UNIVERSIDAD DE QUINTANA ROO

División de Ciencias Políticas y Humanidades

English-Spanish translation of the article "A statistical explanation of MaxEnt for ecologists"

TRABAJO MONOGRÁFICO

En la modalidad de Traducción

Para obtener el grado de:

LICENCIADO EN LENGUA INGLESA

Presenta

Mariana Ruelas García

Asesores

Dra. Griselda Murrieta Loyo

Dra. María del Rosario Reyes Cruz

Dr. Moisés Damián Perales Escudero

Chetumal, Quintana Roo, México, Mayo 2018

UNIVERSIDAD DE QUINTANA ROO

División de Ciencias Políticas y Humanidades

English-Spanish translation of the article "A statistical explanation of MaxEnt for ecologists"

Trabajo monográfico elaborado bajo la supervisión del Comité del Programa de Licenciatura y aprobada como requisito para obtener el grado de

LICENCIADO EN LENGUA INGLESA

COMITÉ DE TRABAJO MONOGRÁFICO

Directora : _____
Dra. Griselda Murrieta Loyo

UNIVERSIDAD DE QUINTANA ROO
DIVISIÓN DE CIENCIAS POLÍTICAS Y HUMANIDADES

Asesor: _____
Dra. María del Rosario Reyes Cruz

Asesor: _____
Dr. Moisés Damián Perales Escudero

Chetumal, Quintana Roo, México, Mayo 2018

UNIVERSIDAD DE QUINTANA ROO
CONTROL ESCOLAR
TITULACIONES

ACKNOWLEDGEMENTS

I would like to thank the Universe for being so generous with me by bringing wonderful people to my life who have shared their knowledge and experiences and for letting me learn from all of them. I appreciate the incredible opportunity that was given to me of studying my second major in the English Language, which has opened me so many professional doors.

First of all, I would like to thank my family for being so supportive with my studies and for always being there for me even if we do not always share the same point of view. Thank you father for I have never needed anything, you are always putting my mother and I first, you are always giving us all you have, you are one of a kind and I hope to be always as generous and responsible as you are. Thank you mother for I have never felt alone in my entire life; you are always there for me, no matter what. You have always supported me unconditionally. You are the most caring and loving human being and I am truly thankful for being your daughter.

I want to thank my professor Griselda Murrieta for being so patient with me through the whole process of this paper and mostly for accepting being part of this journey, thank you for giving me your time and for sharing your knowledge and expertise with me, for correcting me when needed and for believing in me and my abilities; You are an inspirational professor for me and I hope one day I could be as good at teaching as you are, professor. It has been an honor for me to be your pupil all this time.

I truly appreciate the time professor María del Rosario Reyes and professor Moisés Perales have invested in reviewing and correcting my paperwork. To all my professors I happened to be lucky to have during the major. Thank you for giving me the necessary tools to become a well prepared professional.

Lastly but not least, thank you for all the people who have always been there for me, encouraging me and for being part of my life.

CONTENT

CHAPTER I. INTRODUCTORY MATTERS ---------- 7
1.1 Introduction ---------- 7
1.2 Rationale ---------- 11
1.3 Objectives ---------- 12
1.4 Literature review ---------- 13
1.4.1 A brief history of the translation discipline ---------- 14
1.4.2 The main translation approaches ---------- 16
1.5 Method ---------- 20
CHAPTER II TRANSLATION ---------- 24
CHAPTER III TRANSLATION ANALYSIS ---------- 59
3.1 Borrowing ---------- 59
3.2 Calque ---------- 61
3.3 Literal translation ---------- 63
3.4 Transposition ---------- 65
3.5 Modulation ---------- 67
3.6 Equivalence & Adaptation ---------- 68
3.7 Explicitation ---------- 69
3.8 Amplification ---------- 70
3.9 Reduction ---------- 71
CHAPTER IV CONCLUSIONS ---------- 73
REFERENCES ---------- 76
APPENDIX ---------- 79

RESUMEN

La traducción de artículos de revistas especializadas es una herramienta imprescindible para expandir el conocimiento a través del mundo. Esta monografía nos brinda la traducción del artículo "A statistical explanation of MaxEnt for ecologists" cuya versión original fue escrita en el idioma inglés y se tradujo al idioma español en este trabajo. La traducción de este documento formará parte de la bibliografía académica del curso "Ecología de Poblaciones" impartido en el Colegio de la Frontera Sur. Este es un artículo que hace referencia al uso y características del software MaxEnt para ser entendido y utilizado por ecólogos. Esta traducción será de utilidad para profesores y estudiantes del Colegio de la Frontera Sur. Además de traducir el texto al español, el objetivo también fue el de revisar y presentar las diferentes etapas y la evolución que ha tenido la traducción a través del tiempo. De igual manera se describen las principales escuelas y técnicas de traducción de las cuales, a modo de análisis, se hizo énfasis solo en aquellas que fueron aplicadas al momento de traducir el artículo. Se añadieron ejemplos de aplicación de dichas estrategias por diversos autores y ejemplos extraídos de la traducción que se realizó. Se presentan también en esta monografía las conclusiones a las que llegó la autora de este trabajo después de haber realizado la traducción, así como sugerencias para que la traducción final sea la más apropiada. Al final de este documento, se anexa el artículo original que fue traducido.

Palabras clave: MaxEnt, traducción, inglés, español, ecólogos, técnicas, artículo, estadística

ABSTRACT

The translation of articles from specialized journals is a paramount tool to expand knowledge throughout the world. This monograph contains the translation of the article "A statistical explanation of MaxEnt for ecologists" whose original version was written in English and translated to Spanish in this paperwork. The translation of this document will be part of the academic bibliography of the "Ecología de Poblaciones" course given in El Colegio de la

Frontera Sur; this is an article whose content is about the use of the MaxEnt software and its characteristics and it is written in a way to be understood and used by ecologists. This translation will be useful for professors and students of El Colegio de la Frontera Sur. Besides translating the text into Spanish, the objective was also to revise and offer the different stages of translation history and its evolution through time; likewise, the main schools and techniques in translation are described and only the techniques which were applied to translate the chosen article were emphasized to be analyzed. Examples from different authors and taken from this translation were added. The conclusions reached by the author at the end of this translation process are also included in this paperwork as well as suggestions from the author for the final product to be the most suitable one. At the end of this paperwork the original article of MaxEnt written in English is included.
Keywords: MaxEnt, translation, English, Spanish, ecologists, techniques, article, statistics

CHAPTER I
INTRODUCTORY MATTERS

1.1 Introduction

Through history, civilizations have always faced an enormous challenge: to communicate. First, people used signs and symbols to communicate, then, as the time went by oral language became the media to establish communication. Words then became the instrument to represent places, people and objects and to carry on conversations. Written language soon followed oral language. Language turned into a valuable part of human's evolution. Then population expanded and people's language changed into different languages. As a result of the variety of existing languages, translation appeared as a resource to communicate among peoples that did not share the same language.

Translation is an art, it is a human's ticket to get to know different cultures, different manners of thinking, it is our way of interpreting the other language. Translation has been extremely important through the history and its importance will not diminish through time; on the contrary, the relevance of translation is increasing, especially in the science fields where translation is such an important resource for enhancing the knowledge (Montgomery, 2009).

There are many definitions for translation. Understanding this concept is essential as it takes an important role in our lives; therefore, a couple of definitions are presented in the following pages. Levý is a translation theoretician who defines translation as follows:

> Translation is communication. More precisely, translators decode the message contained in the text of the original author and reformulate (encode) it into their own language. The message contained in the translated text is then decoded by the reader of the translation. A binomial chain of communication is established (Levý, 2011, p. 23).

Levy's definition of translation seems to take into account message as the main element of this activity devoid of the contextual factor. This definition locates translation in a communicative setting where it is seen as a chain of messages to be communicated. No reference is made to the strategies used to carry out a translation in this definition but it is still a good explanation on what a translation is about.

Schulte, the founder of the center for Translation Studies at The University of Texas and co-founder of the American Literary Translators Association, presents another idea of what translation is:

> In a deeper philosophical sense, translation deals with the challenge of carrying complex moments across language and cultural borders, and, therefore, translators always navigate in realms of uncertainty… Furthermore, as soon as a word enters into contact with another word, certain new associations of meaning are created that transcend the original definition of a word. Therefore, each translation is the making of yet another meaning that comes to take shape through the interpretive approach and insight of the translator. The premise of all translations remains the same. Each translation is the variation of yet another translation, which excludes the notion of ever arriving at the only definitive translation (Schulte, 2012, paragraph 5).

Schulte, on the other hand, emphasizes the contextual element. He remarks the fact that words meaning may change according to the context in which they are. He goes deep into meaning and its relation with the translator.

Other authors think more freely about translation. For García Yebra, for example, the main point of translation is the idea being and not being said in the L1. In the prologue of "Metafísica de Aristóteles" this author mentions the golden rule for translating according to his ideology and experiences and it goes as following:

> La regla de oro para toda traducción es, a mi juicio, decir todo lo que dice el original, no decir nada que el original no diga, y decirlo todo con la corrección y naturalidad que permita la lengua a la que se traduce [The golden rule for every translation is, to my mind, to say everything the original text says, not to say

> anything the original text does not say, and to say it with the correctness and ease that the target language allows to] (García Yebra, 1998, p. XXVII).

His definition is simple yet concise; for Garcia Yebra the main objective in a translation is to say all the ideas exposed in the original text in the most accurate and natural way as possible, always taking into consideration what the target language allows us to do within its vocabulary and grammatical structures.

After reading these definitions, I can state that translation is a communicative instrument to allow interaction to peoples with different languages; it is a powerful resource to enhance our knowledge when information is only available in one language. Translating is a language process to communicate ideas as accurate as possible from one language to another one. The translation depends mainly on the translator´s interpretation, ability and cultural knowledge.

In my opinion, translations give us the opportunity to know other points of view and other perspectives besides the ones we are familiar with or besides the ones we thought were the only ones. Translation is very important when the most of the advanced knowledge of a topic is usually produced in few languages. Many of the most recent articles on scientific disciplines are mainly written in English (Foyewa, 2015). Thus my intention is to contribute to having up to date texts translated into Spanish. Then I decided to do a monograph on translation as the final paper to graduate from the bachelor's degree program in English.

Translating a text is a process that requires the implementation of certain techniques to reach first a neat interpretation of a discourse and then a L2 text that conveys the L1 text meaning. Translating requires the development of high skills and a wide cultural knowledge base as well as knowledge of several topics of different fields. I thus want to challenge myself with the purpose of interpreting a text written in another language (English) and translating it into my first language (Spanish) without silencing the essence and inner voice of the original text. Translating this text will give me the opportunity to use all the tools and knowledge learnt through all my years in the English major. By translating I put in practice my skills in grammar, semantics, syntax and all the skills obtained from my classes.

As translation is a complex linguistic and communicative process, several different translation techniques and strategies have been developed by different authors, depending on the approach the translators want to work with and depending on what kind of translation is to be

done. Three main schools have developed approaches to address the translation of texts: the Russian, the Canadian and the American schools.

The Russian approach has Shweitser and Reitsker as its main collaborators; this approach describes three types of relationships between the source language and the target language: the correspondence between languages, the context and other transformations needed. The most popular approach is the Canadian one, proposed by Vinay and Darbelnet in 1958; these two authors based their theory on the structure of the language and "the spirit" of the message to be transmitted. Malone (1988) is the main author of the American model in which the techniques serve as tools for the act of translating but also as a resource for the study thereof.

After revising different techniques for the translation process, the translation techniques from the Canadian Approach (proposed by Vinay and Darbelnet in 1958) are the ones in which this translation relies; its techniques are the best-suited ones for the type of text to be translated in this paper. This approach offers a variety of techniques, which are very useful for the translation of the article; the techniques Vinay and Darbelnet offer are: borrowing, calque, literal translation, transposition, modulation, equivalence and adaptation (Fawcett, 2003). All of them will be explained in following chapters of this paper.

There are different kinds of texts: persuasive texts, descriptive texts, informative texts, instructive (British Broadcasting Corporation, 2011). he article to be translated in the following work is a specialized research paper in the Geography and Ecology Field titled "A statistical explanation of MaxEnt for ecologists" and it falls into the category of informative texts. I decided to translate this text because of its implications, especially in the ecology and biogeography field. Moreover, there is a department of engineering and a PhD in Geography at the University of Quintana Roo and at El Colegio de la Frontera Sur that might find very useful this translation due to the fact there is paucity in the existing literature in these fields in the Spanish language.

The translator´s objective is to allow more readers to get the knowledge given in a specific text by translating from a source language to a target one. In this sense, the translation to be done will be useful for the Spanish-speaking audience in the Geography and Ecology field, especially for the students taking the Master program: *Recursos Naturales y Desarrollo rural* in El Colegio de la Frontera Sur; these students sometimes do not have the English level needed to understand and comprehend the chosen article. The resulting translation would be a great

contribution for further studies and would be a useful material for some classes in El Colegio de la Frontera Sur institution.

The article to be translated was published in the Diversity and Distributions journal that specializes in biogeographical principles, theories, and analyses. This monthly journal focuses on the biogeography and ecology of diversity and distributions. This article contains specific terminology and it deals with concerns related to the biogeography and ecology field. Since it is a specialized article, it would be of relevance and interesting for a specific audience mentioned in the paragraph above.

Translating an article of this nature is challenging mainly because of the specialized terms used; besides, it requires investing time to learn about the subject matter of the paper so as to have a more natural and understanding outcome. Furthermore, by translating this article, it gives myself the opportunity to use different translation techniques as well as to put in practice my writing skills; it also gives me the opportunity to make use of my English language knowledge regarding grammar, semantics, vocabulary, just to mention a few ones. This has been a complex task for me.

This paper is divided in four main chapters. In the first one the theoretical framework is found. In here, a brief review of the main stages and approaches of the translation discipline are exposed. I review especially the Canadian Approach to find the strategies and techniques needed to succeed in the translation of the chosen article. In the second chapter, the Spanish translation of the source language article is presented. Chapter three analyses the techniques used to translate the chosen text; a brief description of these techniques is given as well as examples taken from the translation and the analysis of such examples. In the fourth chapter of this paper I give a conclusion about the process of translating into Spanish the chosen article originally written in English; moreover, personal recommendations for novice translators are presented too.

1.2 Rationale

Translation has become a crucial part of science; the necessity of sharing and spreading the information on a global scale has made translation an important tool in order to fulfill that

necessity. Science is in a state of continuous development and to partake in the progress of science, we need to be constantly updated.

There are many scientific and academic fields in which there is a prevalence of foreign literature, especially literature from the United States and some other countries where the English language is the mother tongue. As a result, many of the books and articles are written in English. Even though learning English is part of the curricula of most of the majors, scientific and specialized articles, such as the one translated in this paper, are hard to comprehend for most of the students who study English. This may be because there are many technical concepts and complex structures that, though common in academic papers, are not familiar for students.

At the Colegio de la Frontera Sur, for example, in the Masters' program: *Recursos Naturales y Desarrollo rural*, students are required to read specialized texts in English. Many of those students find it difficult to understand this kind of texts, not only because of the language being used but also because of the terms used. Thus, I think that translating the article would be really helpful not only for these master's students but also for ecologists in general.

For my part, the strongest reason to carry out this translation is to apply knowledge and techniques acquired through all the years of the English Language Major I studied. Translating specialized articles like the one in this paper is an enormous challenge since it has many technical terms and specialized language. In order to succeed I had to apply several translation techniques and I had to learn a little bit about geology, ecology and geography.

1.3 Objectives

This monograph attempted to accomplish the following objectives:

- To provide a faithful and understandable Spanish translation of the source English-written article "A statistical explanation of MaxEnt for ecologists" in order to be used in the course "Ecología de Poblaciones" and in the masters' programs related to Biology from El Colegio de la Frontera Sur.
- To provide an analysis of the techniques used during the translation process as well as to provide examples of those applied techniques

- To provide recommendations for novice translators to translate specialized articles

1.4 Literature review

In this chapter, we first present a brief review of the history of the translation discipline. Then the most important translation approaches, techniques and strategies are explained. Examples of the use of the translation techniques are given. The purpose of this theoretical background is to explain how these techniques have been applied to attempt to convey the original meaning of the source text by presenting some examples.

Different approaches, translation techniques and strategies have been designed to try to explain the objective and the process that should be followed to translate a text from a source into a target language. Those approaches have changed over the time. Montgomery (2000), for example, has commented on the ever changing nature of the discipline and on the ever changing nature of knowledge which is in a constant mobility. He explains that knowledge can be seen as a mobile of culture and changes are brought forth by the discoveries that are made and by the cultural requirements of each target culture. Changes in translation approaches and techniques are also due to the changes in the development of the disciplines.

Based on the disciplines, three main types of translation can be distinguished: Literary, Scientific and Technical translation (Kohoutkova, 2016). Here I will not talk about the literary type since it is not related to the text chosen to be translated in the present monograph. Scientific and technical translation has always played a crucial role in history and in the advancement of human civilization. Pinchuck (1977) has discussed the main differences between these two procedures: on one hand, technical text is designed to convey information in a precise and adequate form; on the other hand, a scientific text will discuss and scrutinize, and also arrange information so the ideas will be explained. Scientific texts pretend to explain and propose theories. Due to these differing aims, the language used in each type of text, and consequently the strategies needed to translate them, may vary significantly.

Gómez and Gómez (2011) have also commented the dichotomy between scientific and technical translation. For them, the distinctive features between the two are that scientific discourse concern itself with theoretical aspects and it is often associated with the study and

description of natural phenomena, whereas technical texts are centered around the practical applications of scientific knowledge. The text chosen to be translated is the result of the practical application of scientific knowledge.

1.4.1 A brief history of the translation discipline

Translating has been a practice for ages; during the Roman period, romans translated from Greek as a way to expand and enrich their cultural knowledge and most of the translations were carried out orally in the beginning (Gentzler, 2014). Nowadays, translation continues to be used as an intercultural practice orally but also written.

Throughout the history, different perspectives regarding this discipline emerged in different places and in different times. The very pioneers of the translation discipline were commentators, such as Cicero (first century, B.C) and St. Jerome (fourth century, B.C). Jerome is best known for his translation of the Bible into Latin. Cicero and Jerome debated on the usage of word-for-word and sense-for-sense translation. The translators at this stage made suggestions on how to translate regarding their own works and techniques (Gentzler, 2014).

Then, in the seventeenth century, John Dryden came into the scene negating metaphrase, word-for-word translation for lacking fluency and arguing in favor of paraphrase, which was more focused on meanings (Ghanooni, 2012). Later on, there were also people who put the emphasis on meaning rather than on phrases. In the eighteenth century, as El-Dali mentioned:

> "The translator was compared to an artist with a moral duty both to the work of the original author and to the receiver". In this context, the translator had to pay attention to the meaning, which means to understand the message and not only translate literally" (2011, p. 30).

In the 1900s, translation was viewed as "an interpretation which necessarily reconstitutes and transforms the foreign text. For scholars as Schleiermacher and Bolt, translation is a creative force in which specific translation strategies serve a variety of cultural and social functions" (quoted in Ghanooni, 2012, p. 78); equivalence was also a recurrent topic in translation. Based on research found in Nida´s book (1964), in the 1930s, courses for specialized translations firstly

appeared in Germany and Russia and teachers of these courses found out difficulties in teaching how to translate scientific texts by using the Literary Translation Theory.

From the 1940s to 1950s the main topic in translation studies was Translatability which is, according to Venuti (2000, p.16), "an essential quality of certain works and …by virtue of its translatability the original is closely connected with the translation…we may call this connection a natural one, or, more specifically, a vital connection." Nida (1945) theorized about the translation problems and when he was working on the translation of Bible he realized most of the solutions to them should take into consideration the acquisition of sufficient "cultural information," to avoid misinterpretations.

