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).

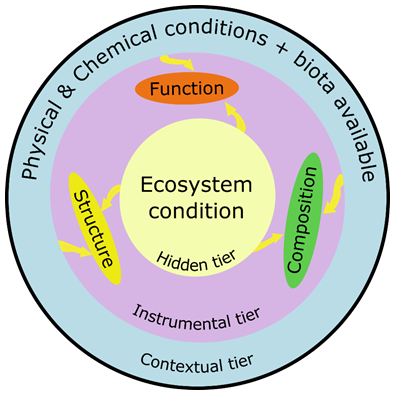

Fig. 1. The three tier conceptual model of ecosystem integrity.

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 season[1] 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 shurbs (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 where trained and tuned, they where 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) where used. The complete available time series of this product was downloaded and then yearly composites where 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).

[1]January, February, March, April and December

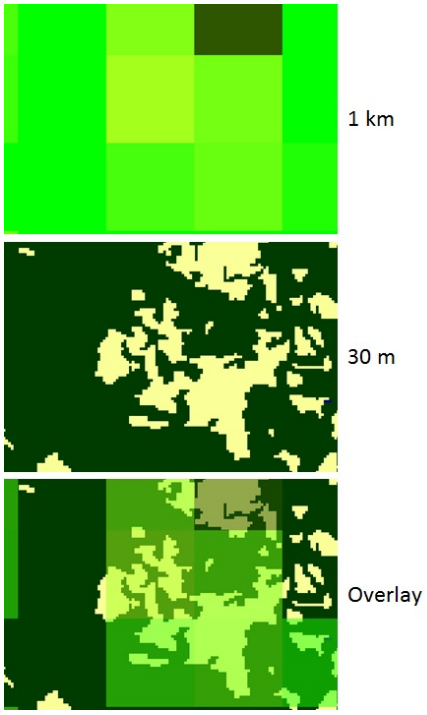

Fig. 2. Example of spatial overlay: 30 m raster on 1 km grid.

The percentage of a pixel covered by forest, rainforest, shrubland were considered as instrumental variables and bareground, 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.

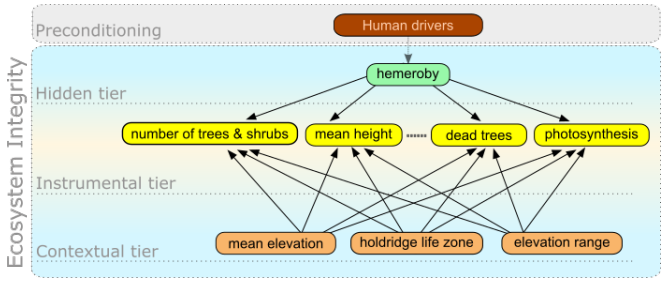

Fig. 3. Generic Ecosystem Integrity Bayesian network.

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