Universidad Nacional Autónoma de México

Coordinación de la Investigación Científica UNAM

C3 Centro de Ciencias de la Complejidad

Bayesian Classifiers:

A universal approach to predicting patterns of land-use change and its causes

un centro TRANSVERSAL para la UNAM

DR. CHRISTOPHER R. STEPHENS

Coordinador de Investigación del C3

Investigador Instituto de Ciencias Nucleares

Grupo CHILAM

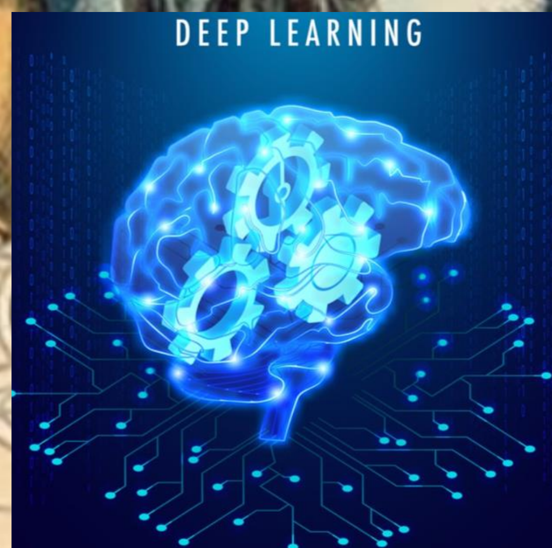
Will I conquer the entire world?

Where will we have the highest levels of deforestation?

When will we reach a tipping point for melting of the Greenland ice sheet?

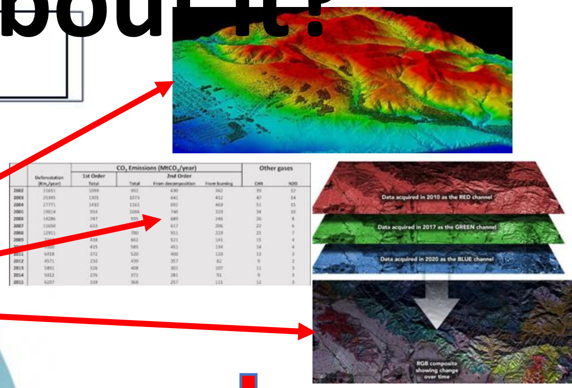
What is this object in my remote sensing data?

Who is most responsible for the deforestation at a given place?



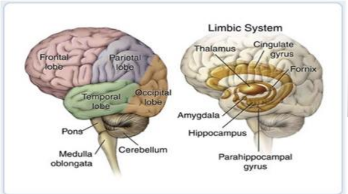
The answers to these questions are all predictions

What do we/can we do about it?



El programa Sembrando Vida es una política en materia agroecológica puesta en marcha por el gobierno de Andrés Manuel López Obrador en el que se propone atender a la población marginada en las regiones de alta biodiversidad mediante un subsidio al cual pueden acceder quienes son propietarios de 2.5 hectáreas que se encuentren deforestadas. El subsidio incluye, además de un apoyo económico y de semillas entre otros recursos, una serie de reuniones con los beneficiarios y los funcionarios involucrados en el programa, explicó Omar Felipe Giraldo, profesor de la Escuela Nacional de Estudios Superiores Unidad Mérida de la Universidad Nacional Autónoma de México (UNAM).

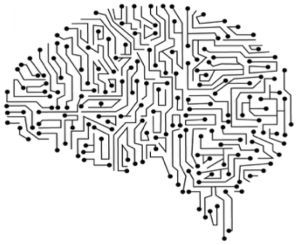
Sin embargo, tras un proceso de análisis y escucha de testimonios, se ha podido identificar una serie de deficiencias que generan que el programa no cumpla con sus metas y que tampoco coadyuve a los desafíos del campo mexicano ya existentes. En primer lugar, el programa no ha llegado a favorecer la reforestación del país, si no que se ha notado que alrededor de 72,830 hectáreas se perdieron debido a Sembrando Vida, esto ya que con el fin de poder obtener el beneficio, algunos propietarios deforestaron sus terrenos para cumplir con el requisito, añadió Giraldo.