In 1958, the concept of translatability is drawn on by Jean-Paul Vinay and Jean Darbelnet (quoted by Ghanooni, 2012). These two authors talked about the "Equivalence of message" referring to the reality and context of the two languages involved in the translation process. While working on the translation process, Vinay and Darbelnet determined six translation techniques very useful even nowadays when translating texts.

Through the 1960s to 1970s the concept "equivalence" was very popular among translation theorists. Nida expressed there are two different types of equivalence: "Formal equivalence focuses attention on the message itself, in both form and content; dynamic equivalence aims at complete naturalness of expression, and tries to relate the receptor to modes of behavior relevant within the context of his own culture." (1964. p. 159)

Another author who tried to explain the concept of translation was George Steiner. In 1975, "unlike linguistic-oriented theories that considered translation as functional communicative, he went back to German Romanticism and the hermeneutic tradition" (Ghanooni, 2012, p. 80) to attempt a model based on meaning to explain translation.

In the 1980s, translation was viewed as an independent form of writing, with differences from the original text to be translated (Venuti, 2004). As this discipline moved towards the present, the level of sophistication and inventiveness did in fact soared and new concepts, methods, and research projects were developed which interacted with this discipline. Gender research, process-oriented and culturally oriented research, skopos and computerized corpora are just some examples of the new researches and concepts (Ghanooni, 2012).

Nowadays, with the invention of the Internet and digital tools, the translation discipline is finding new methods and tools to be more automatic and increased the cultural exchanges. As El-

Dali states, these innovations lead translators to look for technological ways to use more practical techniques that enable them to translate better and in less time (2011).

1.4.2 The main translation approaches

As I mentioned before, there have been several approaches that have tried to explain the discipline of translation. Many scholars have designed techniques and approaches they have found useful to allow a more accurate translation. Some of the most important theories will be briefly explained in the following paragraphs.

The Russian Approach has Yakob Retsker as its major exponent; he describes three types of relationship between a source language and the target one (Fawcett, 2003, pp. 27-33):

1. Equivalence: a one-to-one relationship between the languages involved in the translation process, regardless of context or whatsoever.
2. Variant and contextual correspondence, also known as analogy (renamed by Shveitser): one-to-many correspondences between the languages involved, regarding context. In this case, the context is crucial to apply the correct words and phrases.
3. All other types of translational information, also known as adequacy (renamed by Shveitser): no one-to-one equivalence. In such case, the translator might have to use concretization, logical derivation, antonymic translation and/or compensation as translation techniques in order to succeed.

This taxonomy is part of the attempts to formalize and organize the techniques and procedures involved in the translation discipline. Another famous approach to translation is Canadian one developed mainly by Vinay and Darbelnet (1958). These two theorists advocate seven different techniques which can be applied at the linguistic levels of lexis, grammar and text. These translation techniques are: borrowing, calque, literal translation, transposition, modulation, equivalence and adaptation (Fawcett, 2003). According to Vega Cernuda (1994), Vinay and Darbelnet's approach is the one of the most complete models and the greatest contribution that the traductology discipline has given to translators. In the following paragraphs the dynamic translation model by Vinay and Darbelnet will be presented and examples of its techniques are given.

In their book called "Comparative Stylistics of French and English, a methodology for translation", Jean-Paul Vinay and Jean Darbelnet (1995) explain basic concepts such as linguistic sign, meaning and sense, langue and parole, language and stylistics, and many more but also they mention a methodology for translation which includes strategies the translators use and they mention the following:

> In the process of translating, translators establish relationships between specific manifestations of two linguistic systems, one which has already been expressed and is therefore given, and the other which is still potential and adaptable. Translators are thus faced with a fixed starting point, and as they read the message, they form in their minds an impression of the target they want to reach (Vinay & Darbelnet, 1995, p. 30).

These scholars recognize two methods of translation in their theory: direct or literal translation and oblique translation; In Direct translation it is possible to translate directly (element by element) the message from the source language to the target one. In Oblique translation, due to structural or metalinguistic differences between the languages, complex methods have to be used in order to convey the message properly (Vinay & Darbelnet, 1995). The following are the techniques used in direct translation:

Borrowing: the source-language form is taken into the target language because there is a metalinguistic lacuna (e.g. a word that does not have a (clear) equivalence in the target language). For instance, Mexican food names like *tamales, tacos, tortillas* and so on (ibid) have been borrowed into English. Also, borrowing can be used to create a stylistic effect; for example, the term *perestroika* describes a type of political movement and this word is not translated to keep its exotic flavor (Fawcett, 2003).

Calque: in the words of Vinay and Darbelnet (1955 p. 32), "a calque is a special kind of borrowing whereby a language borrows an expression form of another, but then translates literally each of its elements." An example these authors presented in their book is the following: *Compliments of the season!* is translated as *Compliments de la saison!* (Vinay & Darbelnet, 1995)

There are two types of calque: lexical calque, which respects the syntactic structure of the target language and structural calque, which introduces a new construction into the target language. Here are examples of each of them:
Compliments of the Season! -- Compliments de la saison! is a lexical calque example and Science-fiction -- Science-fiction is a structural calque example (Venuti, 2000).

Moreover, calque can be used for different words such as common collocations, names of organizations and institutions and the components of compounds (Newmark, 1988).

One of the advantages of the calque technique is the application of implicature, a concept discussed by Baker and defined as what the speaker means or implies in its speech (1992).

Literal translation: this technique happens when a text can go from the source language to the target language with no changes, just the ones required by the grammar of the target language (Fawcett, 2003). Example given by Vinay and Darbelnet: *where are you?* is translated as Où êtes-vous? (Vinay & Darbelnet, 1955). When literal translation does not work properly, for instance, when it gives a different meaning than the intended one or when the literal translation means nothing in the target language, translators have to use a different method of translation and it is in these situations when oblique translation is the best option.

The following are the techniques used in oblique translation:

Transposition: this is used to deal with grammatical changes when translating. It is about "replacing one word class with another without changing the meaning of the message… in translation there are two distinct types of transposition: obligatory transposition and optional transposition" (Vinay & Darbelnet, 1995, p. 36). For example, *l´économie n´a cessé de croître* can be translated using transposition as *the economy grew steadily* or *the economy continued to grow* (Fawcett, 2003).

Modulation: it is a variation of the message by changing the point of view without changing the meaning. This variation is justified when the literal or transposed translation is grammatically correct but it is considered awkward in the target language; an example of this strategy is the following: *it is not difficult to show* can be translated as *Il est facile de démontrer* (Vinay & Darbelnet, 1955).

Equivalence: this technique is about conveying meanings when the two languages involved in the translation refer to the same situation in different ways. This strategy is mostly used when idioms and proverbs are part of the translation (Fawcett, 2003). Vinay and Darbelnet

mentioned in their book (1955) proverbs are perfect examples of this technique: *it is raining cats and dogs* is translated to the French language as *Il pleut à seaux* and *too many cooks spoil the broth* is translated as *deux patrons font chavirer la barque.*

Adaptation: this technique is used when the type of situation given in the source language message is unknown or it does not exist in the target language culture. To apply this strategy translators are to create a new situation that has to be considered as equivalent to the other one. Vinay and Darbelnet gave as an example the case in which translating literally *he kissed his daughter in her mouth*, which is pretty common in the English culture as a sign of paternal love and not a sign of romantic love, would be odd in the French culture and adaptation should be used in order to keep the meaning of the original message without misunderstanding it (Vinay & Darbelnet, 1995).

The techniques stated by Vinay and Darbelnet are the ones I am more related to. They were quietly explained during the translations courses at the English Language Major and we also practiced their use when translating certain technical texts. Thus, it was decided to use these techniques to translate the text "A Statistical Explanation of MaxEnt for Ecologists".

There have some other attempts to design translation techniques. In 1988, Malone (the American approach) describes further general techniques that contains specific processes to make great translations (Fawcett, 2003, pp.41-50):

1. Matching: substitution and equation. Basically, substitution encompasses Vinay and Darbelnet´s transposition, equivalence and adaptation while equation is about literal translation.
2. Zigzagging: divergence and convergence. These strategies have to do with the different lexical structuring between languages; Divergence is about one-to-many equivalence while convergence is about many source terms collapsing into only one in the target language. Basically, zigzagging happens when there are two or more words meaning the same thing but, mostly, with some connotation difference.
3. Recrescence: amplification and reduction. Amplification is mainly providing explanations so the reader comprehends better the translated text; reduction is the omission of information the translator considered to be unnecessary or of little importance for the comprehension of the translated text.

4. Repackaging: diffusion and condensation. Diffusion can occur for structural and grammatical reasons but also for translation decision-making in cases when complex concepts cannot be found in one of the languages involved in the translation. Condensation is less used, as translations are usually longer than the source texts.
5. Reordering: this technique is used because sometimes reordering word sequences is necessary for a better comprehension or simply because the source and target language have different grammatical and stylistic structures.

Some of Malone's techniques are quite related to Vinay and Darbelnet's techniques. Reviewing them helped us better understand the process of conveying the meaning of an idea from language into another one. Malone's techniques helped us understand how to justify adjustments of form depending on the semantic, stylistic and communicative requirements of the translated text.

In this chapter, a short review of translation history and some of the translation approaches have been presented as a frame to understand the importance of this communicational tool. I identified the approach I decided to use to translate the text we chose. Finally, some of the most important translation techniques were introduced. The next section explains the methodology used to carry out the translation of the text "A statistical explanation of MaxEnt for ecologists".

1.5 Method

In this chapter, I explain the process followed to carry out the translation of the text chosen. So as to convey the meaning of the source text, first it was necessary to familiarize myself with the topic and type of text I was going to translate. Thus, I engaged first in reading about two main topics: research on Biogeography and on programs for modelling species distribution. Since both topics were unknown for me when reading them I had to check several online sites and dictionaries on Biology and Ecology. I researched the same magazine in which the text chosen was published, downloaded other topic-related papers in Spanish and read them too. I also

reviewed some articles on MaxEnt program. It was necessary to look up some terms that were recurrently used and essential in the source text; here there are some of those terms:

English Source	Spanish Translation
Presence-only	Solo presencia
Relative entropy	Entropía relativa
Species occurrence	Ocurrencia de especies
Machine learning	Aprendizaje automatizado
Modelling method	Método de modelación
Predictor variables	Variables predictoras
Covariate space	Espacio covariable
Landscape of interest	Paisaje de interés
Log likelihood	Probabilidad de registro
Logistic output	Resultado logístico

Reading topic- related papers in Spanish helped me familiarized with the characteristics of the discourse, the written style of the text chosen and the terms used. Reading other papers written by some of the authors of the chapter I translated was also necessary to learn about their written style. Nevertheless, it was also necessary to ask for help to an expert on the topic of programs for modelling species distribution. I had the pleasant opportunity to discuss some technical terms with professor María Angélica Navarro Martinez who has a Phd in Tropical Ecology and a master in Forestry. Professor María A. Navarro is a current professor at El Colegio de la Frontera Sur and she is really involved in the topic; therefore, listening to her comments was remarkably important for my paper.

Some of the resources used to translate the paper on MaxEnt were the following:

Google translator**.** This is one of the most used translator available on internet. However, I strongly recommend this tool to look for single words. When being used to translate a complete idea or paragraph, most of the times you get a literal translation of the sentence or paragraph with no order and broken meaning.

Linguee. This is an online dictionary and a useful tool when trying to understand the use of a word in context; it provides with several examples of use for a word in different contexts. It is highly useful to look up words and phrases.

Wikipedia. Wikipedia is a widely used resource to look for information about almost anything. It is useful to get the general idea of some concepts; nevertheless, it is important to make sure that the information I am checking in this site has been checked by experts before trusting it's appropriateness.

Wiley online library**.** This is an online library with several articles from different specialized magazines, including Diversity and Distributions, which is the topic of the article chosen to be translated in this work. It is useful to get used to the terms applied in the ecology field.

As stated before, the field of translation has had several approaches. Therefore, several techniques and strategies to attempt an accurate translation of the source text has been developed. Regarding my translation paper, after analyzing the different approaches and taking into consideration my experience of the use of those strategies, I decided to use the dynamic translation model by Jean-Paul Vinay and Jean Darbelnet to carry out the translation of the paper at issue.

It is important to mention that even though the dynamic translation strategies by Vinay and Darbelnet were chosen to be used to translate the text, I decided to complement these strategies with some of the techniques in the American model, already explained in previous pages. I decided to complement Vinay and Darbelnet's strategies to improve the translation of the chosen article.

Following, I will explain step by step how I proceeded to translate the text chosen.

1. I read the article several times and made notes regarding the new terms and concepts.
2. I read some articles related to the field and topic of my article to get familiarize with the language used in this type of papers.
3. I made a list of terms used in the paper and looked for their equivalent or translation in different sources.
4. I translated the first 5 pages of the paper and reviewed them line by line with my assessor. Analyzing the translation done helped me clarify some of the ideas from the original text but also allowed me to identify some mistakes I made when writing in Spanish.
5. I started researching about the different translation approaches and techniques in order to better convey the information from the source article.

6. I rewrote some of the paragraphs I had translated and finished translating the rest of the paper. Every single page of the translation was checked with my main advisor.
7. The first translation of the whole article was reviewed by an ecologist who has used the program MaxEnt in some of her researches.
8. She reviewed the article and corrected the use of some of the terms and commented on some ideas that had not been accurately translated.
9. I reviewed the whole translation of the paper and integrated the suggestions the specialist commented.
10. I then did a second version of the translation. Some adjustments were required. This version was reviewed again with my assessor to standardize the use of certain forms and terms.
11. After being analyzed by my main assessor, the translation was given to two other advisors to be checked and the adjustments were done to have a suitable final paper.

Regarding the use of the strategies chosen to translate the text, I must say that I started to translated the source text without having in mind that I had to use the strategies but as soon as I encountered a problem to translate I made use of some of the strategies. Having to analyze the use of the strategies allowed me to determine the form of the language that would better convey the intended meaning of the source text. Eventually and as the translation progressed, I became more analytical and skillful in the use of the translation strategies. At the same time, it became more evident the appropriateness of using a certain strategy.

Once I had the final version of the translation, I started to look for the arguments that would support the use of the strategies and the choice I made of the language structures in Spanish. While analyzing and justifying the translation of certain parts of the original text I sometimes realized that the Spanish being used was not the correct one and had to re write the paragraph. Writing the explanation of the use of the strategies helped me to better understand the translation strategies. By the end of the analysis I realized this was an excellent way to better understand the use of both languages: English and Spanish. Translating enhances the way we produce language in the written form; I personally think that translating specialized articles boosts our knowledge and increases our vocabulary in fields we are not aware of. Grammatical structures are applied in every single sentence or phrase we translate, by translating, we are more conscious of the uses of words and grammatical features in both languages.

CHAPTER II

TRANSLATION

Diversity and Distributions, (Diversity, Distrib.) (2011) 17, 43-57

Investigación sobre la Biodiversidad

UNA EXPLICACIÓN ESTADÍSTICA DE MAXENT PARA ECÓLOGOS

Jane Elith[1]*, Steven J. Philips[2], Trevor Hastie[3] , Miroslav Dudík[4] , Yung En Chee[5] y Colin J. Yates[6]

RESUMEN

MaxEnt es un programa para la modelación de la distribución de especies con base en registros solo de presencia. Este documento va dirigido a ecólogos y describe el modelo de MaxEnt desde una perspectiva estadística, haciendo explícitos los vínculos entre la estructura del modelo, las decisiones requeridas para producir una distribución modelada, y el conocimiento sobre las especies y los datos que podrían afectar las decisiones. Para empezar, discutiremos las características de los datos de presencia de especies, enfatizando las implicaciones para los modelos de distribución. Particularmente, nos enfocamos en los problemas del sesgo muestral y en la falta de información sobre el predominio de especies. La piedra angular de este documento

1 Escuela de Botánica, Universidad de Melbourne, Parkville, Victoria 3010 Australia
2 Laboratorios AT&T, Investigación, Avenida Parque 180, parque Florham, Nueva Jersey 07932, Estados Unidos de América
3 Departamento de Estadísticas, Universidad de Stanford, California 94305, Estados Unidos de América
4 Laboratorios de Yahoo!, Oeste 111, Calle 40 (Séptimo piso). Nueva York, Nueva York 10018, Estados Unidos de América
5 Escuela de Botánica, Universidad de Melbourne, Parkville, Victoria 3010 Australia
6 División de Ciencia, Departamento de Australia Occidental del medio ambiente y su Conservación, LMB 104, Centro de Distribución de Bentley, WA6983, Australia

es una nueva explicación estadística de MaxEnt, la cual muestra que el modelo minimiza la entropía relativa entre dos densidades de probabilidad (una estimada de los datos de presencia y la otra del paisaje), definidas en el espacio de covarianza. Para muchos usuarios este punto de vista es probablemente una forma más accesible para entender el modelo que la forma original basada en los conceptos de aprendizaje automatizado. Enseguida, proporcionamos una explicación detallada de MaxEnt describiendo sus componentes principales (por ejemplo: covariables y características, y la definición de la extensión del hábitat), los mecanismos de ajuste del modelo (por ejemplo: la selección de las características, las dificultades y su regularización) y los resultados. Utilizamos estudios de caso de una especie de *Banksia*, originaria del suroeste de Australia y de un pez ribereño, ajustamos modelos, los interpretamos explorando por qué ciertas decisiones afectan el resultado y lo que esto significa. El ejemplo del pez ilustra el uso del modelo con datos vectoriales para segmentos lineales del río en lugar de datos ráster (datos covariables). Mostramos tratamientos apropiados para el sesgo muestral, datos imprevistos, especies localmente restringidas y predicciones sobre los entornos fuera del rango de los datos de entrenamiento y discutimos nuevas capacidades de este programa. Los apéndices en línea incluyen detalles adicionales del modelo y de las relaciones matemáticas entre explicaciones anteriores y la expuesta en este artículo, código de ejemplo y datos, y demás información de los estudios de caso.

Palabras Clave

Ausencia, nicho ecológico, entropía, aprendizaje automatizado, solo presencia, modelo de distribución de especies

ABSTRACT

MaxEnt is a program for modeling species distributions from presence-only species records. This paper is written for ecologists and describes the MaxEnt model from a statistical perspective, making explicit links between the structure of the model, decisions required in producing a modeled distribution, and knowledge about the species and the data that might affect those

decisions. To begin we discuss the characteristics of presence-only data, highlighting implications for modeling distributions. We particularly focus on the problems of sample bias and lack of information on species prevalence. The keystone of the paper is a new statistical explanation of MaxEnt, which shows that the model minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space. For many users, this viewpoint is likely to be a more accessible way to understand the model than previous ones that rely on machine learning concepts. We then step through a detailed explanation of MaxEnt describing key components (e.g. covariates and features, and definition of the landscape extent), the mechanics of model fitting (e.g. feature selection, constraints and regularization) and outputs. Using case studies for a Banksia species native to south-west Australia and a riverine fish, we fit models and interpret them, exploring why certain choices affect the result and what this means. The fish example illustrates use of the model with vector data for linear river segments rather than raster (gridded) data. Appropriate treatments for survey bias, unprotected data, locally restricted species, and predicting to environments outside the range of the training data are demonstrated, and new capabilities discussed. Online appendices include additional details of the model and the mathematical links between previous explanations and this one, example code and data, and further information on the case studies.

Keywords

Absence, ecological niche, entropy, machine learning, presence-only, species distribution model.

