On an application of fuzzy logic to building interaction networks between species given geographical data

Víctor Anaya[1], Víctor Breña[2], Daniele Colosi[1], Marisol Flores[1], Luis Miguel García Velázquez[1], Adriana Menchaca Méndez[1], Teresa Patiño Cárdenas[2], Miguel Raggi[1], Sergio Rogelio Tinoco Martínez[1], and César Torres Miranda[1]

[1]Escuela Nacional de Estudios Superiores, Unidad Morelia, UNAM
[2]Centro de Ciencias Matemáticas, UNAM

July 4, 2017

victor_anaya@enesmorelia.unam.mx, victorb@matmor.unam.mx, dcolosi@enesmorelia.unam.mx, maria.flores@gmail.com, luism_garcia@enesmorelia.unam.mx, adriana.menchacamendez@gmail.com, lamoradita@gmail.com. mraggi@gmail.com, stinoco@enesmorelia.unam.mx, catomi@gmail.com

No puedo poner los emails!!!!!!! No sé cómo! Alguien arréglelo antes de que rompa la compu!

Abstract

We present a novel way of constructing an interaction network for ecological species using tools from fuzzy logic. We present a few (3???) examples: Quercus Oaxaca (?), Loba-whatever, and redes centro, which were obtained by doing bla and bla César?. We also release computational tools that take as input some observational data and output a network built using the methods described in this paper. Furthermore, the code to analyze said network is also freely released. Keywords: fuzzy logic, ecology, networks

Mathematics Subject Classification: ???

1   Introduction

In [PJ94] and [TM14] the authors do bla bla bla...

In this paper, we improve on that because of bla bla bla, and apply the results to bla bla and bla.

All the source code used for this paper can be freely downloaded and used (under the terms of the license) from

`https://www.github.com/mraggi/FuzzyLogicEcology`

Traditionally, the way to build an interaction network is to use circles or polygons or whatever, but that has the problems of... César

1.1 A brief introduction to the tools of fuzzy logic

In this section we introduce the reader to the fuzzy logic tools we used to build the interaction network. For a more complete treatment on the topic, see [**?**] or [**?**].

We differentiate between regular sets (*crisp* sets) and *fuzzy* sets. A fuzzy set is a set in which membership is not a boolean value, but rather it is a number between 0 and 1, where a membership of 0 means that the element is not in the set and 1 means the element is definitely in the set. Values in between indicate how much an element belongs to a set.

Formally,

Definition 1.1. A **fuzzy set** $A = (X, \mu_A)$ is a set X together with a function $\mu_A : X \to [0, 1]$. In this case we call X the *ground set* and μ_A the *characteristic function* of A. Sometimes we refer to μ_A as the fuzzy set.

We wish to define the standard set operations $(\subset, \cap, \cup, \setminus)$ on fuzzy sets. There are many ways to do so in a consistent and logical manner, but we will stick to the more standard ones.

Definition 1.2. Let A and B be fuzzy sets on the same ground set X. Then,

- $A \subset B$ iff $\mu_A(x) \leq \mu_B(x)$ for every $x \in X$.

- $A \cap B$ is defined by the characteristic function $\mu_{A \cap B}(x) := \mu_A(x)\mu_B(x)$

- $A \cup B$ is defined by the characteristic function $\mu_{A \cup B}(x) := \mu_A(x) + \mu_B(x) - \mu_A(x)\mu_B(x)$

- $A \setminus B$ is defined by the characteristic function $\mu_{A \setminus B}(x) = \max(0, \mu_A(x) - \mu_B(x))$.

It can be easily checked that these definitions satisfy many of the common properties of sets, such as commutativity for $\cap$ and $\cup$, associativity, De Morgan's laws, etc. when the corresponding crisp set properties are satisfied as well paperizar!.

2 Building the network

The main idea behind the method is that instead of considering the distribution area of a species as a crisp set, we consider the distribution as a fuzzy set. The motivation behind this is that given an observation of an individual X of a certain species, the area of effect of this individual in its interaction with other species will decay the further away one moves from X. Este enunciado está terrible. Re-escribir!

Let $\Omega \subset \mathbb{R}^2$ denote the full area of distribution we are considering.