Human Intelligence

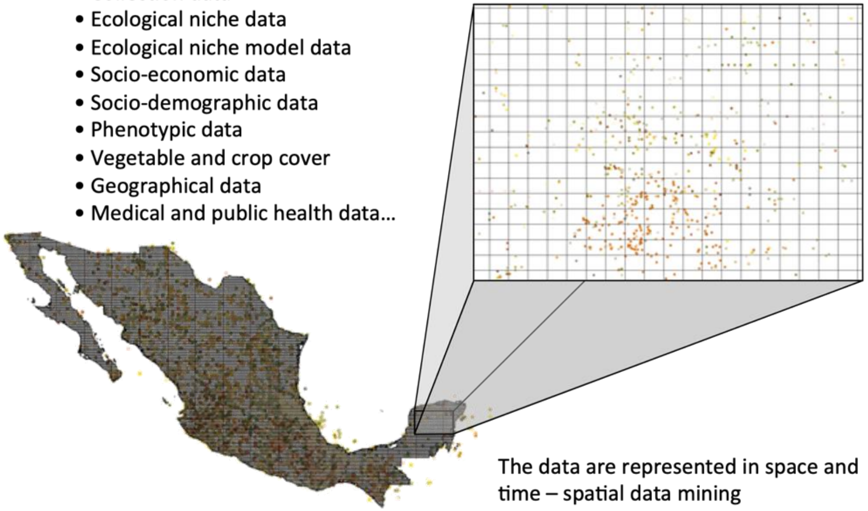
$$P(C|X)$$

Bayesian Classifier

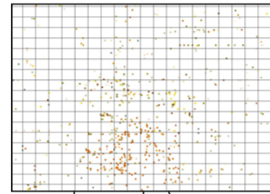


Artificial Intelligence

- Collection data
- Ecological niche data
- Ecological niche model data
- Socio-economic data
- Socio-demographic data
- Phenotypic data
- Vegetable and crop cover
- Geographical data
- Medical and public health data...

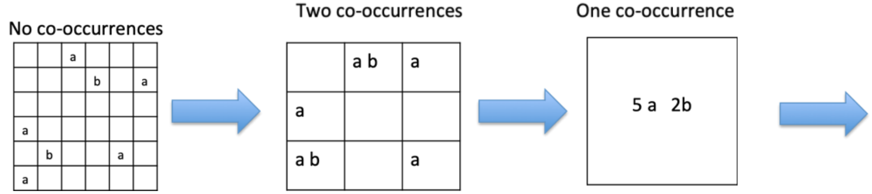


The data are represented in space and time – spatial data mining



Choose a spatial resolution: give everyone one vote there.
The “Senate” versus the “Congress” approach!

Cuadrante	Sigmodon hispidus	Dipetalogaser maxima	Casos Chagas	Precipitación anual	Temperatura promedio	GARP Triatoma maximus	GARP Diptaloster maxima	Perfil agricola
A1	1	3	1	23	18.6	1	1	4
A2	0	1	0	23	18.6	1	1	4
A3	0	2	0	23.7	18.7	1	1	1
A4	0	4	0	23.7	18.7	1	1	3
A5	0	2	1	23.7	18.7	1	1	3
A6	2	5	2	23.7	18.7	1	1	2
A7	0	1	0	23.3	18.4	1	1	5
A8	0	2	0	22.8	18.8	1	1	3
A9	1	3	1	22.8	18.8	1	1	1
A10	0	1	0	22.8	18.8	0	1	1
A11	0	0	0	22.8	18.8	0	1	1
A12	0	0	0	22.8	18.8	0	1	2
A13	0	0	0	22.8	18.8	0	0	4
A14	0	0	0	22.8	18.8	0	0	3
A15	0	2	0	22.8	18.8	0	1	4
A16	0	1	0	22.8	18.8	0	1	2
A17	0	0	0	22.8	18.8	0	1	1
A18	0	0	0	22.8	18.8	0	0	1



N(ab)/N(b)
= P(a|b)