INTRODUCCIÓN

Los modelos de distribución de especies (SDMs por sus siglas en inglés “Species Distribution Models) estiman la relación entre registros de especies en sitios y las características ambientales y/o espaciales de los sitios (Franklin 2009). Son ampliamente usados para diversos propósitos en biogeografía, biología de la conservación y ecología (Elith y Leathwick, 2009a; tabla 1). En las

dos últimas décadas, ha habido un gran desarrollo en el campo de la modelación de la distribución de especies y, hoy están disponibles numerosos métodos. La principal diferencia entre los métodos es el tipo de datos de las especies que usan. En situaciones donde los datos de las especies han sido colectados sistemáticamente – por ejemplo en un estudio biológico formal en el cual un conjunto de sitios son estudiados, registrando la presencia/ausencia o la abundancia de especies en cada sitio – se han usado métodos de regresión con los cuales la mayoría de los ecólogos están familiarizados (por ejemplo el modelos lineal generalizado y los modelos aditivos, GLMs o GAMs por sus siglas en inglés; o los conjuntos de árboles de regresión: bosques aleatorios o árboles de regresión reforzados, BRT por sus siglas en inglés)

Sin embargo, para la mayoría de las regiones, los datos de estudios biológicos sistemático tienden a ser dispersos y/o limitados en cobertura. Los registros en las bases de datos de los herbarios y museos representan registros de presencia de las especies. Muchas de estas bases representan poco más de un siglo de inversión pública y privada en la ciencia biológica y son una fuente muy importante de datos de ocurrencia de especies. El deseo por maximizar la utilidad de tales fuentes ha generado una variedad de métodos SDM para modelar la distribución de las especies solo con datos de presencia. MaxEnt (Philips *et al.*, 2006; Philips y Dudik, 2008) es uno de estos métodos y es el punto central de este documento.

La capacidad predictiva de MaxEnt es consistentemente competitiva con los métodos de mayor desempeño (Elith *et al.*, 2006). Desde el 2004, este método ha sido extensamente utilizado para modelar la distribución de las especies. Ejemplos publicados cubren diversos objetivos (encontrar correlación en la ocurrencia de especies, mapear distribuciones actuales y predecir la distribución de especies en el tiempo y el espacio), a través de diversas aplicaciones en ecología, evolución, biología de la conservación y bioseguridad (Tabla 1). Organizaciones gubernamentales y no gubernamentales también han adoptado MaxEnt para mapear la biodiversidad a gran escala en el mundo real, incluyendo la aplicación en línea del Observatorio de Pájaros de Point Reyes (http://www.prbo.org/) y el atlas de Australia viviente (http://www.ala.org.au/). Jane Elith (JE) y Steven J. Phillips (SJP) participaron en estos programas e identificaron la necesidad de proporcionar una explicación ecológicamente accesible de MaxEnt. Las descripciones existentes incluyen conceptos de aprendizaje automatizado que suelen ser desconocidos para la mayoría de los ecólogos.

En este artículo, explicamos el método de modelación de MaxEnt con énfasis en la explicación de la estadística del método, lo que este asume y los impactos de las decisiones tomadas durante el proceso de modelación. Usamos dos estudios de caso para examinar los efectos de la selección del contexto y la configuración del modelo y, para ilustrar la aplicabilidad del modelo para explorar relaciones ecológicas a pequeña escala, datos ambientales basados en vectores. Nuestro propósito es promover el entendimiento del método y recomendar aproximaciones útiles para la preparación de datos, el ajuste del modelo y su interpretación.

Tabla 1. Ejemplos de estudios publicados utilizando MaxEnt que demuestran variación en sus propósitos, extensiones y organismos.

Propósito principal	Extensión	Organismos	Referencias
Predicción de distribuciones actuales como insumo para la planificación de la conservación, la evaluación de riesgos o listado IUCN, o para nuevos estudios	Los Andes	Colibríes	Tinoco *et al.* (2009)
	Global	Corales pétreos	Tittensor *et al.* (2009)
		Montes submarinos	
		Macrohongos	Wollan *et al.* (2008)
	Noruega	Gato montés europeo	Monterroso *et al.* (2009)
	Portugal		
Comprensión de correlaciones ambientales de ocurrencia de especies, grupos de especies u otros	Nueva Zelanda	Hormigas	Ward (2007a)
	China	Nemátodos	Wang *et al.* (2007)
			Graham & Hijmans (2006)
Predicción de distribución potencial para especies invasivas o para explorar la expansión de la	California	Anfibios y reptiles	Murray- Smith *et al.* (2009)
	Brasil	Mirtáceas	
		19 especies	Verbruggen *et al.* (2009)

distribución		Algas Marinas	Young *et al.* (2009)
	Global	Pájaros	Lamb *et al.* (2008)
Predicción de la riqueza de especies o la diversidad	Los Andes Madagascar Noroeste de Europa Costa Brasileña	Murciélagos Caracoles de estanque Bosques	Cordellier y Pfenninger (2009) Carnaval y Moritz (2008) Yesson y Culham (2006)
Predicción de distribución actual para el entendimiento de la diversidad morfológica/ genética ("filogeografía","estudios filoclimáticos"), dinámica de evolución de nichos, endemismos distribuciones Hindcast para entender patrones de endemismo y vicariancia, etc.	El Mediterráneo y sus alrededores Región del Oeste de Australia Canadá Patagonia	Ciclamen Banksia Mariposas Insectos	Yates *et al.* (2010) Kharouba *et al.* (2009) Tognelli *et al.* (2009) Williams *et al.* (2009) Elith *et al.* (2006)
Pronóstico de distribuciones para entender los cambios con el cambio climático/ transformación de la tierra; incluye estudios retrospectivos	Región Local en California De Local a Nacional	Plantas extrañas Especies variadas	

Prueba de desempeño del
modelo frente a otros
métodos

PREÁMBULO: ¿QUÉ TIENEN DE ESPECIAL LOS CASOS SOLO DE PRESENCIA?

Ampliar el uso de datos de solo presencia para modelar la distribución de especies ha estimulado una amplia discusión acerca de los tipos de distribución (por ejemplo: potencial vs. realizado) que pueden ser modelados con datos solo de presencia de especies en contraste con datos de presencia-ausencia (por ejemplo: Soberón & Peterson, 2005; Chefaoui & Lobo, 2007; Hirzel & Le Lay, 2008; Jiménez- Valverde *et al.*, 2008; Soberón & Nakamura, 2009; Lobo *et al.*, 2010). Como se menciona en varios de estos artículos, el sujeto es complejo debido a la interacción de la calidad de los datos (cantidad y precisión de los datos de especie; relevancia ecológica de las variables predictoras; disponibilidad de información sobre disturbios, limitantes para la dispersión interacciones bióticas), el método de modelación y la escala de análisis. Una revisión comprehensiva de estos aspectos podría ser útil, pero aquí nos limitamos a los puntos importantes para este documento.

Algunas de las publicaciones sugieren que los datos solo de presencia de especies de alguna manera no presentan los problemas de confiabilidad que generan los registros de ausencia de especies (Jiménez-Valverde *et al.*, 2008), enfatizando particularmente que las ausencias producen importantes marcas en las interacciones bióticas, restricciones en la dispersión y disturbios que se podrían excluir en el modelado de distribuciones potenciales (de acuerdo con Svenning & Skov, 2004). Sin embargo, los registros de presencia también están marcados por muchos de los actores que afectan las ausencias. Si una especie está ausente de un área ecológicamente disponible debido a, digámoslo así, disturbios pasados han ocasionado extinciones locales, la señal de esa ausencia será encontrada en los registros de distribución de

presencia: estos no serán registro de presencia en el área perturbada. Independientemente de si las ausencias son usadas en la modelación, el patrón en los registros de presencia sugiere que el área no está disponible y el modelo será afectado por este patrón. De manera similar, si la detectabilidad de una especie particular varía de un sitio a otro, entonces, no solo esto resulta en algunas ausencias falsas en datos de presencia-ausencia sino que también afecta los patrones de presencia en datos de solo presencia. Esto nos lleva naturalmente a la conclusión de que modelar con ausencias no resuelve las limitantes comúnmente atribuidas a los datos de ausencia, tales como el hecho de que las especies no son perfectamente detectables y que pueden no ocupar hábitats idóneos. Este razonamiento significa que abordaremos la descripción del problema de modelado con datos solo de presencia como aquel tratando de modelar la misma cantidad que es modelada con datos de presencia-ausencia, esto es, la probabilidad de presencia de una especie (la cual será definida más adelante).

A partir de este momento, asumimos que los datos disponibles para el modelador son solo de presencia, por ejemplo, un conjunto de localidades dentro de L, el paisaje de interés, donde la especie ha sido observada. Asumamos que $y = 1$ denota presencia de la especie, $y = 0$ denota ausencia, $\mathbf{z}$ indica un vector de covariables ambientales, y que el contexto será definido por todas las localidades dentro de L (o una muestra tomada al azar del mismo). Asumimos que las variables ambientales o covariables $\mathbf{z}$ (las cuales representan condiciones ecológicas) están disponibles a lo largo del paisaje. Se define a $f(\mathbf{z})$ como la densidad de probabilidades de covariables a través de L, y $f_1(\mathbf{z})$ como la densidad de probabilidades de covariables a través de las localidades dentro de L donde la especie está presente, y de forma similar, $f_0(\mathbf{z})$ donde la especie está ausente (densidades- o funciones de densidad de probabilidades - describe la probabilidad relativa de que las variables aleatorias a través de su rango puedan ser univariadas o multivariadas). La cantidad que deseamos estimar es, como con datos de presencia-ausencia, la probabilidad de presencia de la especie, condicionada por el ambiente: $\Pr(y = 1 | \mathbf{z})$. Estrictamente, los datos solo de presencia permiten modelar $f_1(\mathbf{z})$, el cual no puede aproximarse a la probabilidad de presencia. Los datos de paisaje permite modelar tanto, $f_1(\mathbf{z})$ como $f(\mathbf{z})$, y esto obtiene una constante $\Pr(y = 1 | \mathbf{z})$, porque la regla de Bayes nos da la siguiente expresión:

$$\Pr(y = 1 | \mathbf{z}) = f_1(\mathbf{z}) \Pr(y = 1) / f(\mathbf{z}) \quad (1)$$

La única cantidad que falta es la del segundo término, Pr (y = 1), por ejemplo, la prevalencia de la especie (la proporción de los sitios ocupados por la misma) en el paisaje. Formalmente, decimos que la prevalencia no es identificable de los datos de presencia (Ward *et al.,* 2009). Esto significa que la distribución no puede ser exactamente determinada, sin importar el tamaño de la muestra; lo cual es una limitación fundamental de los datos de solo presencia. Notamos, sin embargo, que los datos de ausencia están plagados de problemas de detección de probabilidad (Wintle *et a*l., 2004; MacKenzie, 2005) por lo que inclusive los datos de presencia-ausencia puede no proporcionar una buena estimación de la prevalencia.

Una segunda limitación fundamental de los datos de presencia- ausencia es que la selección del sesgo muestral (donde algunas áreas en el terreno son muestreadas más intensamente que otras) tiene un mayor efecto en modelos de solo presencia que en modelos de presencia- ausencia (Philips *et al.*, 2009). Imagina que $f_1(\mathbf{z})$ está contaminado por una selección de sesgo muestral $s(\mathbf{z})$. Este sesgo ocurrirá más comúnmente en un espacio geográfico (por ejemplo, cerca de caminos) pero podría ocurrir por el medio ambiente (por ejemplo, cerca de barrancos) pero, sin importar, va a aparecer en espacio covariado. Bajo un muestreo sesgado, un modelo de solo presencia da un estimador de $f_1(\mathbf{z})s(\mathbf{z})$ en lugar de $f_1(\mathbf{z})$. Esto es debido a que obtenemos un modelo que combina la distribución de especies con la distribución del esfuerzo de muestreo (Soberón & Nakamura, 2009). En contraste, para los modelos de presencia-ausencia, los sesgos en la selección de la muestra afectan ambos registros, tanto los de presencia como los de ausencia y el efecto de los sesgos se anula (bajo suposiciones razonables, ver a Zadrozny, 2004).

Hasta el momento hemos tratado la presencia o la ausencia como un evento binario, pero en realidad definir la variable de respuesta no es sencillo y, respecto a esto, los datos de solo presencia son un tanto diferentes de los de ausencia (Pearce y Boyce, 2006). La presencia o ausencia de una especie depende de la escala de tiempo y de la escala espacial, por ejemplo, una especie con la facilidad de moverse libremente (tal como un pájaro) puede estar presente en algunas ocasiones, pero en otras no, mientras que una especie de planta puede tener mayores probabilidades para ser encontrada en una parcela grande con ciertas condiciones ambientales que en una parcela pequeña con las mismas condiciones. La ausencia de una especie de plantas en un cuadrante de 1- km^2 alrededor de un punto implica ausencia de un cuadrante en un 1-m^2 alrededor de un punto, pero no viceversa. Con datos de presencia-ausencia, no es difícil

incorporar estas complejidades en la formulación de la variable de respuesta (por ejemplo, la especificación de lo que constituye una muestra), o vía muestras covariables en el modelo, proporcionan los detalles (Leathwick, 1998; MacKenzie & Royle, 2005; Schulman *et al.*, 2007; Ward, 2007b). Sin embargo, con los datos de solo presencia, normalmente tenemos datos que no tienen ninguna escala temporal o espacial asociada. El registro es usualmente un simple registro de la especie en una localidad sin ninguna información acerca del área o del tiempo de la misma.

Con los datos de presencia-ausencia, la definición de la variable de respuesta debería naturalmente ser consistente con el método de muestreo. Por ejemplo, si la información disponible es cuadrantes de 1-m^2, entonces y = 1 debería corresponder a la presencia de la especie en un cuadrante de1-m^2. Con datos de solo presencia, la información disponible no siempre describe el método de muestreo, por lo que el modelador tiene considerables libertades en definir la variable de respuesta. Un enfoque común es asumir implícitamente una unidad de muestreo de tamaño equivalente al de la granularidad de las unidades de información ambiental disponibles (ver la discusión sobre granularidad de Elith & Leathwick, 2009[a]).

En pocas palabras, nosotros planteamos que, con datos de presencia e información de contexto, podemos modelar la misma cantidad que con datos de presencia-ausencia, dependiendo de la constante Pr(y=1). Sin embargo, si los datos de presencia-ausencia están disponibles, creemos es generalmente aconsejable usar el método de modelado de presencia- ausencia, ya que en ese caso, los modelos son menos susceptibles a problemas con los sesgos en la selección de la muestra, el método de muestreo será conocido y podrá usarse para definir apropiadamente la variable de respuesta para el modelado y podemos tomar ventaja de toda la información en los datos. En particular, los datos de presencia ausencia dan mejor información que los de solo presencia acerca de la prevalencia, porque, aún que puede haber algunas dificultades debido a la detección imperfecta, resuelven el mayor problema de no identificación. Volveremos a esto cuando discutamos los resultados del modelo logístico de MaxEnt.

EXPLICACIÓN DE MAXENT

Por primera vez, describiremos la terminología y notación estadística usada en MaxEnt, dejando de lado la terminología de aprendizaje automatizado utilizada en artículos anteriores. Al tiempo

que describiremos el modelo, resaltaremos posibilidades e implicaciones para opciones de modelado y valores predeterminados, y consideraremos como MaxEnt aborda la limitante de los datos de solo presencia identificados en párrafos anteriores. Relegamos las consideraciones más técnicas a los recuadros y la información complementaria, para evitar cortar el flujo de la explicación.

Covariables y características

La mayoría de los ecólogos, que usan la literatura estadística, llaman covariables, predictores o insumos a las variables independientes de un modelo. Los SDMs, incluyen factores ambientales que son relevantes para que el hábitat sea apto (por ejemplo, estimados del clima, de la topografía, del suelo para las plantas; la temperatura, la salinidad y la abundancia de presas para peces marinos). Debido a que la respuesta de la especie a estos tiende a ser compleja, es comúnmente deseable ajustar funciones no lineales (Austin, 2002). En regresión esto puede ser logrado mediante la aplicación de transformaciones a las covariables, por ejemplo, creando funciones base para polinominales y splines, incluyendo funciones lineales definidas en pasos. Los modelos complejos se establecen como combinaciones lineales de estas funciones base en métodos que incluyen GLMs y GAMs (Hastie *et al.,* 2009, capítulo 5). En el aprendizaje automatizado, las funciones base y otras transformaciones de datos disponibles son llamados rasgos -por ejemplo, rasgos son un conjunto amplio de transformaciones de las covariables originales.

En MaxEnt, los rasgos seleccionados se forman "más allá de las escenas", de la misma forma que en regresión, donde la matrix modelo es aumentada por términos especificados en el modelo (ejemplo, polinominales, interacciones). La función ajustada de MaxEnt es usualmente definida mediante muchos rasgos, lo que significa que en muchos modelos habrá rasgos que covarían. Actualmente MaxEnt tiene seis clases de rasgos: linear, producto, cuadrática, eje, umbral y categórica (para más detalles, ver Apéndice S1). Producto son los productos de todas las posibles combinaciones de pares de covariables, permitiendo interacciones simples ajustadas. El modelo de umbral permiten dar "un paso" en el ajuste de la función; el modelo de eje es similar excepto que en este se permite un cambio en la gradiente de la respuesta. Muchos rasgos de

umbral y de eje pueden ser ajustadas por una covariable, proporcionando una función potencialmente compleja. El modelo de eje (que es una función básica para splines lineales), si se usan solos, permiten un modelo aditivo generalizado (GAM por sus siglas en inglés): un modelo aditivo, con funciones no lineales ajustadas de complejidad variable pero sin los pasos de avance de los modelos de umbral. El valor predeterminado de MaxEnt permite todos los tipos de modelos (dependiendo de la cantidad de información que se tenga de la especie); sin embargo, vale la pena considerar los modelos más simples, tal y como se discute posteriormente, bajo implicaciones para el modelado.

El modelo MaxEnt- un repaso rápido

Trabajos anteriores han descrito a MaxEnt como una estimación a la distribución a través del espacio geográfico (Philips *et al.*, 2006; Philips & Dudik, 2008). Aquí hacemos una caracterización diferente (pero equivalente) que se centra en comparar densidades probables en un espacio covariable (Figura 1). Al hacer esto, confiamos en la investigación de doctorado del ex alumno del Profesor en estadística Trevor Hastie, Gill Ward (Ward, 2007b) y agradecemos su contribución. La ecuación 1 nos muestra que, si sabemos la densidad condicional de las covariables en los lugares de presencia, $f_1(\mathbf{z})$, y la marginal (por ejemplo, incondicional) densidad de covariables a través del área de estudio $f(\mathbf{z})$, por ende solo necesitamos conocimiento de la prevalencia de Pr(y=1) para calcular la probabilidad condicional de ocurrencia. MaxEnt hace un estimado del radio $f_1(\mathbf{z})/f(\mathbf{z})$, salida de datos en "bruto" de MaxEnt. Este es el núcleo del modelo de salida de MaxEnt, permite identificar qué características son importantes y la estimación de la relativa idoneidad de un lugar frente al otro. Debido a que la información requerida sobre la prevalencia no está disponible para calcular la probabilidad condicional de ocurrencia, se ha implementado una solución alternativa (denominada producción logística de MaxEnt). Esto permite registrar el producto de salida como una valoración logit: $\eta(\mathbf{z}) = \log (f_1(\mathbf{z})/f(\mathbf{z}))$ y calibrar la intercepción, de manera que la probabilidad implícita de presencia en sitios con condiciones "típicas" para la especie (por ejemplo, donde $\eta(\mathbf{z})$ = el valor promedio de $\eta(\mathbf{z})$ por debajo de f_1) es un parámetro τ. El conocimiento de τ resolvería la no identificabilidad de la prevalencia y, en la ausencia de ese conocimiento, MaxEnt arbitrariamente fija τ a 0.5. Esta transformación logística

es monótona (preservación del orden) en la salida de datos en bruto. Trabajamos en las siguientes secciones de cada parte del modelo MaxEnt, mostrando cómo la elección del paisaje, los datos de las especies y los ajustes seleccionados influyen en los resultados.

El Registro de Especies y el Paisaje

El paisaje de interés (L) es un área geográfica determinada por el problema y definida por el ecologista. Podría, por ejemplo, estar limitada por límites geográficos o por una comprensión de qué tan lejos las especies focales podrían haberse dispersado. Entonces, definimos *L1* como el subconjunto de *L* donde la especie está presente.

La distribución de las covariables en el paisaje es transmitida por una muestra finita, una colección de puntos de L con covariables asociadas, típicamente llamada una muestra de fondo. Estos datos pueden suministrarse en forma de cuadrículas de covariables que cubren una pixelación del paisaje; como un valor predeterminado MaxEnt aleatoriamente muestra 10.000 ubicaciones de fondo de las redes de covariable, los puntos de datos de antecedentes también pueden ser especificados (ver Yates *et al.,* 2010 y estudios de caso presentados más adelante) y las cuadrículas no son esenciales (estudio de caso 2). Cabe mencionar que el ambiente de muestra no tiene en cuenta los lugares donde hay presencia, es simplemente una muestra de *L,* sin embargo, puede incluir por casualidad lugares donde hay presencia. El uso de un ambiente de muestra aleatoria implica la creencia de que la muestra de registros de presencia puede también ser una muestra aleatoria de *L1*. Nos ocuparemos más adelante del caso de las muestras sesgadas.

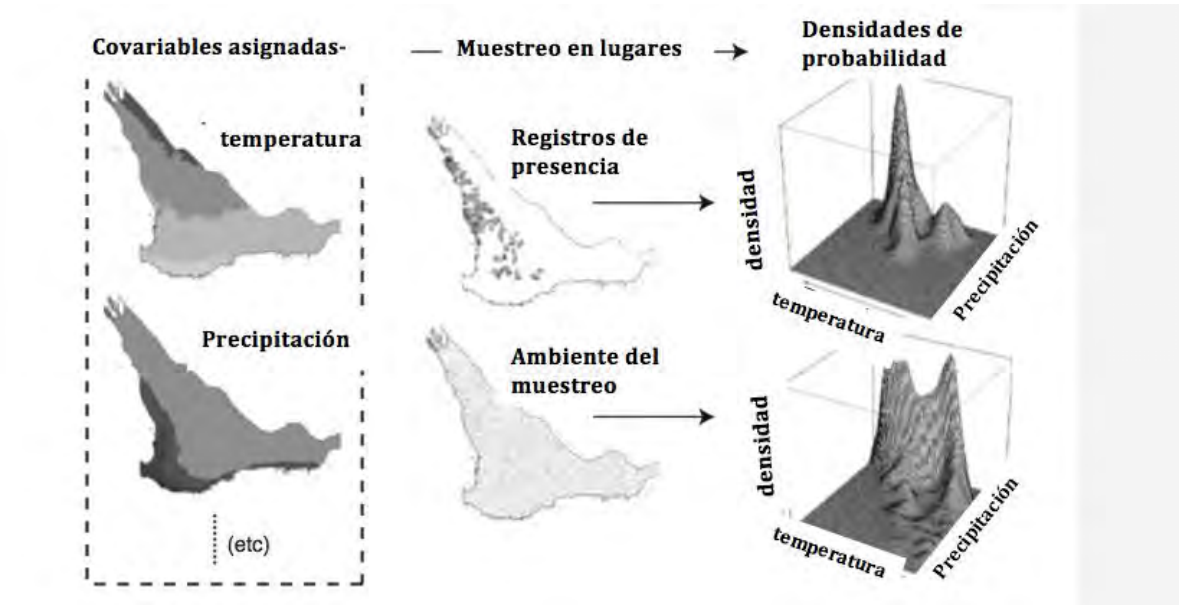

Figura 1. Presentamos una representación diagramática de las densidades de probabilidad relevante para nuestra explicación estadística utilizando datos obtenidos en el estudio de caso 1. Los mapas de la izquierda son dos ejemplos de covariables asignadas (temperatura y precipitación). En el centro se encuentran los lugares de presencia y el ambiente del muestreo. Las densidades estimadas presentadas a la derecha no están ubicadas en el espacio geográfico (en mapa), pero muestran la distribución de valores en el espacio covariable para la presencia (parte superior derecha) y el ambiente del muestreo (parte inferior derecha). Éstas pueden representar las densidades $f_1(\mathbf{z})$ y $f(\mathbf{z})$como un modelo simple con características lineares.

Descripción del modelo

MaxEnt utiliza los datos covariables de los registros de ocurrencia y la muestra de fondo para estimar la relación entre $f_1(\mathbf{z})/f(\mathbf{z})$. Esto lo hace una estimación de $f_1(\mathbf{z})$ consistente con los datos de ocurrencia; muchas de tales distribuciones son posibles, pero elige la que está más cerca de $f(\mathbf{z})$. La minimización de la distancia de $f(\mathbf{z})$ es razonable, porque $f(\mathbf{z})$ es un modelo nulo para $f1(\mathbf{z})$: sin datos de ocurrencia, no tendríamos razones para esperar que la especie prefiriera condiciones ambientales particulares sobre ninguna otra; no podríamos hacer nada mejor que predecir que la especie ocupa las condiciones ambientales proporcionalmente a su disponibilidad

en el paisaje. En MaxEnt, esta distancia de f(**z**) se toma como la entropía relativa de f_1(**z**) con respecto a f(**z**) (también conocida como la divergencia de Kullback-Leibler).

El uso de datos de fondo informa al modelo acerca de f(**z**), acerca de la densidad de covariables en la región y proporciona la base para la comparación con la densidad de covariables ocupada por la especie (por ejemplo, f_1(**z**) (Figura 1). Se imponen restricciones para que la solución sea una que refleje la información de los registros de presencia. Por ejemplo, si una covariable es la lluvia estival, entonces las limitaciones aseguran que la precipitación media de verano para la estimación de f_1(**z**) es cercana a su media a través de las ubicaciones con presencias observadas. De tal manera, la distribución de la especie se estima minimizando la distancia entre f_1(**z**) y f(**z**) sujeto a limitar la precipitación media de verano estimada por f_1 (y los medios de otras covariables) para estar cerca de la media a través de ubicaciones de presencia.

Tenemos en cuenta que los documentos anteriores que describen MaxEnt se centraron en una definición basada en la ubicación en un paisaje finito (normalmente una cuadrícula de píxeles). Llamaremos a esto una definición basada en el espacio geográfico y la comparamos con nuestra nueva descripción, que se centra en el espacio ambiental (covariable). Hay que tener en cuenta, sin embargo, que no estamos insinuando por esta formulación que en ninguna de las definiciones haya alguna consideración de la proximidad geográfica de lugares a menos que se usen predictores geográficos. En la definición original (Phillips *et al.,* 2006), el punto central era $\pi(x) = Pr(x|y = 1)$, es decir una distribución de probabilidad sobre píxeles (o ubicaciones) x. A esto se le llamó la distribución de salida en "bruto" (Phillips *et al.,* 2006) y arrojó la probabilidad, en los casos en los que la especie está presente, de que se encuentre en el pixel x. Maximizar la entropía de la distribución en bruto equivale a minimizar la entropía relativa de f_1(**z**) respecto de f (z), por lo que las dos formulaciones son equivalentes (véase el Apéndice S2 para las ecuaciones que muestran la transición de las definiciones geográficas a las ambientales). El modelo nulo para la distribución cruda fue la distribución uniforme sobre el paisaje, ya que sin datos no tendríamos razón alguna para pensar que la especie preferiría un lugar y no otro. Como se mencionó al comienzo de esta sección, en el espacio ambiental, el modelo nulo equivalente para z es f (z).

Las restricciones se describieron anteriormente en referencia a covariables, pero como se explica en la sección sobre covariables y características, MaxEnt realmente se adapta al modelo con base en las características productos de las transformaciones de las covariables. Esto permite modelar relaciones potencialmente complejas. Las limitaciones se extienden de ser limitaciones

en los medios de covariables a ser restricciones en los medios de las características. Llamaremos h(z) al vector de rasgos y ß (nota, esta notación es diferente a los trabajos anteriores: Tabla 2) al vector de coeficientes. Como se ha explicado, en Phillips *et al.* (2006), minimizar los resultados de entropía relativa da como resultado una distribución de Gibbs (Della Pietra *et al.*, 1997) que es un modelo de familia exponencial:

$$F_1(\mathbf{z}) = f(\mathbf{z})e^{\eta(z)} \qquad (2)$$

Donde $\eta(\mathbf{z}) = \alpha + ß \cdot h(\mathbf{z})$

Y α es una constante de normalización que asegura que $f_1(\mathbf{z})$ integre (sumas) a 1.

A partir de esto, está claro que el objetivo de un modelo MaxEnt es $e^{\eta(z)}$, mismo que estima la relación $f_1(\mathbf{z})/f(\mathbf{z})$. Es un modelo log-linear, similar en forma a un GLM y depende de las muestras de presencia y las muestras de fondo que se utilizan en la formación de la estimación. Por lo tanto, la definición del paisaje está íntimamente ligada a la solución que se da.

Tabla 2. Terminología utilizada en este documento.

Elemento/Concepto	**Definición**	**Observación**
Ambiente	Muestra de sitios del escenario	
Entropía	Indicador de dispersión. Documentos anteriores describían al modelo como la maximización de la entropía en un espacio geográfico; este documento se enfoca en minimizar la entropía relativa en espacio covariable.	
Características	Conjunto amplio de transformaciones de las covariables originales.	
Marcador	Capa cuadriculada de 1/0 dato utilizado para	

	indicar áreas incluidas en el ambiente del muestreo (=1) y excluidas (cero datos). Para ser incluido como indicador. Para poder proyectarlos a toda una región, pero sin incluir parámetros (a modo de ejemplo, 1 a través de toda la región de interés), debería suministrarse junto con todas las redes covariables una cuadrícula llamada marcador.	
Mapa MESS	Superficie de similitud ambiental multivariable (por sus siglas en inglés, MESS). Mide la similitud entre cualquier sitio dado y un conjunto de sitios de referencia, con base en las variables predictoras seleccionadas. Reporta la cercanía del sitio con relación a la distribución de los sitios de referencia, proporciona parámetros negativos de los diferentes sitios y traza estos parámetros a través de toda la región de predicción (Elith *et al., 2010*).	
La prevalencia no es identificable	La prevalencia no puede ser determinada con exactitud a través de datos de presencia de manera aislada sin importar el tamaño de la muestra. Ésta es una limitación fundamental de los datos de presencia de especies.	
Funciones de densidad de probabilidad	Describe la probabilidad relativa de variables aleatorias sobre su rango; puede ser univariado o multivariado.	
Parámetros de regularización (de	Regularización se refiere a facilitar el modelo, hacerlo más regular, esto para evitar la complejidad	ß en documentos anteriores* ʎ en este documento

ajuste)	en la adaptación del modelo. En MaxEnt, los parámetros de regularización pueden ser cambiados si así se requiere.	
Sesgo muestral	Algunas áreas en el escenario son muestreadas más intensivamente que otras. Lo anterior ocurre mayormente debido a las características del espacio geográfico o por cuestiones ambientales.	s(**z**)
Pesos o coeficientes	Parámetros del modelo en los que se mide la contribución de cada característica.	ʎ en documentos anteriores* ß en este documento

***** Philips *et al.* (2006), Phillips y Dudík (2008)

Mecánica de la solución

Para llegar a una solución, MaxEnt necesita encontrar coeficientes (betas) que cubran las limitaciones aunque no al grado de que las sobreajuste y se produzca un modelo de generalización limitada. MaxEnt logra lo anterior ya sea estableciendo un límite en el error o en la desviación máxima permitida de las características de los medios muestra (empírica). MaxEnt primero hace automáticamente una nueva escala de todas las características para obtener el rango 0-1. Después, se calcula un error ligado ($ʎ_j$ en la ecuación 3) por cada característica (de nuevo, nótese el cambio en "observación" de documentos anteriores, Tabla 2). El error ligado reflejará la variación en los valores de esa característica, ajustado por un parámetro calibrado para dicha clase de entidad (Phillips y Dudík, 2008; y ecuación 3). MaxEnt podría estimar errores ligados a la entidad solamente a partir de los datos, por ejemplo**,** utilizando validación cruzada. No obstante, para simplificar el ajuste del modelo y porque los datos están a menudo viciados, MaxEnt usa parámetros calibrados para cada clase de identidad en específico, basados en una amplia base de datos internacional (Phillips y Dudík, 2008). Ese conjunto de datos cubre 226 especies, 6 regiones del mundo, contiene muestras de un tamaño que va de 2 a 5822 y de 11-13 predictores por región (Elith *et al.*, 2016). Es posible que el ajuste pueda no funcionar

adecuadamente en una gran cantidad de bases de datos distintas, por ejemplo, si existen muchos indicadores. El usuario puede cambiar los parámetros calibrados si así lo desea. El pre ajuste también incluye limitantes al conjunto de clases de identidad que serán consideradas para muestreos pequeños.

$$\lambda_j = \lambda\sqrt{\frac{s^2[h_j]}{m}} \qquad (3)$$

donde λ_j es el parámetro de regularización para la característica h_j. Esta variación de la característica es $s^2[h_j]$ sobre m sitios de presencia y su clase de entidad tiene un parámetro calibrado λ. Conceptualmente, λ_j corresponde a la amplitud de los intervalos de confianza y, por lo tanto, toma la forma del error típico (la expresión de la raíz cuadrada) multiplicado por el parámetro λ, de acuerdo con el nivel de confianza deseado.

Las lambdas en la ecuación 3 permiten la regularización, es decir mediante el ajuste de la distribución, lo que permite hacerla regular. Estos errores ligados son una forma específica de regularización llamada regularización- L1 (Tibshirani, 1996) que proporciona soluciones escasas (algunas con muchos ceros, por ejemplo, muchas características removidas). La regularización no es exclusiva de MaxEnt; es un método común y moderno para la selección de modelos. Se podría pensar como una manera de contraer los coeficientes (los betas), tal como penalizarlos, a valores que balanceen forma y complejidad, permitiendo en ambos una predicción precisa y la generalidad de los mismos. En MaxEnt, la forma del modelo es medida en los lugares de incidencia, utilizando una probabilidad de registro (caja 1). Un modelo altamente complejo tendrá una probabilidad de registro alta, pero podría no generalizar bien. El propósito de la regularización es intercambiar la forma del modelo con la complejidad de este (el segundo término en la ecuación 4). En este sentido, MaxEnt establece un modelo de máxima probabilidad penalizado (Phillips y Dudík, 2008; ecuación 4), estrechamente relacionado con otras penalidades por complejidad tal como el criterio de información de Akaike (AIC por sus siglas en inglés,

1974). Maximizar la probabilidad penalizada de registro es equivalente a minimizar la entropía relativa sujeta a restricciones de error ligado.

$$\max_{\alpha,\beta} \frac{1}{m} \sum_{i=1}^{m} \ln(f(\mathbf{z}_i) e^{\eta(\mathbf{z}_i)}) - \sum_{j=1}^{n} \lambda_j |\beta_j| \tag{4}$$

sujeto a $\int_L f(\mathbf{z}) e^{\eta(\mathbf{z})} d\mathbf{z} = 1$

donde **z** es el vector de la característica para el lugar de incidencia i de m lugares,

y para j = 1 … n características.

El resultado logístico de MaxEnt

MaxEnt (de la versión 3 en adelante) proporciona un resultado logístico predeterminado. Es un intento por acercarnos lo más posible a un estimado de la probabilidad de que las especies estén presentes, dado un entorno, Pr(y = 1|z). Ésta es una post-transformación del resultado en bruto de MaxEnt que hace ciertas suposiciones acerca de la prevalencia y del esfuerzo de muestreo (caja 2 y apéndice S3). Estos dos tipos de resultados de MaxEnt (en bruto y logístico) están relacionados monotónicamente, por lo que, si el propósito de un estudio es clasificar los sitios de acuerdo con la idoneidad, no importa qué tipo se utiliza: ambos darán una clasificación idéntica y, por lo tanto, medidas basadas en rangos idénticos (por ejemplo, valores de AUC). La transformación logística de MaxEnt no es un procedimiento estadístico comúnmente usado, por lo que se explica a continuación el fondo y sus implicaciones.

De la ecuación 1, podemos ver que un enfoque simple Pr(y = 1|z) debería ser simplemente multiplicar $e^{n(z)}$ por una constante que estime la prevalencia; este enfoque tiene la desventaja de que $e^{n(z)}$ puede ser arbitrariamente largo, lo que implica que podríamos obtener un aproximado de Pr(y = 1|z) que exceda de 1 (Keating y Cherry, 2004; Ward, 2007b). Los modelos exponenciales pueden comportarse erróneamente cuando se aplican a nuevos datos, por ejemplo, cuando extrapolamos a nuevos entornos. Para evitar estos problemas y la no identificabilidad de

la prevalencia Pr(y = 1), el resultado logístico de MaxEnt, transforma el modelo de un modelo de familia exponencial (ecuación 2) a uno logístico:

$$\Pr(y = 1|\mathbf{z}) = \tau e^{\eta(\mathbf{z})-r}/(1-\tau+\tau e^{\eta(\mathbf{z})-r}) \quad (5)$$

donde $\eta(\mathbf{z})$ es el resultado linear de la ecuación 2, r es la entropía relativa del estimado de MaxEnt de $f_1(\mathbf{z})$ a partir de $f(\mathbf{z})$ y τ es la probabilidad de presencia en lugares con condiciones típicas para la especie (por ejemplo, donde $\eta(z)$ = el valor promedio de $\eta/\mathbf{z})$ bajo f_1). El valor predeterminado para s es arbitrariamente establecido en 0.5. La ecuación 5 es derivada utilizando un "minimax" o la inferencia bayesiana (detalles en el apéndice S3). En áreas inadecuadas, el denominador del resultado logístico es cercano a 1-s, por lo que el resultado es sencillamente una escala linear de resultado en bruto. Para áreas más adecuadas, el efecto del denominador es principalmente enlazar el resultado del modelo por debajo de 1. El resultado logístico con s= 0.5 empíricamente nos da un aproximado mejor calibrado de $\Pr(y = 1|\mathbf{z})$ que los valores en bruto sin transformar (Phillips y Dudík, 2008).

Debido a que la prevalencia de especies $\Pr(y = 1|z)$ no es identificable de entre los datos de incidencia, la prevalencia $\Pr(y = 1|z)$ implicada por el resultado logístico (con el valor predeterminado de s) no convergirá en la prevalencia verdadera, incluso si se tiene un gran número de datos de incidencia. Por otro lado, la prevalencia real depende de la definición de la variable de respuesta y, ésta por sí misma depende del método de muestreo, a menudo desconocido para los datos de presencia (ver preámbulo). Más tarde, si existe información adicional disponible que pudiera ser usada para la estimación de τ, la prevalencia podría ser identificada. Por lo tanto, ofrecemos orientación para la interpretación del resultado logístico de MaxEnt en relación al esfuerzo de muestreo y τ (caja 2).

Implicaciones de la modelización

Estas propiedades del modelo de MaxEnt tiene varias implicaciones sobre cómo debe ser usado.