Figure 2.1: Visualization of the area of distribution of a species

Let $P = (x_0, y_0)$ denote a point with coordinates (x_0, y_0) and $C > 0$ be real constant number. We define $f_P : \Omega \to [0, 1]$ in the following way:

$$f_P(Q) = e^{-C[d(P,Q)]}.$$

where $d(P, Q)$ denotes the (Euclidean) distance from P to Q.

Now, for a certain species $\mathcal{A}$, let A denote the set of observed locations of an individual of that species. It's common to find such data when doing field work of this type and ... César? O quizás en la intro y aquí nada más decir "As we previously mentioned, ...".

Definition 2.1. We define the **area of distribution** of species $\mathcal{A}$ as the fuzzy set given by the following characteristic function:

$$\mu_A = \cup_{a \in A} f_a$$

Following the standard way of constructing interaction networks based on geographical data ([TM14], [?]), we build a weighted directed graph $D = (V, E)$ where V is the set of species, and there is an edge $\mathcal{A} \to \mathcal{B}$ if their respective areas of distribution of $\mathcal{A}$ and $\mathcal{B}$ intersect, and the weight of this edge is defined as

$$w(\mathcal{A} \to \mathcal{B}) = \frac{\text{Area}(\mathcal{A} \cap \mathcal{B})}{\text{Area}(\mathcal{B})}$$

Translating the above definition to the fuzzy world, we may calculate the weight as follows: if μ_A and μ_Y are the respective characteristic functions of $\mathcal{A}$ and $\mathcal{B}$ as defined above,

$$w(\mathcal{A} \to \mathcal{B}) := \left(\int_\Omega \mu_A \mu_B \right) / \left(\int_\Omega \mu_B \right)$$

2.1 Calculating weights numerically

We are left with the following computational problem: Given a set of species and for each species a set of points representing the observational data of the species in question, calculate the weighted digraph as above.

Obviously an analytical approach is a terrible way to do it Bah, justifiquen esto uds. Daniele? Alguien?

So we rely on numerical methods instead. First we'll describe how to find μ_A for each species $\mathcal{A}$, and then we proceed to find $w(\mathcal{A} \to \mathcal{B})$ for each pair of species $\mathcal{A},\mathcal{B}$.

Divide Ω into an $n \times n$ grid. Then, each μ_A can be discretized and represented by an $n \times n$ matrix of real numbers. The higher n is, the better the approximation. For all our experimental results, we found that $n \approx 8000$ is more enough to get an accurate approximation. Alguien se quisiera aventar el análisis de error? No? That's what I thought.

2.2 Finding μ_A

We rewrite μ_A as follows, using the inclusion-exclusion principle:

$$\mu_A = 1 - \prod_{a \in A}(1 - f_a) \tag{2.1}$$

(do we want to expand on this? It took us a while to figure this out from the formulas)

In this way calculating μ_A takes $O(|A|n^2)$ time. To speed up computations by sacrificing (a tiny amount of) accuracy, note that
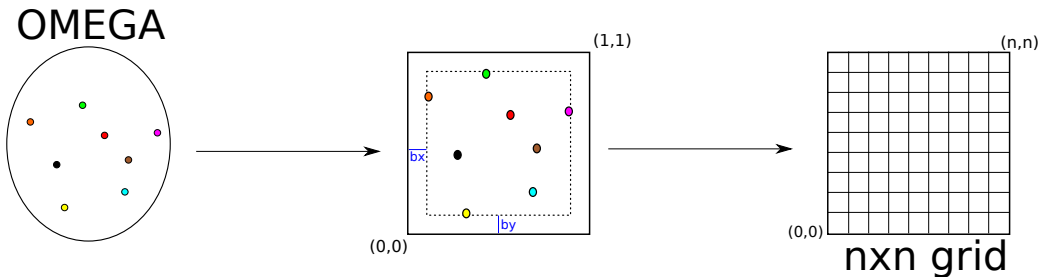
$$\mu_A \leq \sum_{a \in A} f_a$$

Therefore, if we admit an error of $\varepsilon > 0$ for each square in the grid, we only need to compute $f_a(p)$ for points p near a inside a circle of radius R, where

$$R = f_a^{-1}(\varepsilon/|A|) = \sqrt{-\log\left(\frac{\varepsilon}{|A|C}\right)}$$

we note that the exponential decay in the formula for f_a gives a quick way to only affect the area near point a.

With this in mind, we proceed to normalize as follows:

First, let $O = (mx, my)$ and $W = (Mx, My)$, where mx, my, Mx, My denote the minimum and maximum points in the dataset. Let $F = W - O$.

Then, $b_x = R/|F_x|$ and $b_y = R/|F_y|$ are the x and y borders respectively. Any element of the grid outside the border would, by the definition of R, contain a number that is less than ε and therefore can be ignored.

Also, C has to be transformed into C_x and C_y by the equation look it up in the code, because of Daniele said so. I agreed he was correct, but I don't remember.

2.3 Finding w

Once we have μ for every species, it's time to calculate $\int \mu_A \mu_B$ for every pair of species $\mathcal{A}$, $\mathcal{B}$. The straightforward approach has a complexity of $O(s^2 n^2)$, where s is the number of species and n is the size of the grid.

Note the following speedup. Note that

$$\int \mu_A \mu_B \approx \sum_{0 \le x < n} \sum_{0 \le y < n} \mu_A[x,y]\mu_B[x,y]$$

If we flatten the matrix μ_A and μ_B to vectors ν_A and ν_B, this last sum is simply the dot product $\nu_A \cdot \nu_B$.

So consider the $s \times n^2$ matrix

$$M = \begin{bmatrix} \longleftarrow & \nu_0 & \longrightarrow \\ \longleftarrow & \nu_1 & \longrightarrow \\ \dots & \dots & \dots \\ \longleftarrow & \nu_{s-1} & \longrightarrow \end{bmatrix}$$

Then, MM^T is an $s \times s$ matrix whose (i,j)-th entry contains $\int \mu_A \mu_B$. This can be done efficiently with fast matrix multiplication algorithms ([**?**]) (Marisol, tú te los sabes seguramente, pon unas palabritas de ellos y di que son super chidos y que no se que).

Note also that the algorithms described here are highly parallelizable, and indeed the code uses as many cores as available.

3 Experiments and results

3.1 Digraphs

The following digraphs were obtained as a result of the algorithms described in the previous section. (mm... damos las gráficas? La más chica es una matriz de 56×56, aunque podríamos dar sólo las aristas más pesadas... no sé)

3.2 Analysis of the digraphs

This graphs are cool because here is the robustness and the analysis revealed deep and important truths (Yo les doy los resultados del análisis. Uds. escriben por qué, para qué, cómo, cuándo, etc.)

4 Conclusions

Finally, we conclude this awesome work. Please accept this paper!!

Acknowledgements

References

[PJ94] Fulano Perez and Sutano Juarez, *Fuzzy logic and bla bla bla*, Ecology today! **10** (1994).

[TM14] César Andrés Torres Miranda, *Importancia de los ensambles de especies y de los procesos geológico-climáticos en la comprensión de los mecanismos de evolución de la sección lobatae del género quercus en américa del norte*, Ph.D. thesis, Instituto de Investigaciones en Ecología, Universidad Nacional Autónoma de México, 2014.

A On the mathematical expression of the area distribution

We consider a species A and denote by N the number of its individuals; then the expression (2.1) reads:

$$\mu_A = \sum_{i=1}^{N} f_i - \sum_{i \neq j} f_i f_j + \sum_{i \neq j \neq k, i \neq k} f_i f_j f_k + \cdots + (-1)^{N-1} \prod_{i=1}^{N} f_i, \tag{A.1}$$

or, in more compact notation,

$$\mu_A = \sum_{n=1}^{N} (-1)^{n+1} \sum_{1=i_1 < \cdots < i_n}^{N} f_{i_1} \cdots f_{i_n}. \tag{A.2}$$

Now we compute the integral of μ_A over all of space, i.e. over $\mathbb{R}^2$:

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, \mu_A(x, y)$$

$$= \sum_{i=1}^{N} \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_i(x, y) - \sum_{i \neq j} \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_i(x, y) f_j(x, y)$$

$$+ \sum_{i \neq j \neq k, i \neq k} \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_i(x, y) f_j(x, y) f_k(x, y) + \cdots + (-1)^{N-1} \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \prod_{i=1}^{N} f_i(x, y),$$

$$= \sum_{n=1}^{N} (-1)^{n+1} \sum_{1=i_1 < \cdots < i_n}^{N} \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_{i_1}(x, y) \cdots f_{i_n}(x, y). \tag{A.3}$$

Notice that if the f_i are Gaussian distributions, then the first integral in the r.h.s. of the second line of the above expression results to be

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_i(x, y) = \frac{\pi}{C}. \tag{A.4}$$

We consider the integral of the product of n Gaussians:

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_{i_1}(x, y) \cdots f_{i_n}(x, y) = \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, \exp\left[-C \sum_{\alpha=1}^{n} \left((x - x_{i_\alpha})^2 + (y - y_{i_\alpha})^2\right)\right]. \tag{A.5}$$

Notice that

$$\sum_{\alpha=1}^{n} (x - x_{i_\alpha})^2 = nx^2 - 2x \underbrace{\sum_{\alpha=1}^{n} x_{i_\alpha}}_{X_{1,n}} + \underbrace{\sum_{\alpha=1}^{n} x_{i_\alpha}^2}_{X_{2,n}}. \tag{A.6}$$

So, using the obvious notation,

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, f_{i_1}(x, y) \cdots f_{i_n}(x, y)$$

$$= \int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, \exp\left[-Cn(x^2 + y^2) + 2C(X_{1,n}x + Y_{1,n}y) - C(X_{2,n} + Y_{2,n})\right],$$

$$= \frac{\pi}{nC} \exp\left[\frac{C(X_{1,n}^2 + Y_{1,n}^2)}{n} - C(X_{2,n} + Y_{2,n})\right],$$

$$= \frac{\pi}{nC} \exp\left[\frac{C\left((\sum_{\alpha=1}^{n} x_{i_\alpha})^2 + (\sum_{\alpha=1}^{n} y_{i_\alpha})^2\right)}{n} - C \sum_{\alpha=1}^{n} (x_{i_\alpha}^2 + y_{i_\alpha}^2)\right],$$

$$= \frac{\pi}{nC} \exp\left[-\frac{C}{n} \sum_{1=\alpha<\beta}^{n} \left((x_{i_\alpha} - x_{i_\beta})^2 + (y_{i_\alpha} - y_{i_\beta})^2\right)\right],$$

$$= \frac{\pi}{nC} \exp\left[-\frac{C}{n} \sum_{1=\alpha<\beta}^{n} d_{i_\alpha i_\beta}^2\right], \tag{A.7}$$

where $d_{i_\alpha i_\beta}$ is the (euclidean) distance between point $P_{i_\alpha} = (x_{i_\alpha}, y_{i_\alpha})$ and point $P_{i_\beta} = (x_{i_\beta}, y_{i_\beta})$. Finally

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, \mu_A = \sum_{n=1}^{N} (-1)^{n+1} \sum_{1=i_1<\cdots<i_n}^{N} \frac{\pi}{nC} \exp\left[-\frac{C}{n} \sum_{1=\alpha<\beta}^{n} d_{i_\alpha i_\beta}^2\right]. \tag{A.8}$$

Analogously we obtain the integral of the product of two membership functions

$$\int_{\mathbb{R}^2} \mathrm{d}x \, \mathrm{d}y \, \mu_A \mu_B = \sum_{n=1}^{N_A} \sum_{\tilde{n}=1}^{N_B} (-1)^{n+\tilde{n}} \sum_{1=i_1<\cdots<i_n} \sum_{1=\tilde{i}_1<\cdots<\tilde{i}_n} \frac{\pi}{(n+\tilde{n})C}$$

$$\exp\left[-\frac{C}{n+\tilde{n}} \left(\sum_{1=\alpha<\beta}^{n} d_{i_\alpha i_\beta}^2 + \sum_{1=\alpha<\beta}^{\tilde{n}} d_{\tilde{i}_\alpha \tilde{i}_\beta}^2 + \sum_{\alpha=1}^{n} \sum_{\tilde{\alpha}=1}^{\tilde{n}} d_{i_\alpha \tilde{i}_{\tilde{\alpha}}}^2\right)\right]. \tag{A.9}$$