1. For a given class C and set of predictors $\mathbf{X}$, partition a region of space, $\mathcal{R}$, and interval of time, $\mathcal{T}$, into cells, α, to form an ensemble $\mathcal{E}$ of N cells
2. Convert all variables C, $X_i \in \mathbf{X}$ into a set of binomial variables X_{ij} on $\mathcal{E}$, i.e., at the same spatial and temporal resolution
3. For each cell α, determine the presence/absence of each $X_{ij}(\alpha)$
4. Define co-occurrences $(C(\alpha), X_{ij}(\alpha), (X_{ij}(\alpha), X_{i'j'}(\alpha)) \ldots (C(\alpha), X_{ij}(\alpha), X_{i'j'}(\alpha), \ldots)$ on a cell α
5. Count on $\mathcal{E}$, $N(C)$, $N(X_{ij})$, and the co-occurrences $N(CX_{ij})$, $N(CX_{ij}X_{i'j'})$ etc.
6. Calculate $P(C)$, $P(X_{ij})$, $P(C|X_{ij})$, $P(X_{ij})|C)$, $P(X_{ij}X_{i'j'})$, $P(C|X_{ij}X_{i'j'})$ etc.
7. Generate a Bayesian classifier $P(C|\mathbf{X})$

What is my ensemble? 2479 Mexican municipalities

What are my C? The 10% of municipalities with the highest…

1. Loss of forest cover in 2020 in total number of hectares of the municipality — Extensive
2. Percentage loss of forest cover in 2020 relative to the municipal area — Intensive
3. Percentage loss of forest cover in 2020 relative to the municipal forest cover in 2000. — Intensive

What are my X?

- 98 interval variables divided into 6 types: 39 Social, 19 Economic, 18 Governmental, 19 Climatic, 2 Forest cover, 1 Area.
- Extensive and intensive predictors considered: e.g., “Population in extreme poverty” vs. “% of population in extreme poverty”
- All variables are mapped to the municipal level
- Municipalities are rank ordered for each variable and divided into deciles to determine 10 features. In total we have 1058 features.
- Two Mexican Government programs considered “Sembrando Vida” and “Pago por Servicios Ambientales (PSA)”

PSA: This program aims to promote the recognition of the value of the environmental services provided by forest, agroforestry ecosystems and natural resources, in addition to supporting the creation of markets for these services through various programs that have been evolving in the current decade. These programs support communities, ejidos, Regional Forestry Associations, and forest landowners. https://www.gob.mx/conanp/documentos/pago-por-servicios-ambientales-en-areas-naturales-protegidas

What is my P(C|X)?

$P(C|\mathbf{X}) = P(C|X_1, X_2, \ldots, X_{98})$ $\mathbf{X} = (X_1, X_2, \ldots, X_{98})$ 98 dimensional Deforestome space, in reality its 1058 dimensional

As a number this can tell us the where and when of deforestation. It won't tell us the "why".

For the Decision Cycle we need to know the "whys" - which X_i are:

- Important
- Believable (statistically)
- Causal (direct, indirect, very indirect,…)
- Actionable (is it a variable we can change, e.g., climate = no, government program = yes)

Use a Bayesian approach: $P(C|\mathbf{X}) = P(\mathbf{X}|C)P(C)/P(\mathbf{X})$. And then $P(C|\mathbf{X},\mathbf{X}') = P(\mathbf{X}'|C,\mathbf{X})P(C|\mathbf{X})/P(\mathbf{X}'|\mathbf{X})$. And then…

$P(\mathbf{X}|C) \rightarrow \prod_{i=1}^{M} P(\xi_i|C)$ Assume a factorization of the likelihood (weight of evidence)

Importance

$$S(C|X) = \ln\left(\frac{P(C|X)}{P(\bar{C}|X)}\right) = \sum_{i=1}^{M} S_i + S_0. \qquad s_i = \ln(P(\xi_i|C)/P(\xi_i|\bar{C})).\ s_0 = \ln\left(\frac{P(C)}{P(\bar{C})}\right)$$

$$\varepsilon(C|\xi_i) = N(\xi_i)(P(C|\xi_i) - P(C))/\sqrt{N(\xi_i)(P(C)(1 - P(C))}$$

So far, so good using "AI"

Believability

Let's start simple with M = N: Total factorization – Naïve Bayes approximation

$$P(X_i) = \frac{N(X_i)}{N}$$

$$P(X_i|C) = \frac{N(CX_i)}{N(C)}$$

$$P(X_i|\bar{C}) = \frac{(N(X_i) - N(CX_i))}{(N - N(C))}$$

$$P(C) = N(C)/N$$

The “Building Blocks”

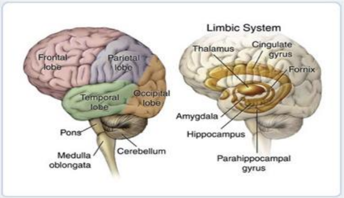
Note that all the top drivers are extensive

Believability Importance Null hypothesis

Table 1: Top/bottom: most predictive drivers in ha

Bin	Description	range	$\varepsilon(C\|X_{ij})$	$s(X_{ij})$	$P(C\|X_{ij})$	$P(C)$
	Most significant niche drivers for deforestation in ha					
p10	Forest area 2000 (ha)	46 167.49 : 137 9018.17	21.15	2.21	0.51	0.10
p10	Sembrando vida program 2020 (mdp)	44 698 : 955 155	16.10	2.55	0.59	0.10
p10	Sembrando vida program 2019-2020 (mdp)	72 869 : 1 863 655	15.65	2.49	0.58	0.10
p10	Sembrando vida program 2019 (mdp)	58 515 : 908 500	14.89	3.70	0.82	0.10
p10	Precipitation of the driest quarter(mm)	92 : 577	14.76	1.71	0.38	0.10
p10	Precipitation of the driest month (mm)	26 : 166	14.55	1.70	0.38	0.10
p10	Population lacking access to basic services in housing	23300.5 : 220919	14.55	1.70	0.38	0.10
p10	Precipitation of the warmest quarter (mm)	510 : 1330	14.34	1.68	0.38	0.10
p10	Precipitation of the coldest quarter (mm)	148 : 976	13.92	1.64	0.37	0.10
p10	Annual precipitation (mm)	1400 : 4887	12.86	1.56	0.35	0.10
p09	Mean temperature of the wettest quarter (°C)	27.37 : 28.35	12.44	1.52	0.34	0.10
p01	Precipitation Seasonality (%)	39.7 : 59.58	12.23	1.50	0.34	0.10
p10	Precipitation of the wettest month(mm)	298 : 885	11.60	1.45	0.32	0.10
p10	Precipitation of the wettest quarter (mm)	788 : 2561	10.97	1.39	0.31	0.10
p02	Mean Diurnal Range (°C)	11.54 : 12.97	10.75	1.37	0.31	0.10
p10	Minimum temperature of the coldest month (°C)	16.1 : 22.2	10.75	1.37	0.31	0.10
p10	Mean temperature of the coldest quarter (°C)	23:27.8	10.12	1.31	0.30	0.10
p09	Forest area 2000 (ha)	22 153.50 : 46 167.49	9.99	1.31	0.29	0.10
p10	Percentage of forest area	84.73 : 99.23	9.95	1.30	0.29	0.10
p10	Mean Diurnal Range (°C)	2.05 : 11.54	9.70	1.28	0.29	0.10
	Most significant anti-niche drivers for deforestation in ha					
p02	Population lacking access to basic services in housing	409 : 866	-5.26	-8.62	0.00	0.10
p02	Population with at least three social deficiencies	422.5 : 973	-5.26	-8.62	0.00	0.10
p01	Municipality area (ha)	221 : 3 400.80	-5.28	-8.62	0.00	0.10
p09	Mean Diurnal Range (°C)	17.68 : 18.33	-5.28	-8.62	0.00	0.10
p10	Mean Diurnal Range (°C)	18.33 : 20.98	-5.28	-8.62	0.00	0.10
p02	Mean Temperature of Driest Quarter	12.65 : 14.53	-5.28	-8.62	0.00	0.10
p01	Annual Precipitation (mm)	44 : 240	-5.28	-8.62	0.00	0.10
p02	Annual Precipitation (mm)	17.68 : 18.33	-5.28	-8.62	0.00	0.10
p01	Precipitation of Wettest Month	9 : 48	-5.28	-8.62	0.00	0.10
p01	Precipitation of Wettest Quarter (mm)	20 : 125	-5.28	-8.62	0.00	0.10
p06	Precipitation of Driest Quarter (mm)	23 : 29	-5.28	-8.62	0.00	0.10
p01	Precipitation of Warmest Quarter (mm)	3 : 110	-5.28	-8.62	0.00	0.10
p01	Total population	81 : 1598.50	-5.28	-8.62	0.00	0.10
p01	Population with at least one social deprivation	75 : 1462	-5.28	-8.62	0.00	0.10
p00	Sembrando vida porgram 2019-2020 (mdp)	0	-7.31	-0.90	0.04	0.10
p00	Sembrando vida 2019-2020/ municipality area	0	-7.31	-0.90	0.04	0.10
p00	Sembrando vida 2019-2020/ Forest area	0	-7.31	-0.90	0.04	0.10
p00	Sembrando vida 2020 (mdp)	0	-7.37	-0.90	0.04	0.10
p00	Sembrando vida 2020 / municipality area	0	-7.37	-0.90	0.04	0.10
p00	Sembrando vida 2020 / forest area	0	-7.37	-0.90	0.04	0.10

What about causality?



What about actionability?

What about causality?

Strength (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.

Bradford Hill 1965

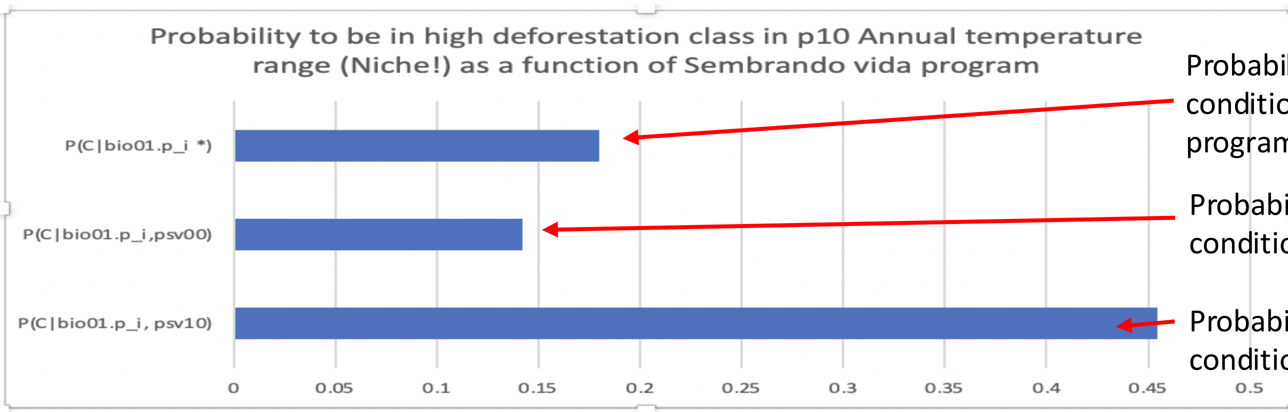
The bigger the score the more likely that it is causal

For two potential causes represented as binary variables we have:
P(C|11), P(C|10), P(C|01), P(C|00)
“AB testing” – look for where cause 1 is present without cause 2, and vice versa, where they´re both present (to look for “epistasis”) and where they´re both absent (to look for other causes)

$$\varepsilon(C|X_{ij}X_{kl}) = \frac{N(X_{ij}X_{kl})(P(C|X_{ij}X_{kl}) - P_{nh})}{\sqrt{N(X_{ij}X_{kl})P_{nh}(1-P_{nh}))}}$$

$P_{nh} = P(C)$, $P(C|X_{ij})$ and $P(C|X_{kl})$

This will tell us whether the associations are statistically significant



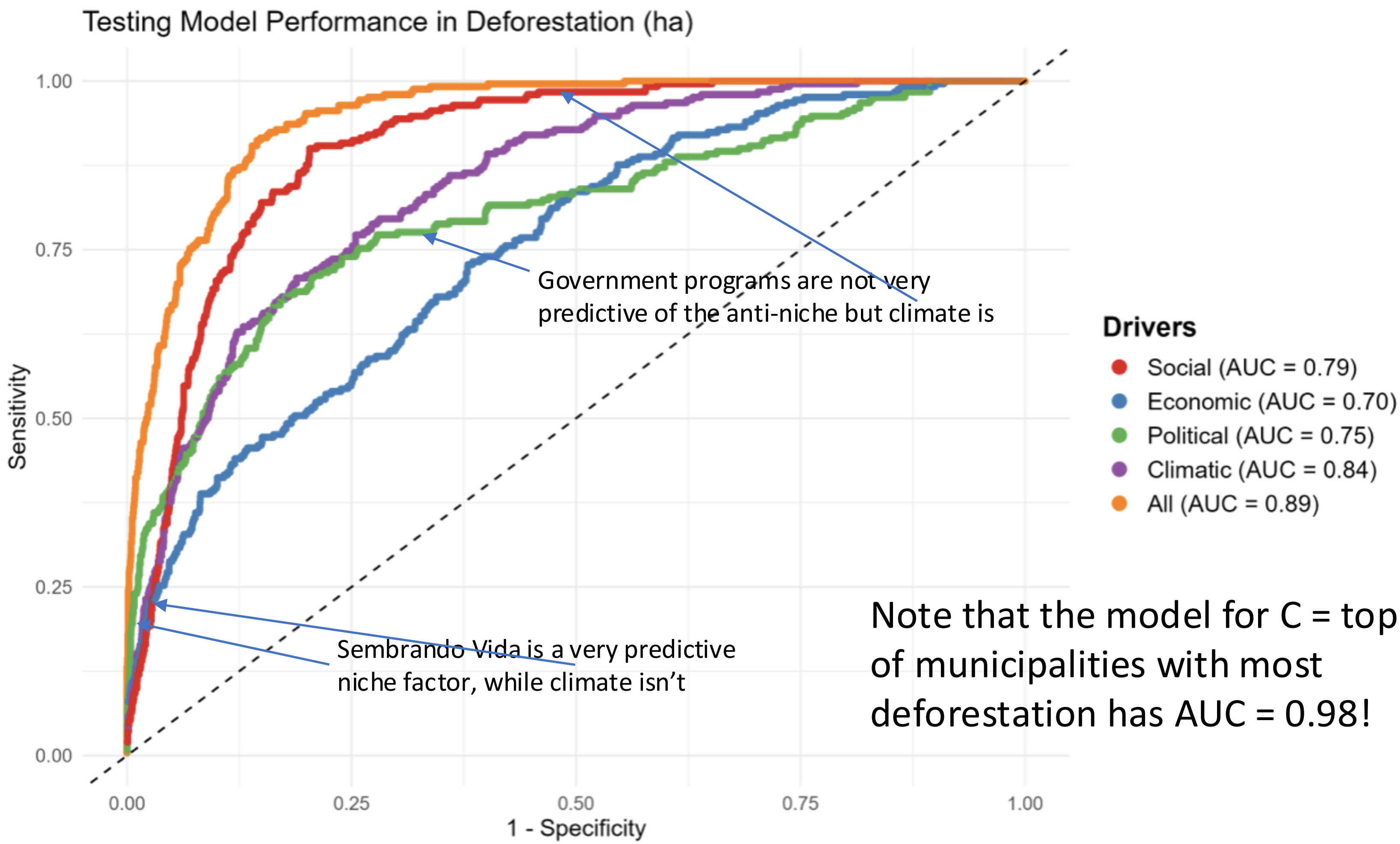
Probability given you’re in climatic niche conditions and you don’t care about the program (marginal probability)

Probability given you’re in climatic niche conditions and NO program support

Probability given you’re in climatic niche conditions and HIGH program support

How good are the models?

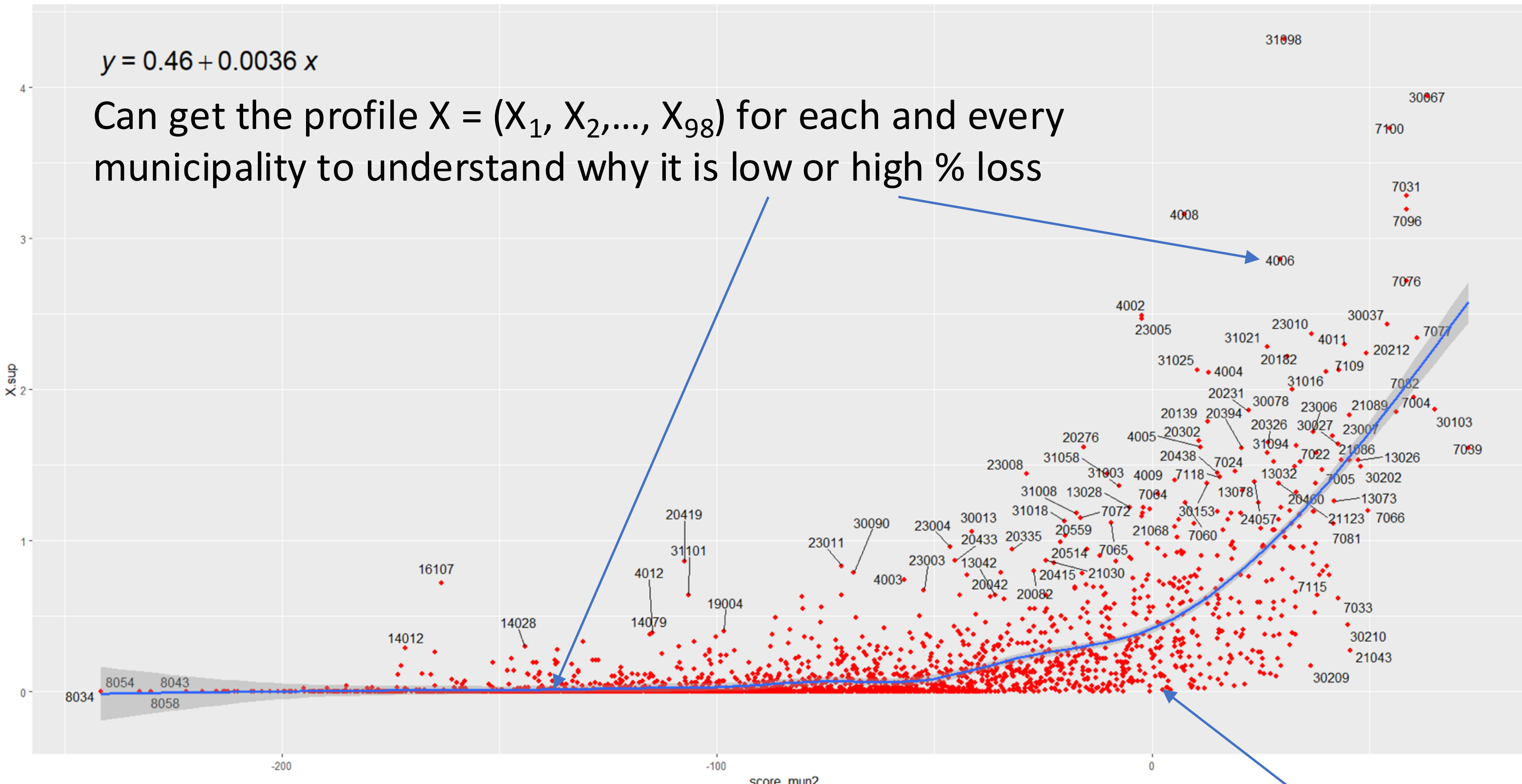
Very good, but you're only as good as the predictability of the data you're working with



Note that the model for C = top 2% of municipalities with most deforestation has AUC = 0.98!

We can use our Bayesian classifier to create a regression model for any continuous variable Y(**X**), such as area deforested, by regressing on S(C|**X**):

Example: Y(**X**) = F(S(C|**X**)); Y(**X**) = Area lost as a percentage of municipal area

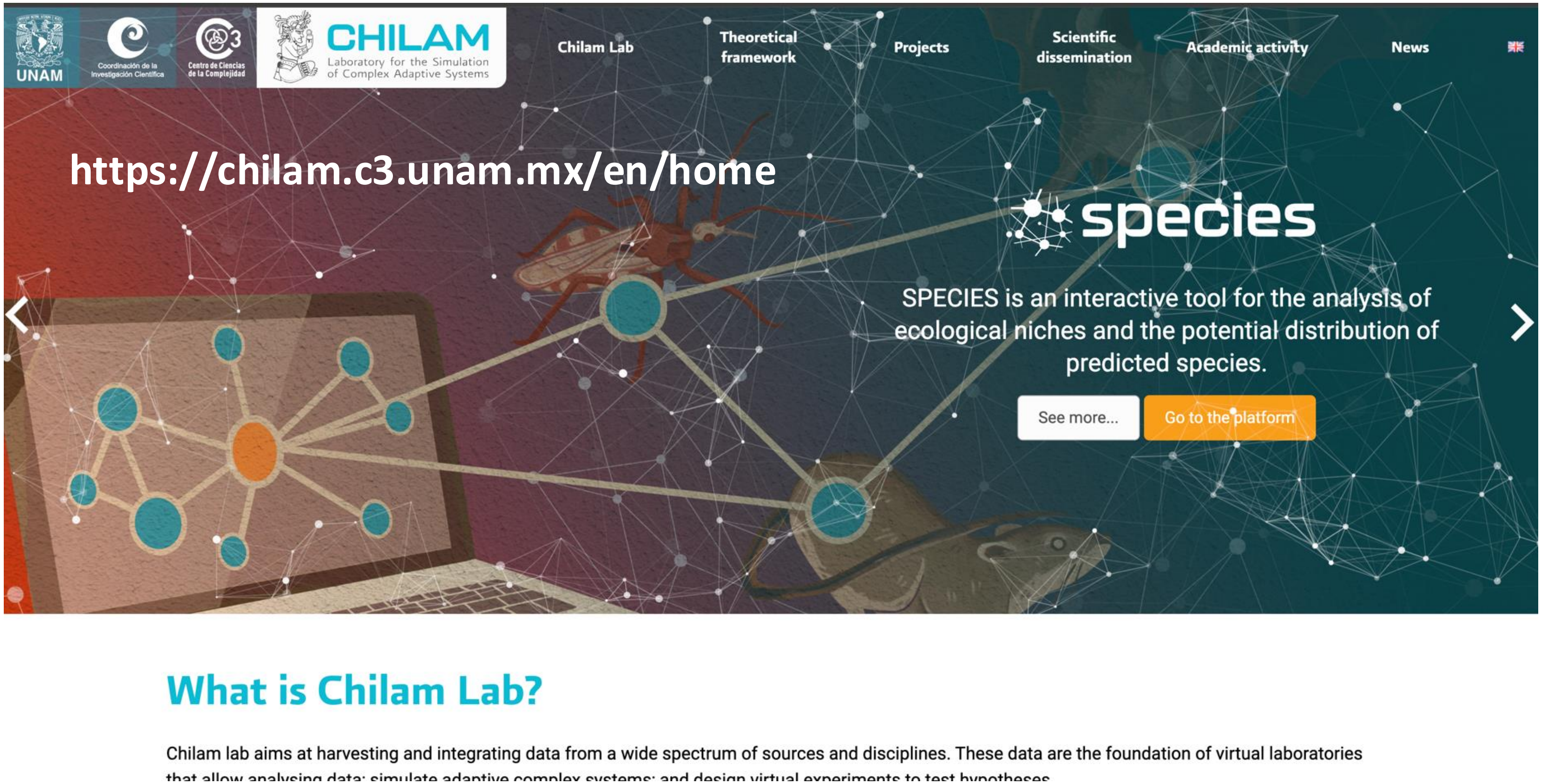


Can get the profile X = (X_1, X_2,…, X_{98}) for each and every municipality to understand why it is low or high % loss

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4585940  0.0106426   43.09   <2e-16 ***
score_mun2  0.0036013  0.0001054   34.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3143 on 2467 degrees of freedom
Multiple R-squared:  0.3214,   Adjusted R-squared:  0.3211
F-statistic:  1168 on 1 and 2467 DF,  p-value: < 2.2e-16
```

Why are these municipalities "high" score but little deforestation?



Chris Stephens: stephens@nucleares.unam.mx

Work in collaboration with: Constantino González Salazar, José Garcia, Romel Calero, Pedro Romero Martínez, José Luis Gordillo, Gabriel Garcia

And the rest of the CHILAM (SPECIES, EPISPECIES, EPIPUMA, PROYECTO42,…) team