MaxEnt se basa en una muestra no sesgada (al igual que todos los métodos de modelización de especies), por lo que los esfuerzos de recolectar un conjunto completo de

registros de presencia (limpiado para eliminar errores y duplicaciones) y los esfuerzos en lidiar con parcialidades son críticos (Newbold, 2010). Se implementan los métodos para lidiar con registros sesgados de especies (ver el estudio de caso 1 y Dudík *et al.*, 2006; Phillips *et al.*, 2009; Elith *et al.*, 2010). Las principales alternativas son brindar antecedentes con sesgos similares a aquellos en los datos de presencia (por ejemplo, al utilizar lugares ya estudiados por otras especies en el mismo grupo biológico) o utilizar una red de sesgo que indique los sesgos en los datos del estudio (ver el tutorial que proporciona MaxEnt a modo de ejemplo). Todos los valores en la red deberán ser positivos (o especificados como no datos) y deberán ser mostrados a escala para representar el esfuerzo relativo del estudio a través del paisaje L. Hay una consideración importante adicional. Si las redes covariables están desprotegidas (por ejemplo, latitud y longitud en grados, como por ejemplo información de la página web WorldClim http:// www.worldclim.org/), cualquier región que cubra un rango en latitud no trivial (digamos más de 200 km, especialmente que se encuentre lejos del ecuador) tendrá rejillas de red de área variante. Por ejemplo, en Australia, las celdas en el norte son aproximadamente 1.3 veces más el área de las celdas en el sur. MaxEnt saca muestras de las celdas aleatoriamente, asumiendo implícitamente áreas de celdas iguales. Las soluciones se dan para proyectar las redes a una proyección de un área igual, para crear una red mostrando las variaciones en el área de las celdas que pueden ser usadas luego como red sesgada, o para crear tu propia muestra de fondo con pesos de muestreo apropiado (estudio de caso 1).

Caja 1 Probabilidad de registro

En estadística, una probabilidad de registro describe el registro de la probabilidad de un resultado observado. Varía de 0 [ln(1)] a infinito negativo [ln(0)]. Si el espacio de resultados es continuo, medimos la densidad de probabilidad en el resultado observado en lugar de medir la probabilidad. Con datos de presencia los únicos resultados conocidos son los de presencia, por lo que cuando medimos probabilidades, el cálculo es hecho únicamente en lugares de incidencia (comparado con la regresión logística donde son calculados en lugares de presencia y de ausencia). Para un conjunto de observaciones, se realiza la estimación de la probabilidad del registro promedio. Cuando ajustamos un modelo MaxEnt de la interface del software, aparece un cuadro de diálogo que reporta el ajuste en el registro promedio que ha sido penalizado, comparado con un modelo nulo.

Caja 2 El estudio de caso del jaguar: reconciliando resultado logístico con el esfuerzo de muestreo

El jaguar (Panthera onca) y el Pecarí (Pecari tajacu) tienen una gran distribución similar en América del Sur y Centroamérica, por lo que los modelos de MaxEnt de ambas especies serían similares utilizando el valor predeterminados. Sin embargo, el jaguar es una especie mucho más rara que el pecarí, ¿cómo podrían compararse los resultados? La respuesta es que la probabilidad de la presencia es descrita solamente en relación a una definición dada de presencia/ausencia (por ejemplo, la escala temporal y espacial de una muestra: ver Preámbulo). Por ejemplo, para una especie rara como el jaguar un registro de presencia es común que derive de un muestreo por un tiempo prolongado y/o de un área más amplia (por ejemplo, utilizando trampas fotográficas por varios meses) y no como en el caso del pecarí que es más común y fácil de observar. Dado que los datos de presencia no muestran usualmente información sobre el esfuerzo de la muestra, esta elasticidad por definición es ampliamente conceptual y explica como pensar acerca del significado de las probabilidades en las especies. Cuando s es 0.5, los lugares típicos de presencia tendrán un resultado logístico cerca de 0.5. Esto es razonable siempre que podamos interpretar el resultado logístico como correspondiente a una escala temporal y espacial del muestreo que resulte en un 50% de probabilidad de que las especies estén presentes en áreas adecuadas. Ver apéndice S3 para más información.

Alternativamente, si el valor de s está disponible para un cierto nivel de esfuerzo de muestreo, podría usarse en lugar de los valores predeterminados y entonces las predicciones para las dos especies serían directamente comparables. Tau mide una forma de peculiaridad (Rabinowitz *et al.*, 1986). El jaguar tiene muy poca abundancia local inclusive en áreas adecuadas dentro de su rango, por lo que un valor muy pequeño de s es apropiado para todo menos para los esquemas más intensivos de muestreo. El aproximado de s podría provenir de conocimiento especializado o de encuestas específicas. Mientras s es determinado por la prevalencia, y viceversa, s es presuntamente más intuitivo ecológicamente hablando, ya que es una propiedad característica de las especies mientras que la prevalencia depende, en gran medida, del área de estudio elegida.

La solución de MaxEnt es afectada por el paisaje (la región) que se usa en el muestreo de fondo, tal y como lo demuestra VanDerwal *et al.,* (2009). Conceptualmente, ese terreno deberá incluir el rango ambiental total de las especies y excluir áreas que definitivamente no hayan sido investigadas (a menos que la razón por la que no se ha investigado sea que se está seguro de que

la especie no tiene incidencia ahí). Una especia endémica local, por ejemplo, propensa a ser restringida geográficamente debido a barreras de dispersión, debería ser modelada usando información de áreas en las que se haya posiblemente dispersado antes. Excluiremos las áreas despejadas que no serán estudiadas porque no exista hábitat para las especies. Se usarán marcadores para excluir las áreas del ambiente de muestreo, tal y como se explica en el tutorial en línea para MaxEnt (y ver la tabla 2). No obstante, es posible hacer predicciones de las áreas excluidas, utilizando herramientas para hacer proyecciones. Discutiremos algunas advertencias sobre los conceptos generales de la selección de antecedentes en el primer estudio de caso.

MaxEnt incluye una gama de tipos de entidades, y subconjuntos de éstos pueden ser utilizados para simplificar la solución. Por defecto, el programa restringe el modelo a entidades simples cuando pocas muestras están disponibles (siempre se usa el tipo linear; la cuadrática se usa con por lo menos 10 muestras; la de bisagra con por lo menos 15; la de umbral y la de producto con por lo menos 80) porque, tal como con cualquier método de modelado, pocas muestras proveen información limitada para determinar las relaciones entre las especies y su medio ambiente (Barry & Elith, 2006; Pearson *et al.,* 2007). En tales casos, también es buena idea primero reducir el conjunto de indicador candidato utilizando entendimiento ecológico de la especie (Elith & Leathwick, 2009b). Las entidades de bisagra tienden a volver redundantes a las entidades lineares y de umbral, y una forma de crear un modelo con funciones establecidas relativamente uniformes, más parecido a GAM, es utilizar solamente entidades de bisagra (ejemplo, Elith *et al.,* 2010 y el estudio de caso 1). Excluir entidades de producto crea un modelo aditivo que es más fácil de interpretar, aunque es menos capaz de modelar interacciones complejas.

MaxEnt tiene un método inherente para la regularización (Regularización- L1) que es confiable y conocido por desempeñarse bien (Hastie *et al.,* 2009). Éste lidia implícitamente con la selección de atributos (relegando algunos coeficientes a cero), es poco probable que se mejore y más probable que se degrade, por procedimientos que usan otros métodos de modelado para pre seleccionar variables (por ejemplo, Wollan *et al.,* 2008). En particular, este método es más estable en la fase de variables correlacionadas que en la regresión escalonada, por lo que es menos necesario remover variables correlacionadas (a menos que algunas de ellas sean conocidas por ser irrelevantes ecológicamente), o covariables pre procesadas utilizando PCA y seleccionando algunos ejes dominantes. No obstante, es importante resaltar que muy a menudo

existen otras variables, una muy buena idea es la pre selección especializada de un conjunto de variables candidatas (Elith & Leathwick, 2009b). Seleccionar variables proximales es propenso a ser particularmente importante cuando se van a utilizar modelos en diferentes regiones o climas. Si se requieren modelos más afinados, los parámetros de regularización pueden ser incrementados por el usuario (por ejemplo, ver Elith *et al.,* 2010).

Si se comparan modelos para diferentes especies se necesita cierta atención en el uso de las producciones logísticas porque la probabilidad de presencia es solamente definida en relación a un nivel de esfuerzo de muestreo dado, el cual por defecto se asume que es uno que resulta en un 50% de probabilidad de observar las especies en áreas adecuadas (caja 2). El esfuerzo que requiere el muestreo, por ende, depende de la especie. Esto presenta algunos desafíos para las comparaciones entre especies de áreas habitables, mismos que son resultado directo de usar datos de presencia y no son problemas exclusivamente de MaxEnt. Algunos usuarios podrían de hecho ver la escala de especies específicas como una oportunidad, ya que la literatura en cuanto a funciones favorables (por ejemplo, Real *et al.*,2006) afirma que es difícil trabajar con la probabilidad de presencia.

UTILIZANDO MAXENT

Estudio de Caso 1: Modelamiento de distribuciones actuales y futuras de una planta

Este análisis predice la distribución actual de Banksia prionotes, después utiliza el modelo para identificar donde son propensos a ocurrir los ambientes adecuados para la especie en condiciones de cambio climático. En esto, subrayamos la importancia de la selección del terreno y de lidiar con el sesgo muestral, eliminando el sesgo de muestras de antecedentes de redes de covariables no proyectadas, empleando el uso de un conjunto reducido de tipos de entidades para un modelo más afinado y herramientas para calificar los ambientes en nuevos tiempos y lugares.

Una *Banksia prionotes* puede ser desde un arbusto leñoso hasta un árbol pequeño nativo del suroeste de Australia Occidental (WA por sus siglas en inglés). Se distribuye ampliamente a lo largo de esa área y muestra una preferencia por suelos arenosos profundos. Comúnmente es

una planta dominante en maleza y en bosques bajos, es una fuente importante de néctar para los melifágidos y una especie sobre saliente de flores ornamentales.

Métodos

Aquí, utilizamos datos de especie del atlas de Banksia (Taylor & Hopper, 1988; Yates *et al.,* 2010), con 361 registros para B. Prionotes de los 4631 lugares a través del suroeste de la región florística de Australia (SWAFR, por sus siglas en inglés) que fueron estudiados para Banksia y para los cuales tenemos datos ambientales completos. El atlas es el resultado de un proyecto comunitario de ciencias y los registros pueden ya sea ser interpretados como de presencia o de presencia-ausencia, dependiendo de las suposiciones que se hagan acerca de la búsqueda de patrones de contribuidores. Aquí se les trata como datos de presencia, pero se usa el conjunto completo de locaciones como un tratamiento "de fondo". Para demostrar el efecto de esta decisión, fueron evaluados dos fondos de alternativas (por ejemplo, definiciones del terreno): una muestra de 10000 lugares dentro del SWAFR (Yates *et al.*, 2010; y figura 2) y una muestra de 20000 lugares por toda Australia. El gran número de lugares en Australia fue utilizado para asegurar una buena representación de todos los ambientes y estuvo basado en pruebas anteriores sobre los efectos del tamaño del fondo de muestreo en la estructura de modelo para estos predictores (J. Elith, datos no registrados). Debido a que los datos covariables para este estudio no están proyectados, estas muestras fueron pesadas aleatoriamente de acuerdo con el área de la celda (ver métodos en Apéndice S4).

Utilizar lugares aleatoriamente dentro de la región florística implica que los registros de presencia son una muestra aleatoria de todas las ubicaciones donde la especie está presente en la región, lo cual es poco probable porque los registros fueron tomados de los parches existentes de la vegetación en entornos probablemente apropiados (la región ha sido extensamente despejada para la agricultura y algunas de las áreas más interiores son demasiado áridas para muchas especies de Banksia). Utilizar lugares aleatorios a lo largo de Australia implica que la especie pudo haberse dispersado a cualquier lugar a lo largo del continente y, por ello, implica que todo el continente sea considerado disponible para el muestreo. Esto es cuestionable porque las áreas

desérticas hacia el norte y este del área inhabitada son vistas como barreras para la dispersión. Volveremos a las implicaciones más tarde.

Yates *et al.,* (2010) identifica importantes factores climáticos para las plantas del suroeste de Australia Occidental. Basamos nuestro conjunto candidato de predictores en su estudio, pero usamos una fuente de datos diferente para que podamos entrenar y predecir a través de toda Australia. Tal como se describe en el Apéndice S4, nuestras covariables (todas sin ser proyectadas, en 0.01 grados o aproximadamente 1km de resolución de red) incluyen cinco variables climáticas: la isotermalidad (ISOTHERM por sus siglas en inglés), que significa la temperatura media del trimestre más húmedo (TEMPWETQ, por sus siglas en inglés), la temperatura media del trimestre más cálido (TEMPWARMQ), la precipitación anual (RAIN, por sus siglas en inglés), la precipitación del trimestre más seco (RAINDRYQ) y un aproximado de la capacidad de retención de agua disponible en la planta solum (SOLWHC por sus siglas en inglés). Presentamos esto solamente como un estudio de demostración y reconocemos que para la aplicación rigurosa en esta región y para obtener predicciones precisas, son necesarios datos de suelos de mejor calidad además de predictores que representen transformaciones en la tierra (Yates *et al.,* 2010). El ambiente futuro fue representado por cambios predichos bajo el escenario A1FI para estimar 2070 sobre el conjunto de 23 GCMs en IPCC AR4 (Solomon *et al.,* 2007); el SOLWHC se asumió para permanecer tal y como está ahora.

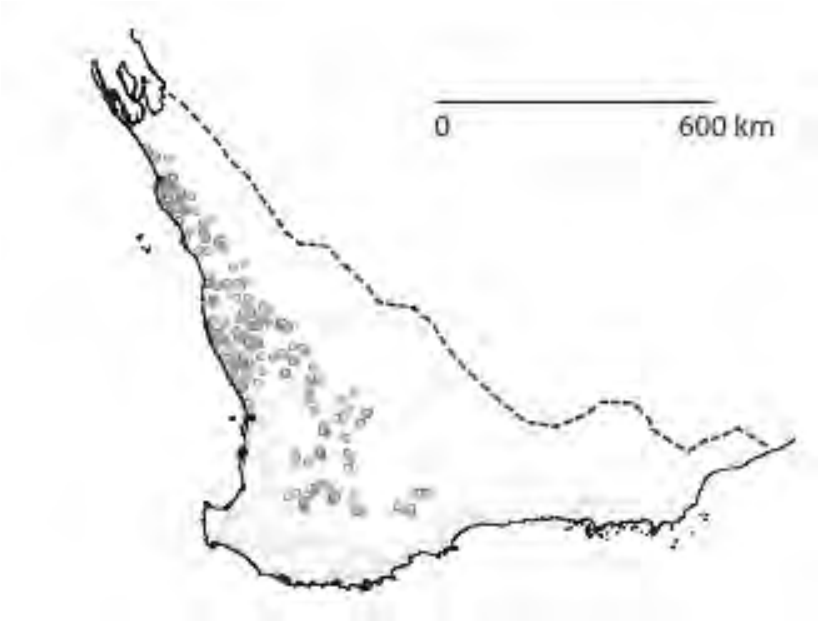

Figura 2. Se señalan todos los lugares del atlas Banksia (en gris) con incidencia de Banksia

prionotes en círculos grises.

Los modelos fueron ajustados y proyectados a climas actuales y futuros (Figura 3) usando solamente características de bisagra, con parámetros de regularización por defecto (véase el Apéndice S5 para los detalles del modelo y para una comparación con modelos equipados con todos los tipos de entidades). Ajustamos todos los modelos en los conjuntos de datos completos pero también utilizamos 10 veces la validación cruzada para estimar los errores de las funciones ajustadas y el rendimiento predictivo de los datos presentados. Esta última es una buena prueba para cada modelo pero, dados los diferentes terrenos, no funciona para comparar a través de los modelos. Es necesario resaltar que la AUC (por sus siglas en inglés) en este caso se calcula en presencia versus datos de fondo (Phillips *et al*., 2006). Para comparar los modelos en base a datos consistentes, podemos dividir los datos del atlas en conjuntos de entrenamiento y de pruebas para una validación manual cruzada de 5 veces, probando cada modelo en base a datos idénticos, retenidos a través de dos estadísticos de prueba (área bajo la curva característica de funcionamiento del receptor (AUC por sus siglas en inglés), y correlación, COR por sus siglas en inglés; detalles en el Apéndice S4). El código de ejemplo para realizar dichos análisis se encuentra disponible en línea (Apéndice S4).

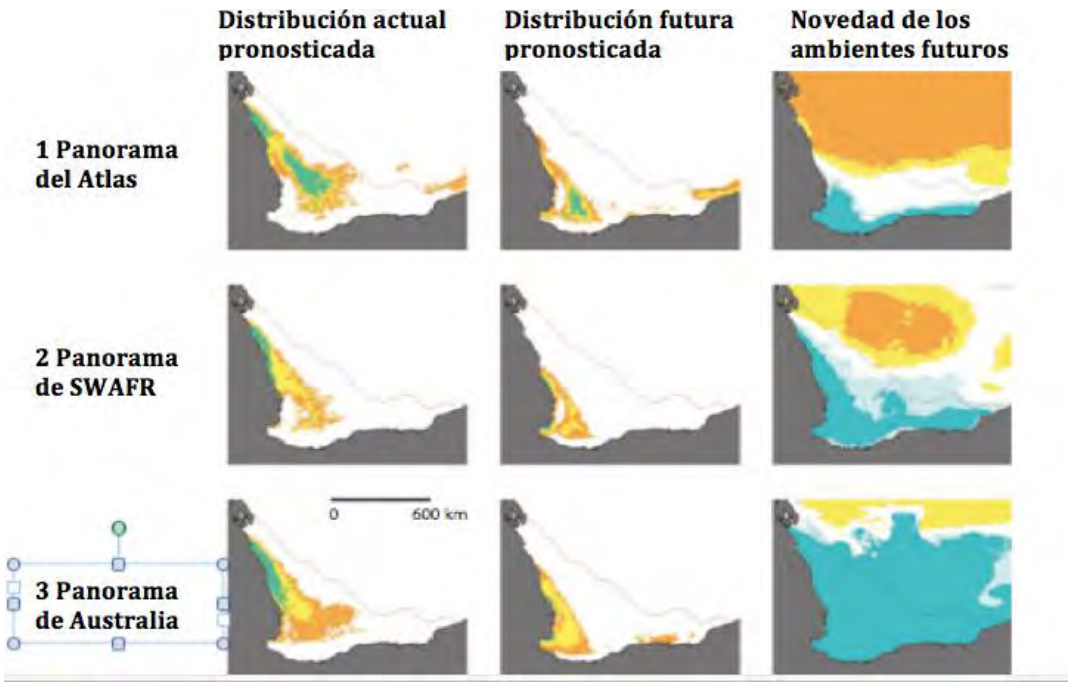

Figura 3. Los resultados del modelo para el estudio de caso 1, mostrados por los tres conjuntos de datos (en filas): distribuciones actuales y futuras pronosticadas, y grado de extrapolación en comparación con los datos de formación. Las distribuciones pronosticadas son resultados logísticos, desde valores bajos (blanco, 0-0.2) hasta naranja, amarillo, verde y azul (0.8-1.0). Para mapas de extrapolación, colores templados indican que la extrapolación está ocurriendo, el color naranja indica lo más extremo. El color gris indica al océano.

Resultados

El panorama del atlas (modelo 1) produce una distribución mapeada en la región habitada con más énfasis en el oeste, comparado con otros tratamientos de panoramas (figura 3). El sesgo del oeste en la distribución de los lugares estudiados (figura 2) afecta las distribuciones pronosticadas por el modelo 2 y 3 (panorama aleatorio a través de SWAFR o Australia) pero fue tomado en cuenta al usar el panorama del atlas (modelo 1). La distribución del este es más consistente con la ecología conocida de la especie y con la distribución observada (Taylor y Hopper, 1988). La importancia de la variable varía con el conjunto de datos, con TEMPWETQ siendo mucho más sobresaliente al utilizar un panorama de toda Australia que al utilizar solamente la parte del suroeste. De la misma manera, figuras de las funciones ajustadas varían a través de los conjuntos de datos (Apéndice S5). Esto es de esperarse porque cada conjunto de datos implica una cuestión de modelado diferente (por ejemplo, el panorama de toda Australia hace que uno se pregunte: ¿por qué esta especie se encuentra solo en ambientes del suroeste?).

Un número creciente de aplicaciones de SDM implica pronósticos de nuevos panoramas (por ejemplo, de nuevos lugares o tiempos; Elith y Leathwick, 2009a). Estas son aplicaciones contenciosas que hacen pronósticos fuertes (Dormann, 2007) y usualmente requieren de un pronóstico de entornos no muestreados por los datos de entrenamiento. MaxEnt se ha ampliado para incluir nuevas capacidades para informar a los usuarios sobre la predicción de nuevos ambientes (Elith *et al.*, 2010). MaxEnt ya proporciona información mapeada sobre el efecto del "clamping" del modelo, es decir, el proceso por el cual las entidades están limitadas para permanecer dentro del rango de valores en los datos de entrenamiento. Esto identifica las ubicaciones donde los pronósticos son inciertos debido al método de extrapolación, al mostrar

donde afecta el clamping sustancialmente al valor pronosticado. Consideramos que se debe tener cuidado extremo cada que se extrapole fuera del entrenamiento, para que cálculos nuevos ("mapas MESS", por ejemplo, interfaces de similitud ambiental multivariable) muestren diferencias entre el panorama de entrenamiento y de pronóstico (figura 3). En este caso, muestran que, al comparar con ambientes de lugares del atlas, las partes del norte de SWAFR experimentarán nuevos climas en el 2070 (Figura 3 modelo 1). Los modelos basados en panoramas aleatorios a lo largo de SWAFR o del continente (modelo 2 y 3) requieren menos extrapolación (ya que una amplia cantidad de muestreo de sitios de panorama trae consigo un muestreo más amplio de ambientes) pero, dados los problemas con el realismo de estos tratamientos, no vemos el resultado como una ventaja necesaria para futuros pronósticos.

Los apéndices S5 y S6 incluyen más información sobre cómo pronostican estos modelos en todo el continente, tanto para los climas actuales como para los futuros. Proporcionan ideas interesantes sobre la variación del modelo a través de escalas, regiones y conjuntos de datos, y destacan la importancia de la elección del panorama (véase el comentario, Apéndice S5). En particular, es interesante que el modelo 3 limite los pronósticos del área general correcta, tenga la validación cruzada AUC más alta de 10 veces más (tabla 3) y, sin embargo, obtenga la justificación ecológica más pobre para la elección del fondo y sea menos probable su utilidad para manejar las especies localmente. La ventaja de limitar el panorama en áreas locales y alcanzables (modelo 1 y 2) es que el enfoque del modelo es contrastar panoramas habitados e inhabitados del área local y, particularmente, empleando datos del panorama a una escala fina, se puede lograr una diferenciación útil en la escala del manejo. Es también muy probable que sea la opción más realista ecológicamente para una gran cantidad de especies restringidas a lo local. Por otro lado, si los modelos deben ser proyectados muy fuera del área geográfica local, el uso de terrenos locales traerá consigo la penalidad de que el pronóstico de otras áreas es propenso a provocar una extrapolación considerable. Claramente se requiere alguna compensación.

	Importancia de la variable							
Modelo (Panorama)	RAIN DRYQ	RAIN	TEMP-WARMQ	TEMP-WETQ	ISO-THERM	SOL-PWHC	AUC (10 veces más CV pero varían los conjuntos de datos)	AUC; COR (5 Veces más datos de atlas)
1 (ATLAS)	57.9	30.7	7.9	0.4	1.1	2.0	0.92	0.96; 0.62
2. (SUROESTE)	45.3	35.4	4.7	3.4	9.9	1.4	0.90	0.93:0.52
3 (AUSTRALIA)	19.7	17.7	5.3	54.0	3.0	0.3	0.99	0.91: 0.45

Tabla 3 La importancia de la variable estadística de la evaluación para el estudio de caso 1. Los nombres de las variables y las abreviaciones para la estadística de la evaluación son consistentes con el texto.

Estudio de caso 2: Modelamiento de las distribuciones de peces en ríos

Este análisis predice la distribución actual de la especie, Gadopsis bispinosus, el pez negro de doble espina, en ríos del sureste de Australia. En el preámbulo, presentamos un caso en el que con presencia y datos del fondo podemos modelar la misma cantidad que se podría modelar con datos de presencia- ausencia, hasta la constante Pr (y= 1). Una implicación de esto es que debemos ser capaces de usar los mismos tipos de datos, incluyendo relaciones de modelos ecológicos a pequeña escala con información detallada – por ejemplo: necesitamos no limitarnos solo a celdas de cuadrillas gruesas y a variables básicas del clima. Aquí, utilizamos información ecológica detallada en la escala del segmento del río para modelar la distribución de la especie del pez nativo. Hasta donde tenemos conocimiento, este es el primer ejemplo usando MaxEnt con datos de vectores (del segmento del río).

Gadopsis bispinosus es un pez nativo de agua dulce endémico del sureste de Australia. Se reproduce en el altiplano frío y despejado o en arroyos montanos con abundantes corrientes. Es más común en arroyos de medianos a grandes que están lo suficientemente profundos para reducir las velocidades de la corriente y en cuencas forestadas con insumos sedimentarios relativamente pequeños (Lintermans, 2000).

Métodos

Los datos de especies surgen de encuestas (descritas más adelante en el Apéndice S7) de los drenajes internos de ríos en el noroeste de Victoria, en Australia. En esta área, hay diez sistemas principales de ríos, agrupados en cuatro regiones que inician en terrenos de colinas o montañosos y que drenan con dirección al norte. G. Bispinosus fue registrado en 255 sitios. Utilizamos datos

covariables de la captura de los 255 sitios como nuestra muestra de L_1, y una muestra al azar de 10,000 de 240,000 segmentos de ríos para nuestra muestra de L, como datos de respaldo.

El conjunto de predictores de candidatos comprendía 20 variables que resumían la información a través de tres escalas espaciales jerárquicamente anidadas (segmento, cuenca inmediata y toda un área del río cuesta arriba) y también de río abajo hacia el gran sistema fluvial que drena al océano. Las variables ambientales estiman clima, pendiente del río, vegetación ribereña y características de la cuenca (ver Apéndice S7). El sistema fluvial también fue incluido para cuantificar la variación espacial en las características de la tierra y las alteraciones que no hayan sido cubiertas por el conjunto de candidatos predictores.

Estos datos basados en segmentos (sin rejillas) son modelados usando el formato SWD (muestras con datos, por sus siglas en inglés) de MaxEnt- esto implica presentar resúmenes de tipo de hoja de cálculo de los entornos, tanto en sitios de presencia como en sitios de fondo**.** Todas las variables ambientales fueron continuas excepto la covariable categórica del sistema fluvial. Se usaron configuraciones por defecto para las entidades y para la regularización del modelo de práctica y una validación cruzada de 10 veces para obtener estimaciones fuera de la muestra del rendimiento predictivo y estimaciones de la incertidumbre alrededor de las funciones ajustadas. Para mapear, el modelo fue proyectado a un área seleccionada en la cuenca de Goulburn- Broken. Técnicamente, esto se logró al proyectar en el formato de datos SWD, luego relacionando las predicciones con los segmentos relevantes del río en un sistema de información geográfica (SIG). El apéndice S8 incluye datos y código para replicar este estudio de caso, incluyendo información de cómo ejecutar MaxEnt desde archivos por lotes.

Resultados

Consistente con el conocimiento ecológico acerca de las especies, el modelo predice que G. Bispinosus ocurrirá más frecuentemente en ríos más grandes o en áreas montañosas (Figura 4). Estas ubicaciones están identificadas en aquellas cuyas elevadas cuencas de agua tengan relativamente más precipitación en el período más cálido y en las cuestas máximas más inclinadas. Entre ellos, el énfasis en los segmentos con las temperaturas máximas más cálidas de verano sirvió para excluir las corrientes frías de mayor elevación (Figura 5). Las pruebas de

Jacknife de importancia de la variable ayudan a identificar aquellas con efectos individuales importantes; los tres predictores individuales más importantes fueron la longitud sumada de todos los enlaces ascendentes (TOTLENGTH_UCA), la pendiente máxima de aguas elevadas (US_MAXSLOPE) y la cantidad de cobertura de ribera aguas elevadas (UC_RIP_TRECOV); y el predictor con la mayor información que no estuvo presente en las otras variables es el de la temperatura máxima, basada en el segmento del mes más cálido (MAXWARMP_TEMP). Muchos predictores tuvieron de pequeños a mínimos impactos en el modelo final. El modelo muestra una fuerte discriminación en los datos presentados, con una validez cruzada AUC de 0.97.

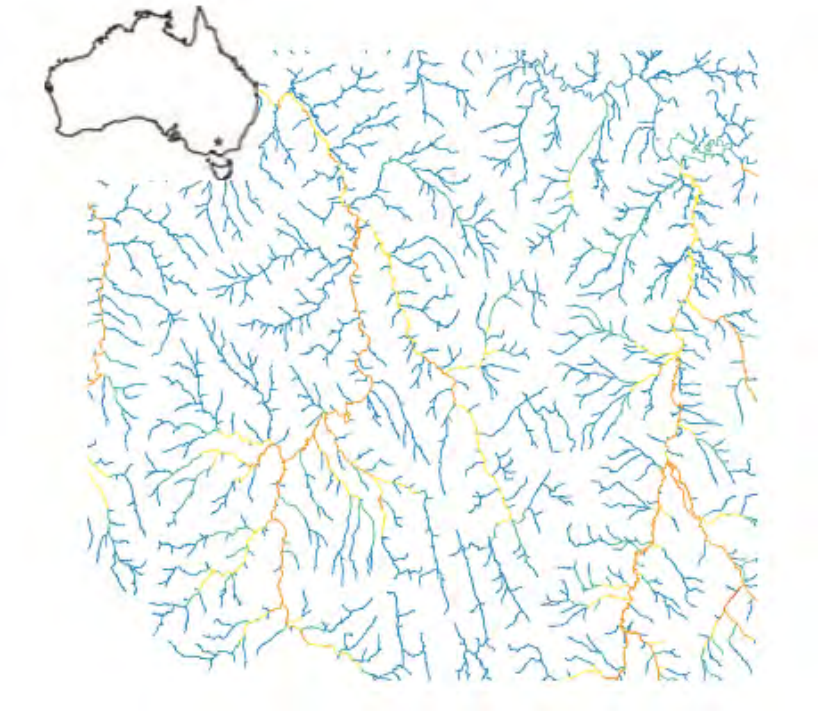

Figura 4. Distribución predicha de Gadopsis bispinosus que muestra predicciones logísticas de resultados de MaxEnt. Leyenda: predicciones en intervalos equivalentes desde 0 hasta 1, de azul (bajo) a verde- Amarillo- naranja (alto). Escala: de este a oeste, el mapa de los ríos abarca 45 km. La estrella en el recuadro muestra la ubicación.

Extensiones/alternativas

Dado que los registros de un sistema fluvial podrían compartir un entorno más similar al de los sistemas diferentes, un enfoque alternativo al de la validación cruzada sería probar las predicciones de forma iterativa en ríos retenidos. Nosotros escogimos no hacerlo en este caso porque los registros de presencia estuvieron concentrados en relativamente pocos sistemas fluviales para que los conjuntos de práctica fueran substancialmente reducidos y los conjuntos de pruebas relativamente pocos.

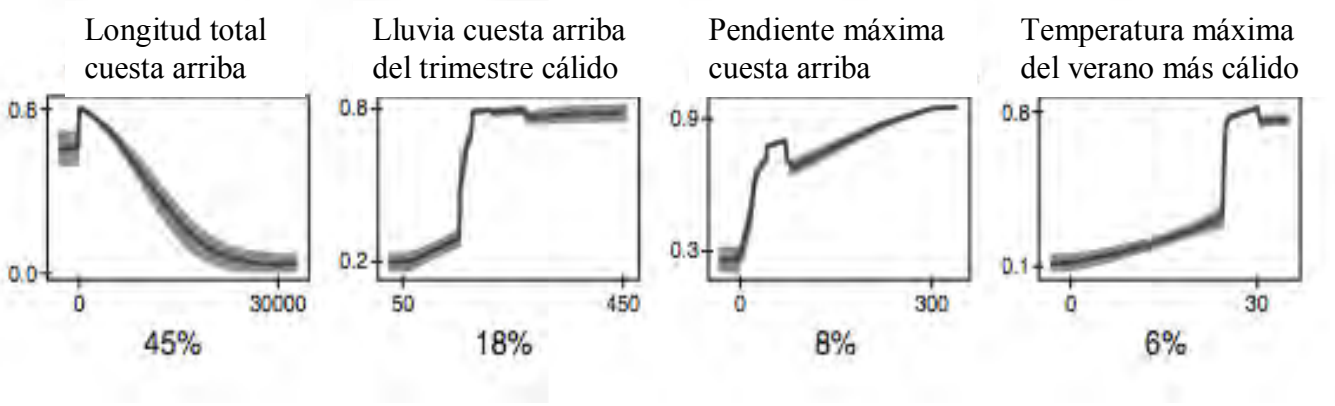

Figura 5. Parcelas parciales de dependencia que muestran la respuesta marginal de Gadopsis bispinosus a las cuatro variables más importantes (por ejemplo: para valores constantes de las otras variables), con importancia variable debajo de cada gráfica. Los ejes e indican los resultados logísticos.

CONCLUSIONES

Hemos descrito MaxEnt desde un punto de vista estadístico, mostrando que el modelo minimiza la entropía relativa entre dos densidades de probabilidad definidas en un espacio de funciones. Una comprensión del modelo conduce naturalmente a recomendaciones para la implementación y las nuestras incluyen la importancia de brindar muestras de fondo apropiadas para lidiar con los sesgos muestrales y ajustar el modelo- a través de la selección del tipo de entidad y de ajustes de

regularización- para adaptarse a los datos y a la aplicación. Los datos de presencia son un recurso valioso y potencialmente puede ser usado para modelar las mismas relaciones ecológicas como con datos de presencia-ausencia, siempre que se puedan abordar los sesgos y la no identificabilidad de la prevalencia. MaxEnt es actualizado regularmente, usualmente para incluir nuevas capacidades para adaptarse a las aplicaciones en expansión y también algunas veces para cambiar el programa predeterminado para aquellos mayormente usados en la práctica. Nuevas capacidades recientes incluyen la validación cruzada y mapas MESS (por ejemplo, estima como se compara el espacio ambiental en tiempos y lugares predichos con aquellos de los datos de la práctica) demostrados en el estudio de caso 1. Además, nuevos mapas seleccionables permiten a los usuarios interrogar predicciones espacialmente, brindando información para cualquier rejilla en los componentes de la predicción (por ejemplo: lo que contribuye a su valor particular) y donde se asientan las condiciones ambientales en las funciones ajustadas. Mapas de factores limitantes muestran la variable que influencia más a la predicción en cada rejilla (apéndice S6). Para más detalles, véase Elith *et al.* (2010) y el más reciente tutorial en línea (http:// www.cs.princeton.edu/~schapire/maxent/). SDMs pueden proporcionar información útil para explorar y predecir distribuciones de especies y estamos deseosos de ver su continuo desarrollo y uso para aprender y conservar la biodiversidad del mundo.

CHAPTER III
TRANSLATION ANALYSIS

As it was mentioned in the first chapter, the analysis of this translation has its foundation on Vinay and Darbelnet's direct and oblique translation techniques. Although the Canadian Approach is the chore of this analysis, due to the difficulty of this paper, there was a need of using some other techniques from the approach of Malone. For the purpose of this section, I present the analysis of the different techniques from the Canadian Approach used for the translation of the article chosen by first giving a brief insight of what each of the techniques is about, then examples taken from my translation are presented. After the brief explanation, it is explained why the word, sentence or phrase was translated the way it was by using some theoretical support.

The examples are presented divided in two sections as a way of contrasting: one containing the source part of the text, the other presents the Spanish translation. The explanation on how the technique was applied in that specific part of the text is given below the examples.

Due to the scientific nature of the article, it was notorious throughout this translation that the application of the oblique techniques was scarcer than the one from the direct techniques. A comment about this situation is presented at the end of this chapter.

3.1 Borrowing

In the borrowing technique, the source-language word is translated in the target language as it is written. The reason is that the target language has a gap in its lexicon regarding this word. Another reason to borrow a word is to create a stylistic effect on the text. If the intention of the translator is to demonstrate specificity on some terms, borrowing is a good idea even if there is an existent translation for the term. The advantage of the borrowing technique is that it brings an original connotation to the word in TL (Fawcett, 2003).

The following are some examples of the borrowing technique found in the translation of "A statistical explanation of MaxEnt for ecologists". Then a brief explanation of the application of this technique is given.

Source Text	Target Language
It is a log-linear model, similar in form to a GLM…	Es un modelo log-lineal, similar en forma a un GLM…

In this example, we used a borrowing of the word log from the English language. There is no word in Spanish with the same or equivalent meaning. The morphological structure of the compound word is also a borrowing from the English morphology. As we can see, the word order for the noun and adjective is the one used in the source language: adjective + noun. In Spanish, the predominant and the most natural word order for this situation is noun + adjective, meaning that the adjective position is next to the noun and not before it (Real Academia Española, 2010). The specialist who reviewed the translation confirmed the most used form for this word in Spanish when talking about this topic was log-lineal.

Source Text	Target Language
The lambdas in equation 3 allow regularization…	Las lambdas en la ecuación 3 permiten la regularización…

According to the Dictionary of the Spanish Language (2017) the word lambda was borrowed from the Greek alphabet to many other languages. Being said that, "the lambdas" is a borrowing from the Greek alphabet to the source language of the article and finally, the word was borrowed to the target language of the translation.

Source Text	Target Language

Equation 5 is derived using a "minimax" or …	La ecuación 5 es derivada utilizando un "minimax" o …

According to the encyclopedia of mathematics (2014), a minimax can be interpreted (for example, in decision theory, operations research or statistics) as the least of the losses that cannot be prevented by decision making under the given circumstances. There is no equivalence for this word in the target language of this translation. Some articles in Spanish use the same borrowing from the English language, which means that the term is understood by Spanish readers even though it is not written in Spanish. The term is mostly used in game theory but the meaning is also perfectly applied in statistics.

3.2 Calque

When applying this translation technique, which is similar to the borrowing one, it can be appreciated a literal translation at a phrase and word level and it is mostly applied in philosophical and scientific texts (Fawcett, 2003).

These are some examples of the calque technique taken from the translation of the chosen article.

Source Text	Target Language
Species distribution models (SDMs) estimate the relationship between…	Los modelos de distribución de especies (SDMs por sus siglas en inglés " Species Distribution Models) calculan la relación entre …

A calque from the phrase "species distribution models" was used into the target language of this paper; in specialized texts new words and expressions are continuously created so to keep the essence of the message using a calque becomes an appropriate technique. This phrase refers to a specific model used in the program explained in this document. The phrase was translated word for word but the syntax slightly changed since

in Spanish nouns generally precede adjectives; this position of the adjective is more natural in most of the registers in Spanish (Real Academia Española, 2010).

Source Text	Target Language
Then not only does this result in some false absences in presence-absence data, it also affects the pattern of presences…	Entonces, no solo esto resulta en ausencias falsas en datos de presencia-ausencia sino que también afecta el patrón de presencias…

Calque works in this case because this phrase can be found in specialized and scientific texts related to Ecology and Geology. This is an example of a lexical calque as it follows almost the exact word order. Thus this technique was chosen to be applied in this case since it is quietly used when translating scientific and specialized texts (Fawcett, 2003). We found the same translation being used in specialized text in Spanish.

There is a slight change in the syntax of the phrase as the three components of the phrase are nouns and that type of structure is not a common structure in Spanish. In Spanish, nouns can be converted into adjectives by adding a preposition, in this case the Spanish preposition "de" made possible the adjectival conversion of the words presencia-ausencia (Alarcos, 1995)

Source Text	Target Language
Selected features are formed "behind the scenes", in the same way as in regression…	las características seleccionadas se forman "detrás de cámaras", de la misma forma que en regression…

The calque technique is used in this example to express what the speaker actually implies rather than what s/he writes; that is the reason why the word was replaced by one that would better convey the speaker's meaning. In Spanish the expression used to mean this idea is: detrás de cámaras. Thus, instead of translating scenes literally as escenas, it is translated as cámaras, which is totally related to the meaning of the word scenes and the implied context in both languages. This is an example of implicature, which is part of the calque technique (Baker, 1992).

Source Text	Target Language
The null model for the raw distribution was the uniform distribution over the landscape…	El modelo nulo para la distribución cruda fue la distribución uniforme sobre el paisaje…

The concept "raw distribution" is widely used in specialized statistics articles in English and Spanish. Nevertheless, it was necessary a slight variation in the syntax of the phrase because in the Spanish language as mentioned before the noun precedes the adjective (Real Academia Española, 2010).

3.3 Literal translation

Fawcett (2003, p.36) describes this translation technique as "when a text can go from one language to another with no changes other than those required by the target language grammar". A literal translation is reversible and complete in itself (Venuti, 2000).

These are a few examples of this technique.

Source Text	Target Language
we describe MaxEnt using statistical terminology and notation…	[…] describiremos MaxEnt utilizando terminología y notación…

we fit models and interpret them, exploring why certain choices affect the result and what this means.	[…] ajustamos modelos y los interpretamos explorando por qué ciertas decisiones afectan el resultado y lo que esto significa.
we assume that the data available to the modeller are…	[…] asumimos que los datos disponibles para el modelador son…
we can model the same quantity as with…	[…] podemos modelar la misma cantidad que con…
we believe it is generally advisable to use a	[…] creemos es generalmente aconsejable usar…

As it can be seen, these sentences and/or parts of sentences were literally translated with no changes of sentence order except for the omission of the pronoun. In Spanish the pronoun of a sentence is usually omitted since it is implied in the conjugation of the verb; the verb is the chore of the sentence in the Spanish language (Alarcos, 1995).

Source Text	Target Language
This analysis predicts the current distribution of Banksia prionotes…	Este análisis predice la distribución actual de Banksia prionotes…

The literal translation technique was used in this part of the text because in both languages theses sentences follow the same syntax: both of the sentences have a subject and a predicate, which are the main syntactic components of a sentence in Spanish (Larousse, 2005). In this case there is no need of word order rearrangement to keep the meaning of the original text.

3.4 Transposition

The transposition technique deals with grammatical changes when translating. It involves "replacing one word class with another without changing the meaning of the message… there are two types of it: obligatory transposition and optional transposition" (Venuti, 2000). The elements that can be transformed are: Words, phrases and parts of sentences (simple or compound).

The following are some examples of transposition found in the translation of "A statistical explanation of MaxEnt for ecologists".

Source Text	Target Language
Where species data have been collected systematically – for instance, in formal biological surveys in which a set of sites are surveyed and the presence/ absence or abundance of species at each site are recorded – regression methods familiar to most ecologists (e.g., generalized linear or additive models, GLMs or GAMs; or ensembles of regression trees: random forests or boosted regression trees, BRT) are used.	Se usan métodos de regresión con los cuales la mayoría de los ecologistas están familiarizados (ejemplo: modelos lineales generalizados o modelos aditivos generalizados **(**GLMs o GAMs por sus siglas en inglés**)**: o conjuntos de árboles de regresión: bosques aleatorios o árboles de regresión reforzados, BRT por sus siglas en inglés) en situaciones en las cuales los datos han sido recolectados sistemáticamente.

In this example the transposition technique was used as there was the need of rearrangement of the parts of the sentence. In the Spanish language, the general structure of a compound sentence has the subject mostly at the beginning of each sentence (Hualde et al, 2001), not at the end like in the English language. Another reason why the word order of the sentence changed is that in Spanish the reflexive form, unlike in English, is more frequently use instead of the passive voice. (Real Academia Española, 2010)

Source Text	Target Language

Most ecologists, following the statistical literature, call the independent variables in a model the covariates, predictors or inputs.	La mayoría de los ecologistas, en apego a la literatura estadística, llaman covariables, predictores o insumos a las variables independientes de un modelo.

The transposition technique is applied in this example to have a better stylistics in the target language but keeping the same meaning of the original message. There is a word class shift, we replaced the verb form by an adverbial phrase, what we call in Spanish "locución adverbial"; adverbial phrases are very common in use, especially in Spanish (Holecková, 2007).

Source Text	Target Language
Understand environmental correlates of species occurrences, groups of species, or other.	Comprensión de correlaciones ambientales de ocurrencia de especies, de grupos de especies u otros.

In this phrase a class shift was made by changing the word from being a verb into a noun. The decision of changing the verb "understand" into a noun was based on the fact that there is a need of conciseness when writing charts in academic Spanish texts which is meant by the noun (American Psychological Association, 2010). This shift does not affect the meaning of the intended message.

Source Text	Target Language
Similarly, if the detectability of a particular species varies from site to site	De manera similar, si la detectabilidad de una especie en particular varía de lugar a lugar…

In this example, there is a slight variation on the word form Similarly. Similarly is an adverb and it was changed into an adverbial phrase as in the target language the use of certain adverbial phrases is preferred from its regular adverbial form (Real Academia Española, 2010).

Source Text	Target Language
MaxEnt (from version 3 onwards) gives a logistic output as its default.	MaxEnt (de la versión 3 en adelante) proporciona un resultado logístico predeterminado.

In this example, the transposition technique is used by shifting the phrase "as its default" into a past participle that has the function of an adjective: predeterminado (Alarcos, 1995). By doing so, the resulting translation flows better than if we had used literal translation.

3.5 Modulation

The modulation technique can be put into practice by changing the point of view of perspective and even a change of thought in the text. Here is an example: Il n'a pas hesité - He acted at once' (Newmark, 1988). Modulation can be changed from abstract to concrete terms, between a part and a whole, from positive to negative, cause for effect, passive voice to active voice (Fawcett, 2003).

Modulation is a useful tool because in some cases even though transposition results are grammatically correct, they can be considered awkward.

These are examples of the modulation technique.

Source Text	Target Language

Some of the published discussion suggests that presence- only data in some sense release us from the problems of unreliable absence records	Parte del análisis publicado sugiere que los datos de presencia de especies de alguna manera no presentan los problemas de confiabilidad que generan los registros de ausencia de especies

This is a precise example of modulation, from a positive sentence to a negative sentence without changing the intended message. By doing so, we modified the form of the message but not its purpose. A re-stylization of the source message was done (Levy, 2011).

Source Text	Target Language
If comparing models for different species some care is needed in use of the logistic outputs because probability of presence is only defined …	Si se comparan modelos para diferentes especies se necesita cierta atención en el uso de las producciones logísticas porque la probabilidad de presencia es solamente definida…

In this example, the modulation technique is used by shifting from the passive voice into the reflexive passive voice. In Spanish, the reflexive passive voice with the use of "se" is more common than the passive voice and it is commonly use to talk about generalities (Real Academica Española, 2010). The phrase "care is needed in use of the logistic outputs…" can be considered as a generality in this specific process of use; therefore, the reflexive passive voice was used to translate it.

3.6 Equivalence & Adaptation

On the one hand, equivalence technique happens "when two languages refer to the same situation in totally different ways" (Fawcett, 2003, p. 38). It happens even with no formal or semantic correspondence. According to Ni (2009, p. 81) equivalence "this strategy describes the same situation by using completely different stylistic or structural methods for producing equivalent texts". This kind of strategy is usually used at the lexical level and it is commonly used for proverbs and idioms. Formal discourses like the one used in the article translated hardly used idioms or proverbs thus that may be the reason why we did not find the need to use any equivalence strategy when translating this article.

Another strategy that was not used when translating this article was adaptation. Adaptation takes place when the target culture has almost no similarities with the source culture or when the terms are unknown for the target culture context. It is used if the terms in the target language are not clear enough or if they prevent the target reader from comprehending the message. Here is an example given by Vinay and Darbelnet: cyclisme by either baseball or cricket because in the target culture there is no such thing called cyclisme (Fawcett, 2003). Adaptation can be described also as a situational equivalence (Venuti, 2001).

Neither equivalence nor adaptation examples were found throughout the analysis of this translation. When dealing with scientific or specialized texts it is a highlight the objectiveness, the concise, concrete and straightforward form is the best option when translating these types of texts. No emotional load is applied in specialized texts (Mastnà, 2010); therefore, the mentioned techniques were not used in this translation as they deal with a more emotional and contextual use rather than with linguistic features. In specialized texts such as the article translated in this paperwork there is no conflict between the comprehensions of the cultures because it has indeed scientific content, concrete terms, technical descriptions and technical examples; this kind of content does not require any cultural interpretation because it does not generate any vagueness for the comprehension of the target culture.

Other translation techniques

3.7 Explicitation

The use of Vinay and Darbelnet translation techniques was not enough to translate the text chosen. We encountered other translation problems that could not be solved by applying the previous techniques such as calque and transposition, just to mention a couple of them. Thus, we looked for other techniques that would help us transfer the meaning of the text. There were some ideas that needed to be expanded in order to be more understandable.

In order to have a more accurate translation, a couple of other translation techniques were applied besides the Vinay and Darbelnet´s ones. Explicitation is one of them. It introduces more information or details for clarification and for a better comprehension of the message (Pym, 2005).

An example of explicitation found in the translation is presented in the following lines.

Source Text	Target Language
For instance, for a rare species like the jaguar a presence record is likely to derive from sampling over a longer time and/or larger area (e.g., using camera traps over months) than it would for the peccary, which is fairly common and easier to observe	Por ejemplo, para una especie rara como el jaguar un registro de presencia es común que derive de muestreo por un tiempo prolongado y/o de un área más amplia (por ejemplo utilizando trampas fotográficas por varios meses) y no como en el caso del pecarí que es más común y fácil de observar

An explicitation was needed in this case as a way to clarify the message on why the peccary is mentioned in this part of the text. If a literal translation would have been used, the translation would have been confusing for the target reader; on that account, by changing to a negative sentence the idea is clearer and more concrete in the target language to compare the ideas presented in the source text, giving us a good explicitation of the intended message.

3.8Amplification

The amplification technique presented by Malone in the American approach is the addition of one or more words to reinforce the meaning of the message (Fawcett, 2003). This technique was

used to translate as a side tool to enhance some parts of the target text that in my consideration were a little bit confusing with the way they were presented. By amplifying, the message becomes clearer.

This is an example of the amplification technique taken from the translation.

Source Text	Target Language
This viewpoint is likely to be a more accessible way to understand the model than previous ones that rely on machine learning concepts.	este punto de vista, a diferencia de otros que se basaban en conceptos de aprendizaje automático, tiende a ser una forma más accesible de entender el modelo.

Amplification was used to translate this message to make the idea clearer and more understandable. In this specific example, we decided to add the phrase "a diferencia de otros" to emphasize the fact that this new viewpoint is different from the previous viewpoints. It could be said that an amplification of the meaning of the conjunction *than* was done with the purpose of providing a clearer idea.

3.9Reduction

Reduction is a technique presented in the American Model and it is the omission of unnecessary information or of little importance to the target reader (Fawcett, 2003). As a way to enhance the translation of the chosen article, reduction was applied during the translating process.

These are two examples of the reduction technique found in the translation.

Source Text	Target Language
Exponential models can be especially badly behaved when applied to new data, for instance, when extrapolating to new environments.	Los modelos exponenciales pueden comportarse erróneamente cuando se aplican a nuevos datos, por ejemplo,…

In this example the adverb "especially" is omitted in Spanish. It was omitted because in Spanish adverbs are not used modify another adverb; furthermore, the omission of an attached adverb does not affect the verb neither the intended message (Real Academia Española, 2010).

Source Text	Target Language
These are the parameters of the model that weight the contribution	Parámetros del modelo en los que se mide la contribución de cada característica.

As this is an example of a sentence given in a chart inside the article, an omission of the "these are" was done because of the need of stylistics and conciseness of a chart format in Spanish (American Psychological Association, 2010). According to the American Psychological Association, it is not appropriate to use long statements inside the charts; on the contrary, it is appropriate to use phrases or more concise statements to express the ideas.

Having analyzed the use of certain translation techniques, a conclusion was reached on which techniques were the most used ones. Transposition and calque techniques as well as the literal translation appear to be the most recurrent techniques to convey the intended message of the source text. As the article translated is a specialized one, it is not surprising that these were the most used techniques because the article has a concise content; it does not need figurative interpretation nor cultural interpretation. It has an absence of cultural connotations. The most applicable changes were the ones dealing with word order, as grammatically speaking, Spanish language differs from the source language, English.

CHAPTER IV
CONCLUSIONS

Translating requires having many skills which makes it an ambitious and challenging endeavour, it requires specific translation skills but also entrepreneurial ones. Among those we find the ability to write well and adhere to a given framework. These skills demand from us the ability to transfer style, tone and cultural elements accurately from one language to another, to make adjustments for the grammatical guidelines of the target language, specialized knowledge in technical, commercial, industrial or scientific areas, IT skills in order to locate useful and truthful sources of information in the web. Unexpectedly, the skill I struggled the most at the beginning of this difficult endeavor was the ability to write well in Spanish. As a native speaker of Spanish, I have never thought I would be challenged to write properly in my own language. Secondly but very related to the previous skill, I found difficult to adhere to the framework of the type of text I was translating. Again, being a native speaker of the target language of the translation allowed me to believe I knew how to convey the meaning of the source text; nonetheless, being a native speaker does not guarantee we can write scientific prose naturally in the target language. Writing with a high proficiency takes time and dedication. The process of having to translate, being revised and correcting several times the outcome text, made me comprehend that translating is a cycling process that requires long term training skills.

Regarding the ability to transfer style, tone and cultural elements accurately from one language to another, this was a skill for which I widely informed myself by reading about the style and purpose of the text and the field of the text. Reading similar texts helped me understand the social setting of the people that would read this kind of text and get familiarized with the knowledge and specialist terminology being used in the source text. Identifying useful websites and resources to help with the translation was not as challenging as having to read many articles related to the topic of the chosen text. The more I progressed in the translation of the source text, the more I realized translation is first a process that allows novice practitioners to realize of the need to train themselves in areas they have never doubted they would need to.

Apart from the skills needed to translate the text, I found myself with the fact of having to explain the use of certain techniques. Even though they were very useful, at the very beginning of

the process it was not clear why I would make certain changes so as to convey the meaning of the original text. Applying the techniques seemed to be automatically done; the challenge was to understand the reason for applying those techniques. Thus, this lack of knowledge led me to research and learn about Spanish grammar and syntax.

I realized the translation process is time consuming and the analysis requires a great effort in order to produce a comprehensive translation. Translating a specialized article is more than just literal translation. Furthermore, I realized there are some similarities in both Spanish and English grammatical structures and that in the Spanish language the use of adverbs as in English is common but in Spanish it is mostly used in the form of adverbial phrases.

While translating, I encountered many difficulties; I identified technicalities in the source article and had to translate sentences and terms that would not followed a typical structure in Spanish. These structures are examples of a technical specialized text type. Thus, the difficulties were not only understanding the meaning of ideas but also how to convey those using an appropriate Spanish structure.

In specialized articles, such as the one presented in this paper, the translator faces the challenge of giving creative efficient solutions. The terminology had to be carefully chosen so as to fit the need of the audience to whom the article has been addressed. Since the article was written by specialized people on the field of ecology, it was the aimed at a highly educated and demanding audience on the same or related fields. I was obliged to look up in various sources and the text had to be reviewed by an expert in the topic so as to make sure the source text would accurately convey the meaning of the original text. Vinay and Darbelnet´s techniques were remarkably useful to accomplish the translation.

Thus, the translation of the article “a statistical explanation of MaxEnt for Ecologists” is the result of research, recommendations from professionals and academics specialized in the translation, ecology and geology fields and the use of translation techniques. Translating became then a constant research process about many aspects of the language and the discourse itself.

The objectives established for this monograph were achieved, as the result was an understandable and cautiously accurate translation of the article “A statistical explanation of MaxEnt for ecologists” as well as an analysis of the different techniques applied throughout the English-Spanish translation. It may be said that this translation is faithful to the contents and style of the original text but it may not be the only way this text could be translated. This means that

there may be other outcomes if any other person would translate the same text. There is not only one way of translating a text, other translators may find other structures in Spanish to convey the message of the original text. In addition, I consider this translation as greatly useful for academic purposes in El Colegio de la Frontera Sur where non-English speakers would be able to read and comprehend the content of the article.

After this experience, I would advise people interested in translating specialized texts and to novice translators to become researchers in the area of languages and on the topic of the source text. It is highly important to have a general idea beforehand of what terms and phrases you might encounter in the article; by doing so, translating a text would be easier than if you start from zero knowledge of what the text is about. Another important consideration when translating is to be aware of the different translation techniques and find the most adequate ones for the type of translation is being done. It is then highly important to read many times the original article. Reading the article one time is never enough; by reading it more than once, the translator would better understand and comprehend the main idea; this is fundamental to be able to produce a comprehensive and faithful translation. When the final product is done, proofreading is essential to achieve a neat translation. The advice from experts in the field as well as experts in linguistics and in the translation field is remarkably significant to reach a trustful translation. I totally recommend novice translators to ask for the advice of experts because their expertise will be so beneficial for the paper.

All in all, this journey of translating gives me the sense that in light of the importance of translation, as long as there is an involvement of the written language in the communication process among different languages, translation will always be a paramount tool.

REFERENCES

Alarcos, E. (1995). *Gramática de la lengua española.* Madrid: Espasa Calpe.

American Psychological Association. (2010). *Manual de publicaciones de la American Psychological Association.* Sexta Edición. México: El manual moderno.

American Psychological Association. (2013). *Publication manual of the American Psychological Association*. Sixth Edition. United States of America: Author.

Baker, M. (1992). *In other words: A coursebook on translation*. New York: Routledge.

British Broadcasting Corporation. (2011). Types of text. Retrieved from http://www.bbc.co.uk/skillswise/factsheet/en03text-l1-f-different-types-of-text

Cambridge University Press. (2018). Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org

El-Dali, H.M. (2011). Towards an understanding of the distinctive nature of translation studies. *Journal of King Saud University- Languages and Translation.* United Arab Emirates. Retrieved from http://www.sciencedirect.com/science/article/pii/S2210831910000056?via%3Dihub

Encyclopedia of mathematics. (2014). Encyclopedia of mathematics: minimax. Retrieved from https://www.encyclopediaofmath.org/index.php/Minimax

Fawcett, P. (2003). *Translation and language: linguistic theories explained*. Manchester: St. Jerome Publishing.

Foyewa, R.A. (2015). English: the international language of science and technology. *International Journal English Language and Linguistics Research*. Retrieved from http://www.eajournals.org/wp-content/uploads/English-The-International-Language-of-Science-and-Technology.pdf

García, V. (1998). *Metafísica de Aristóteles*. Madrid: Gredos. Retrieved from https://es.slideshare.net/HerlyTorres/la-metafsica-de-aristteles

Gentzler, E. (2014). Translation studies: Pre-discipline, discipline, interdiscipline, and post-discipline. *International Journal of Society, Culture and Language.* United States of America. Retrieved from http://ijscl.net/article_5620_848.html

Ghanooni, A. R. (2012). A review of the history of translation studies. *Theory and Practice in Language Studies vol. 2 no. 1.* DOI: 10.4304/tpls.2.1.77-85

Holecková, M. (2007). *Locuciones adverbiales en el español*. (Master´s thesis, Masaryk University). Retrieved from https://is.muni.cz/th/177786/ff_m/diplomova_prace.pdf

Hualde, I. J. et al. (2001). *Introducción a la lingüística hispánica*. Cambridge: Cambridge University Press.

Kohoutkova, A. (2016). *Machine translation and its use in technical translation from English into Czech*. (Bacherlor thesis, Masaryk University). Retrieved from https://is.muni.cz/th/428669/ff_b/BP_AndreaKohoutkova.pdf

Larousse. (2005). *Sintaxis lengua española*. España: VOX. Retrieved from https://es.slideshare.net/alfonsomarecamiralles/sintaxis-de-la-lengua-espaola-1-vox-libro

Levý, J. (2011). *The art of translation*. Amsterdam: John Benjamins Publishing Company.

Linguee (2018). English Spanish Dictionary online. Retrieved from https://www.linguee.com/english-spanish

Mastná, E. (2010). The nature of scientific/technical texts from viewpoint of translation studies. (Bachelor thesis, Tomas Bata University). Retrieved from http://digilib.k.utb.cz/bitstream/handle/10563/12466/mastná_2010_bp.pdf?sequence=1

Montgomery, S.L. (2000). *Science in translation: Movements of knowledge through cultures and time.* Chicago and London: The University of Chicago Press.

Montgomery, S.L. (2009). English and science: realities and issues for translation in the age of an expanding lingua franca. *The Journal of Specialized translation.* Retrieved from http://jostrans.org/issue11/art_montgomery.pdf

Murrieta, G. (1997). Trabajo práctico educativo: Translating and language as discourse tomado del libro Discourse and the translator de Basil Hatim and Ian Mason. (Bachelor thesis, University of Veracruz).

Newmark, P. (1988). *A textbook of translation*. New York: Prentice Hall.

Ni, L. (2009). For translation and theories. *English Language Teaching Journal vol. 2 no. 2.* Retrieved from https://pdfs.semanticscholar.org/57d5/8f97bb016491932cc46fc492b28f0a6581a8.pdf

Nida, E. A. (1945). Linguistics and ethnology in translation problems. DOI: 10.1080/00437956.1945.11659254.

Nida, E. A. (1964). *Towards a science of translating: with special reference to principles and procedures involved in Bible translating*. Netherlands: Brill.

Oxford University Press. (2018). *Spanish Oxford Living Dictionaries*. Retrieved from https://es.oxforddictionaries.com

Pinchuck, I. (1977). *Scientific and technical translation.* Westview Press.

Pym, A. (2005). Explaining explicitation. Retrieved from http://usuaris.tinet.cat/apym/on-line/translation/explicitation_web.pdf

Real Academia Española. (2017). *Diccionario de la lengua española*: *lambda*. Madrid: Autor. Retrieved from http://dle.rae.es

Real Academia Española. (2010). *Nueva gramática de la lengua española: Manual*. México: Espasa.

Schulte, R. (2012). What is translation? *Translation Review.* Retrieved from http://translation.utdallas.edu/essays/what-is-translation.html

Venuti, L. (2000). *The translation studies reader*. London and New York: Routledge.

Venuti, L. (2004). *The translation studies reader*. London and New York: Routledge.

Vinay, J.-P., Darbelnet J. (1995). *Comparative stylistics of French and English: A methodology for translation*. Amsterdam: John Benjamins Publishing Company.

APPENDIX

Diversity and Distributions, (Diversity Distrib.) (2011) 17, 43–57

BIODIVERSITY RESEARCH

A statistical explanation of MaxEnt for ecologists

Jane Elith[1]*, Steven J. Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1] and Colin J. Yates[5]

[1]School of Botany, The University of Melbourne, Parkville, VIC 3010 Australia, [2]AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA, [3]Department of Statistics, Stanford University, CA 94305, USA, [4]Yahoo! Labs, 111 West 40th Street (17th Floor). New York, NY 10018, USA, [5]Science Division, Western Australian Department of Environment and Conservation, LMB 104, Bentley Delivery Centre, WA6983, Australia

*Correspondence: Jane Elith, School of Botany, The University of Melbourne, Parkville, VIC 3010 Australia.
E-mail: j.elith@unimelb.edu.au

ABSTRACT

MaxEnt is a program for modelling species distributions from presence-only species records. This paper is written for ecologists and describes the MaxEnt model from a statistical perspective, making explicit links between the structure of the model, decisions required in producing a modelled distribution, and knowledge about the species and the data that might affect those decisions. To begin we discuss the characteristics of presence-only data, highlighting implications for modelling distributions. We particularly focus on the problems of sample bias and lack of information on species prevalence. The keystone of the paper is a new statistical explanation of MaxEnt which shows that the model minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space. For many users, this viewpoint is likely to be a more accessible way to understand the model than previous ones that rely on machine learning concepts. We then step through a detailed explanation of MaxEnt describing key components (e.g. covariates and features, and definition of the landscape extent), the mechanics of model fitting (e.g. feature selection, constraints and regularization) and outputs. Using case studies for a *Banksia* species native to south-west Australia and a riverine fish, we fit models and interpret them, exploring why certain choices affect the result and what this means. The fish example illustrates use of the model with vector data for linear river segments rather than raster (gridded) data. Appropriate treatments for survey bias, unprojected data, locally restricted species, and predicting to environments outside the range of the training data are demonstrated, and new capabilities discussed. Online appendices include additional details of the model and the mathematical links between previous explanations and this one, example code and data, and further information on the case studies.

Keywords

Absence, ecological niche, entropy, machine learning, presence-only, species distribution model.

INTRODUCTION

Species distribution models (SDMs) estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites (Franklin, 2009). They are widely used for many purposes in biogeography, conservation biology and ecology (Elith & Leathwick, 2009a; Table 1). In the last two decades, there have been many developments in the field of species distribution modelling, and multiple methods are now available. A major distinction among methods is the kind of species data they use. Where species data have been collected systematically – for instance, in formal biological surveys in which a set of sites are surveyed and the presence/absence or abundance of species at each site are recorded – regression methods familiar to most ecologists (e.g., generalized linear or additive models, GLMs or GAMs; or ensembles of regression trees: random forests or boosted regression trees, BRT) are used.

However, for most regions, systematic biological survey data tend to be sparse and/or limited in coverage. Species records are available though in the form of presence-only records in herbarium and museum databases. Many of these databases

DOI: 10.1111/j.1472-4642.2010.00725.x
http://wileyonlinelibrary.com/journal/ddi

Diversity and Distributions
A Journal of Conservation Biogeography

Table 1 Examples of published studies using MaxEnt, showing variation in purpose, extent and organism.

Primary purpose	Extent	Organisms	Refs
Predict current distributions as input for conservation planning, risk assessments or IUCN listing, or new surveys	Andes	Humming-birds	Tinoco *et al.* (2009)
	Global	Stony corals seamounts	Tittensor *et al.* (2009)
Understand environmental correlates of species occurrences, groups of species, or other	Norway	Macrofungi	Wollan *et al.* (2008)
	Portugal	European wildcat	Monterroso *et al.* (2009)
Predict potential distributions for invasive species, or explore expanding distributions	New Zealand	Ants	Ward (2007a)
	China	Nematode	Wang *et al.* (2007)
Predict species richness or diversity	California	Amphibians and reptiles	Graham & Hijmans (2006)
	Brazil	Myrtaceae 19 species	Murray-Smith *et al.* (2009)
Predict current distributions for understanding morphological / genetic diversity ("phylogeography", "phyloclimatic studies"), endemism and evolutionary niche dynamics	Global	Seaweeds	Verbruggen *et al.* (2009)
	Andes	Birds	Young *et al.* (2009)
	Madagascar	Bats	Lamb *et al.* (2008)
Hindcast distributions to understand patterns of endemism, vicariance, etc	NW Europe	Pond snails	Cordellier & Pfenninger (2009)
	Brazilian coast	Forests	Carnaval & Moritz (2008)
Forecast distributions to understand changes with climate change / land transformation; includes retrospective studies	Mediterr'n + surrounds	Cyclamen	Yesson & Culham (2006)
	Regional W. Australia	Banksia	Yates *et al.* (2010)
	Canada	Butterflies	Kharouba *et al.* (2009)
Test model performance against other methods	Patagonia	Insects	Tognelli *et al.* (2009)
	Local region in California	Rare plants	Williams *et al.* (2009)
	Regional to national	Many species	Elith *et al.* (2006)

represent well over a century of public and private investment in biological science and are a hugely important source of species occurrence data. The desire to maximize the utility of such resources has spawned an array of SDM methods for modelling presence-only data. MaxEnt (Phillips *et al.*, 2006; Phillips & Dudík, 2008) is one such method and is the focus of this paper.

MaxEnt's predictive performance is consistently competitive with the highest performing methods (Elith *et al.*, 2006). Since becoming available in 2004, it has been utilized extensively for modelling species distributions. Published examples cover diverse aims (finding correlates of species occurrences, mapping current distributions, and predicting to new times and places) across many ecological, evolutionary, conservation and biosecurity applications (Table 1). Government and non-government organizations have also adopted MaxEnt for large-scale, real-world biodiversity mapping applications, including the Point Reyes Bird Observatory online application (http://www.prbo.org/) and the Atlas of Living Australia (http://www.ala.org.au/). JE and SJP's involvement in such programs identified a need for an ecologically accessible explanation of MaxEnt. Existing descriptions include concepts from machine learning that tend to be outside the common experience of many ecologists.

In this article, we explain the MaxEnt modelling method with emphasis on a statistical explanation of the method, on what it assumes, and on the impacts of choices made in the modelling process. We use two case studies to examine the effects of background selection and model settings, and to illustrate the applicability of the model for exploring ecological relationships with fine-scale, vector-based environmental data. Our aim is to promote understanding of the method and recommend useful approaches to data preparation and model fitting and interpretation.

PREAMBLE: WHAT IS SPECIAL ABOUT THE PRESENCE-ONLY CASE?

Expanding use of presence-only data for modelling species distributions has prompted wide discussion about the sorts of distributions (e.g., potential vs. realized) that can be modelled with presence-only data in contrast to presence-absence data (e.g., Soberón & Peterson, 2005; Chefaoui & Lobo, 2007; Hirzel & Le Lay, 2008; Jiménez-Valverde *et al.*, 2008; Soberón & Nakamura, 2009; Lobo *et al.*, 2010). As mentioned in several of these articles, the subject is complex because of the interplay of data quality (amount and accuracy of species data; ecological relevance of predictor variables; availability of information on disturbances, dispersal limitations and biotic interactions), modelling method and scale of analysis. A comprehensive review of the issues would be useful, but here we restrict ourselves to key points important for this paper.

Some of the published discussion suggests that presence-only data in some sense release us from the problems of unreliable absence records (e.g., Jiménez-Valverde *et al.*, 2008), particularly emphasizing that absences bear such strong imprints of biotic interactions, dispersal constraints

and disturbances that they may preclude modelling of potential distributions (sensu Svenning & Skov, 2004). However, the presence records are also imprinted by many of the factors affecting absences. If a species is absent from an environmentally suitable area because, say, past disturbances have caused local extinctions, the signal of that absence will be found in the distribution of presence records: there will be no presence records in the disturbed area. Regardless of whether absences are used in modelling, the pattern in the presence records will suggest the area is unsuitable, and the model will be affected by this patterning. Similarly, if the detectability of a particular species varies from site to site, then not only does this result in some false absences in presence-absence data, it also affects the pattern of presences in presence-only data. This leads naturally to the conclusion that dispensing with absences does not address the limitations often attributed to absence data, such as the fact that species are not perfectly detectable and may not occupy all suitable habitat. This thinking means that we will approach the description of the presence-only modelling problem as one that is trying to model the same quantity that is modelled with presence-absence data, that is, the probability of presence of a species (to be defined more carefully below).

From here on, we assume that the data available to the modeller are presence-only, i.e., a set of locations within L, the landscape of interest, where the species has been observed. Let $y = 1$ denote presence, $y = 0$ denote absence, $\mathbf{z}$ denote a vector of environmental covariates, and background be defined as all locations within L (or a random sample thereof). Assume the environmental variables or covariates $\mathbf{z}$ (representing environmental conditions) are available landscape wide. Define $f(\mathbf{z})$ to be the probability density of covariates across L, $f_1(\mathbf{z})$ to be the probability density of covariates across locations within L where the species is present, and similarly, $f_0(\mathbf{z})$ where the species is absent (densities – or probability density functions – describe the relative likelihood of random variables over their range and can be univariate or multivariate). The quantity that we wish to estimate is, as with presence-absence data, the probability of presence of the species, conditioned on environment: $\Pr(y = 1|\mathbf{z})$. Strictly presence-only data only allow us to model $f_1(\mathbf{z})$, which on its own cannot approximate probability of presence. Presence/background data allows us to model both $f_1(\mathbf{z})$ and $f(\mathbf{z})$, and this gets to within a constant of $\Pr(y = 1|\mathbf{z})$, because Bayes' rule gives:

$$\Pr(y = 1|\ \mathbf{z}) = f_1(\mathbf{z})\Pr(y = 1)/f(\mathbf{z}) \quad (1)$$

The only quantity that is lacking is the second term, $\Pr(y = 1)$, i.e., the prevalence of the species (proportion of occupied sites) in the landscape. Formally, we say that prevalence is not identifiable from presence-only data (Ward *et al.* 2009). This means that it cannot be exactly determined, regardless of the sample size; this is a fundamental limitation of presence-only data. As an aside we note, however, that absence data are plagued by issues of detection probability (Wintle *et al.*, 2004; MacKenzie, 2005) so that even presence-absence data may not yield a good estimate of prevalence.

A second fundamental limitation of presence-only data is that sample selection bias (whereby some areas in the landscape are sampled more intensively than others) has a much stronger effect on presence-only models than on presence-absence models (Phillips *et al.*, 2009). Imagine that $f_1(\mathbf{z})$ is contaminated by a sample selection bias $s(\mathbf{z})$. This bias will most commonly occur in geographic space (e.g., close to roads) but could be environmentally based (e.g., visiting wet gullies) but, regardless, will map into covariate space. Under biased sampling, a presence-only model gives an estimate of $f_1(\mathbf{z})s(\mathbf{z})$ rather than $f_1(\mathbf{z})$. That is, we get a model that combines the species distribution with the distribution of sampling effort (Soberón & Nakamura, 2009). In contrast, for presence-absence models, sample selection bias affects both presence and absence records, and the effect of the bias cancels out (under reasonable assumptions, see Zadrozny, 2004).

So far we have treated presence or absence as a binary event, but in reality defining the response variable is not straightforward, and in this regard, presence-only data are quite different from presence-absence data (Pearce & Boyce, 2006). Presence or absence of a species is dependent on the time frame and spatial scale – for example, a vagile species (such as a bird) may be present at some times but not others, while a plant species will be more likely to be found in a large plot with given environmental conditions than in a small plot with the same conditions. Absence of a plant species from a 1-km^2 quadrat around a point implies absence in a 1-m^2 quadrat around that point, but not vice versa. With presence-absence data, it is not hard to incorporate these complexities in the formulation of the response variable (i.e., the specification of what constitutes a sample), or via sampling covariates in the model, provided survey details are available (Leathwick, 1998; MacKenzie & Royle, 2005; Schulman *et al.*, 2007; Ward, 2007b). However, with presence-only data, we typically have occurrence data that do not have any associated temporal or spatial scale. The record is usually simply a record of the species at a location, with no information on search area or time.

With presence-absence data, the definition of the response variable should naturally be consistent with the sampling method. For example, if the available data are surveys of 1-m^2 quadrats, then $y = 1$ should correspond to the species being present in a 1-m^2 quadrat. With presence-only data, the available data do not usually describe the survey method, so the modeller has considerable leeway in defining the response variable. A common approach is to implicitly assume a sampling unit of size equal to the grain size of available environmental data (see Elith & Leathwick, 2009a for discussion of grain).

To summarize, we posit that with presence and background data, we can model the same quantity as with presence-absence data, up to the constant $\Pr(y = 1)$. However, if presence-absence survey data are available, we believe it is generally advisable to use a presence-absence modelling method, since in that case the models are less susceptible to problems of sample selection bias, the survey method will often be known and can be used to appropriately define the response variable for modelling, and we take advantage of all information in the

data. In particular, presence-absence data give us much better information about prevalence than presence-only, because – even though there may be some difficulties because of imperfect detection – they solve the major problem of non-identifiability. We will come back to this when we discuss the logistic output of MaxEnt.

EXPLANATION OF MAXENT

Here for the first time, we describe MaxEnt using statistical terminology and notation, providing a break from the machine learning terminology in previous papers. As we describe the model we will highlight possibilities for – and implications of – modelling choices and defaults, and consider how MaxEnt addresses the limitations of presence-only data identified above. We relegate the more technical considerations to boxes and Supporting Information, to avoid interrupting the flow of the explanation.

Covariates and features

Most ecologists, following the statistical literature, call the independent variables in a model the covariates, predictors or inputs. In SDMs, these include environmental factors that are relevant to habitat suitability (e.g., estimates of climate, topography, and soil for plants; temperature, salinity and prey abundance for marine fishes). Since species' responses to these tend to be complex, it is usually desirable to fit nonlinear functions (Austin, 2002). In regression this can be achieved by applying transformations to the covariates – for instance, creating basis functions for polynomials and splines, including piecewise linear functions. Complex models are fitted as linear combinations of these basis functions in methods including GLMs and GAMs (Hastie *et al.*, 2009, Chapter 5). In machine learning, basis functions and other transformations of available data are termed features –i.e., features are an expanded set of transformations of the original covariates.

In MaxEnt, selected features are formed "behind the scenes", in the same way as in regression, where the model matrix is augmented by terms specified in the model (e.g., polynomials, interactions). The MaxEnt fitted function is usually defined over many features, meaning that in most models there will be more features than covariates. MaxEnt currently has six feature classes: linear, product, quadratic, hinge, threshold and categorical (further details in Appendix S1). Products are products of all possible pair-wise combinations of covariates, allowing simple interactions to be fitted. Threshold features allow a "step" in the fitted function; hinge features are similar except they allow a change in gradient of the response. Many threshold or hinge features can be fitted for one covariate, giving a potentially complex function. Hinge features (which are basis functions for piecewise linear splines), if used alone, allow a model rather like a generalized additive model (GAM): an additive model, with nonlinear fitted functions of varying complexity but without the sudden steps of the threshold features. MaxEnt's default is to allow all feature types (conditional on sufficient species data being available), but it is worth considering simpler models, as discussed later under implications for modelling.

The MaxEnt model – a short overview

Previous papers have described MaxEnt as estimating a distribution across geographic space (Phillips *et al.*, 2006; Phillips & Dudík, 2008). Here, we give a different (but equivalent) characterization that focuses on comparing probability densities in covariate space (Fig. 1). In doing so, we rely strongly on the PhD research of TH's past student, Gill Ward (Ward, 2007b), and acknowledge her contribution. