducible, a relevant contribution in support of evidence based decision-making.

II. THE THREE TIER MODEL FOR ECOSYSTEM INTEGRITY

Ecosystem integrity emerges from the driving of both natural and anthropogenic processes which operate concurrently

over the ecosystem. However, in order to enhance clarity we follow an analytical strategy of separating these processes taking advantage of the modularity of Bayesian network that allows for an object-oriented approach [7]. Thus, we developed a three tier model that accounts for the condition in which the ecosystem is, based on a referent of non-human intervention. In this model, observations obtained by sampling in the field or through remote sensing, are allocated to the “instrumental tier” (Figure 1). We assume that the actual values of the variables in this tier are a result of the simultaneous effect of two components: a) the physical and chemical conditions of existence (also conditioning the evolutionary lineages present in the area) and b) the current condition of the ecosystem. The former constitutes a “contextual tier” in our model and the latter a “hidden tier”. The contextual tier accounts for the physicochemical conditions within which the ranges of values of the variables from the instrumental layer express themselves (conditioning expected values). The hidden layer defines the level of ecosystem integrity based on the values of the instrumental and contextual layer (expected outcome of human intervention).

Human intervention is added as an extra tier that allows the coupling with key drivers that can be hypothesized are preconditions acting over an environmental unit, that are likely to affect ecosystem integrity (not shown in Figure 1).

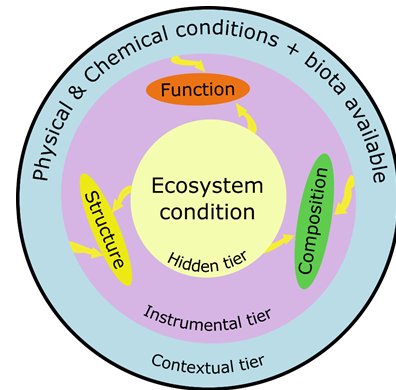


Fig. 1. The three tier conceptual model of ecosystem integrity.

III. METHODS

A. Contextual tier

Data on precipitation, biotemperature and potential evapotranspiration ratio were used to assess physicochemical conditions and calculate [thirty one life zones](#) for the Mexico according to Holdridge (1967), using the [nomenclature](#) proposed by the International Institute for Applied Systems Analysis. [Data](#) from a digital elevation model was included in order to

account for the physical conditions imposed by elevation and local hilliness.

B. Instrumental tier

1) *Field data*: Field data were obtained from INFyS, which Mexico has been conducted since 2004 over a grid of more than 22,000 conglomerates. These are sampled iteratively over a cycle of five years. The grid is evenly spaced depending on the type of ecosystem in question, 5 km for forests and 25 km for arid and semi-arid ecosystems. INFyS includes the sampling of over 150 variables. From these, we selected ten to produce wall-to-wall cartography at 250 m by means of machine learning based spatiotemporal interpolation. To achieve this, the *In Situ* INFyS data was associated to several covariates, which are available continuously over Mexico. A first batch of these were remote sensing derived, yearly composites of Modis Vegetation Indices 16-Day L3 Global at 250 m (MOD13Q1 and MYD13Q1) were produced: $P_{0.05}$, $P_{0.20}$, mean, $P_{0.80}$, $P_{0.95}$ mean of the dry season¹ and mean of the wet season for both NDVI and EVI indices. The second batch of covariates are climatic and topographic: high resolution (90 m) bioclimatic surfaces[8] and a digital elevation model (mean elevation and range).

INFyS variables which were continuous in nature: number of trees and shrubs (diameter ≥ 7.5 cm) per ha, average tree height, standard deviation of tree heights, average tree crown diameter, standard deviation of tree crowns diameter, average stem height, standard deviation of stem heights, average diameter at breast height and standard deviation of diameters at breast height, were used to fit XGBoost[9] regression models. Variables which were discrete: presence/absence of tree pests, presence/absence of standing dead trees, were used to fit Random Forest classification models[10]. Once these predictive models were trained and tuned, they were used to predict on the whole of Mexico to produce the desired cartography.

2) *Remote sensing*: Additionally, as a proxy for vegetation function on the ground, Modis Net Photosynthesis products (MOD17A2 and MYD17A2) were used. The complete available time series of this product was downloaded and then yearly composites were created: mean net photosynthesis, standard deviation of net photosynthesis and mean of the wet and dry seasons.

C. MAD-MEX land cover

Medium resolution (30 m) MAD-MEX[11] (Monitoring Activity Data – Mexico) land cover classification maps were used to generate coarse resolution proportion of cover maps at 250 m. First, the original MAD-MEX scheme was aggregated to IPCC classes, except that forest, rainforest and shrubland were separated and grassland and agriculture aggregated. Then, these 30 m resolution classification maps were overlaid on a 250 m grid and the proportion of each class contained on each 250 m pixel calculated (Figure 2).

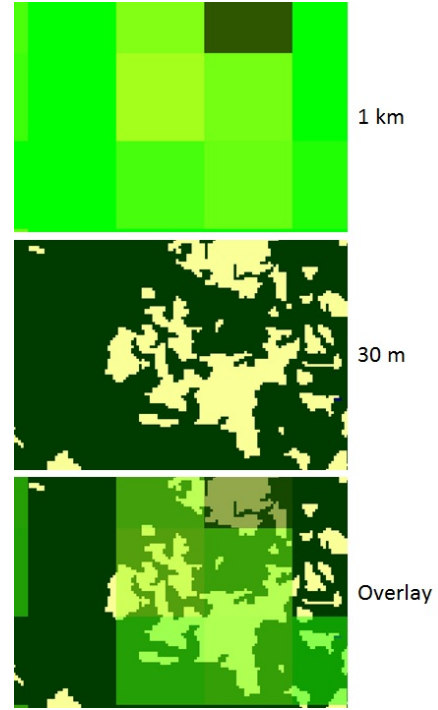


Fig. 2. Example of spatial overlay: 30 m raster on 1 km grid.

The percentage of a pixel covered by forest, rainforest, shrubland were considered as instrumental variables and bare-ground, grassland or agriculture and human settlements, as preconditions that affect ecosystem integrity.

IV. BAYESIAN NETWORK MODELING

All variables were automatically imported to Netica as continuous nodes then discretized to 10 levels, based on the particular histogram pattern of each node. A Bayesian network model was used to describe the influence among the variables as well as to provide estimates of the conditional probability matrices. Bayesian networks are represented by an directed acyclic graph with variables as nodes linked by arrows pointing in the direction of the influence (Figure 3). The model further specifies the dependencies by means of matrices of conditional probability that account for the set of dependencies the variables have, one matrix per node, which completes the quantitative specification of the model.

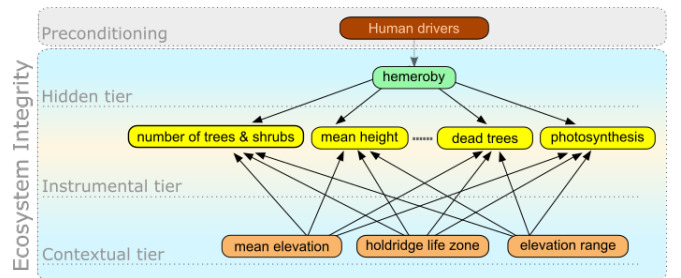


Fig. 3. Generic Ecosystem Integrity Bayesian network.

For network structure we followed a mixed strategy. First, we applied a Tree Augmented Naïve algorithm (TAN)[12] to

¹January, February, March, April and December

help on discovering co-dependency pattern in the data set, then we superimposed the three tier structure (contextual and hidden layers → instrumental layer) plus preconditioning so as to produce the largest correlation with the ecosystem condition primes. All Conditional probabilities tables were obtained using the Counting-Learning algorithm embedded in Netica.

In order to prime the hidden tier we prepared a “Delta Primary Vegetation” (DeltaPV) map, based on the hemeroby concept [13], that estimates the amount of transformation that vegetation shows by comparing **current** land cover obtained from classified satellite imagery of the year 2008 versus **expected** “primary vegetation”, as judged by expert opinion. The qualitative amount of change was evaluated following the criteria showed in Figure 4.

QUALITATIVE AMOUNT OF CHANGE FROM PRIMARY VEGETATION TO "CURRENT" LAND USE AND VEGETATION		
DeltaPV	DESCRIPTION:	STATUS:
0	NO CHANGE OF VEGETATION OR PRIMARY STATE	STASIS
1	NO CHANGE IN PRIMARY STATE CONDITION, THE CHANGE IS IN VEGETATION TYPE BUT INSIDE THE SAME ECOVARIANT (e.g. FOREST, SHRUBLAND, GRASSLAND)	PSEUDOESTASIS
2	NO CHANGE IN PRIMARY STATE CONDITION, THE CHANGE IS IN VEGETATION TYPE TO ANOTHER ECOVARIANT BUT WITH INCREASE IN THE HEIGHT OF THE DOMINANT STRATUS (e.g. FROM SHRUBLAND TO WOODLAND)	PSEUDOESTASIS
3	NO CHANGE IN PRIMARY STATE CONDITION, THE CHANGE IS IN VEGETATION TYPE TO ANOTHER ECOVARIANT BUT KEEPING THE HEIGHT OF THE DOMINANT STRATUS (e.g. FROM OAK FOREST TO TROPICAL DECIDUOUS FOREST)	PSEUDOESTASIS
4	NO CHANGE IN PRIMARY STATE CONDITION, THE CHANGE IS IN VEGETATION TYPE TO ANOTHER ECOVARIANT BUT DECREASING THE HEIGHT OF THE DOMINANT STRATUS (e.g. FROM SHRUBLAND TO GRASSLAND)	DEGRADATION
5	CHANGE FROM PRIMARY FORESTS TO ABBREVIATED SECONDARY VEGETATION	DEGRADATION
6	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO INDUCED FOREST OR TREE PLANTATION	DEGRADATION
7	CHANGE FROM PRIMARY FORESTS, TROPICAL FORESTS, WOODLANDS OR SHRUBLANDS TO SHRUBBY SECONDARY VEGETATION	DEGRADATION
8	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO SAVANNA VEGETATION	DEGRADATION
9	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO HERBACEOUS SECONDARY VEGETATION	DEGRADATION
10	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO INDUCED PALM GROVE, OR INDUCED OR CULTIVATED PASTURE	DEGRADATION
11	CHANGE FROM NATURAL BODY OF WATER TO AQUACULTURAL USE	DEGRADATION
12	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO RAINFED AGRICULTURE	DEGRADATION
13	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO IRRIGATION AGRICULTURE	DEGRADATION
14	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO HUMIDITY AGRICULTURE	DEGRADATION
15	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO BODY OF WATER	DEGRADATION
16	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO BARE SOIL	DEGRADATION
17	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO HUMAN SETTLEMENTS OR IMPERVIOUS SURFACES	DEGRADATION
18	CHANGE FROM ANY KIND OF PRIMARY VEGETATION TO URBAN ZONE OR IMPERVIOUS SURFACE	DEGRADATION
-9999	NO DATA	NA

Fig. 4. DeltaPV: Deviation between **primary** and current vegetation based on INEGI vegetation series IV, as judge by expert opinion.

V. BAYESIAN NETWORK PREDICTION

The final Bayesian network may be used to estimate the level of Ecosystem integrity for any year where the variables of the three tier model are available. The original Bayesian network was trained on data comprising the year 2008 but may be used to extrapolate to, for example, the year 2004 (Figure 5) or to any recent year. It should be noted that even though the Bayesian network in question is trained in classification mode (DeltaPV is an ordinal variable), the expected value along its numeric range can be calculated, which is what we refer to by the Ecosystem Integrity Index.

A Bayesian network can be query in multiple ways, which delivers different “reasoning types” (mainly diagnostic, causal or intercausal) [7]. In a causal reasoning, for instance we could ascertain the values of all nodes that correspond to a prime condition to yield reference values.

VI. AUTHORS

Miguel Equihua, Octavio Pérez-Maqueo, Julián Equihua, Pedro Maeda, Michael Schmidt, Melanie Kolb, Nashieli Garcia-Alaniz, Mariana Munguía-Carrara, Sergio Villela, Oliver López, Everardo Robredo, Santiago Martínez.

VII. ACKNOWLEDGEMENTS

We are grateful by the support to “Integralidad Gamma” project (i-Gamma) provided by FORDECyT fund from CONACyT, Mexico under grant number 296842.



Fig. 5. Ecosystem integrity index projected to 2018, using the model trained on 2008 data.

REFERENCES

- [1] S. Hampton, C. Strasser, J. Tewksbury, B. Gram, W.K., A. A.E., Batcheller, C. Duke, and J. Porter, “Big data and the future of ecology,” *Front. Ecol. Environ.*, 2013.
- [2] S. Hay, D. George, C. Moyes, and J. Brownstein, “Big data opportunities for global infectious disease surveillance,” *PLoS Med*, 2013.
- [3] N. Garcia-Alaniz, M. Equihua, O. Pérez-Maqueo, J. Equihua Benítez, P. Maeda, F. Pardo Urrutia, J. Flores Martínez, S. Villela Gaytán, and M. Schmidt, “The mexican national biodiversity and ecosystem degradation monitoring system,” *Curr. Opin. Environ. Sustain.*, 2017.
- [4] B. Halpern, C. Longo, D. Hardy, K. McLeod, J. Samhouri, S. Katona, K. Kleisner, S. Lester, J. O’Leary, M. Ranelli, A. a. Rosenberg, C. Scarborough, E. Selig, B. Best, D. Brumbaugh, F. Chapin, L. Crowder, K. Daly, S. Doney, C. Elfes, M. Fogarty, S. Gaines, K. Jacobsen, L. Karrer, H. Leslie, E. Neeley, D. Pauly, S. Polasky, B. Ris, K. St Martin, G. Stone, U. Sumaila, and D. Zeller, “An index to assess the health and benefits of the global ocean,” *Nature*, 2012.
- [5] G. Tierney, D. Faber-Langendoen, B. Mitchell, W. Shriver, and J. Gibbs, “Monitoring and evaluating the ecological integrity of forest ecosystems,” *Front. Ecol. Environ.*, 2009.
- [6] J. Pearl, “Bayesian networks: A model of self-activated memory for evidential reasoning,” *Report, Los Angeles University of California Computer Science Department Volume 850021 of UCLA CSD*, 1985.
- [7] D. Koller and A. Pfeffer, “Object-oriented bayesian networks,” *Proceedings of the 13th conf. on Uncertainty in artificial int.*, 1997.
- [8] C.-R. A. P., O. Téllez-Valdés, M. Gómez, C. Venegas-Barrera, J. Manjarrez, and E. Martínez-Meyer, “An update of high-resolution monthly climate surfaces for mexico. international journal of climatology,” *International Journal of Climatology*, 2013.
- [9] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [10] L. Breiman, “Random forests,” *Machine Learning*, 2001.
- [11] S. Gebhardt, T. Wehrmann, M. Ruiz, P. Maeda, J. Bishop, M. Schramm, R. Kopeinig, O. Cartus, J. Kellndorfer, R. Ressler, L. Santos, and M. Schmidt, “Mad-mex: Automatic wall-to-wall land cover monitoring for the mexican redd-mrv program using all landsat data,” *Remote Sensing*, 2014.
- [12] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Machine Learning*, 1997.
- [13] A. Machado, “An index of naturalness,” *Journal for Nature Conservation*, vol. 12, no. 2, pp. 95–110, 2004.

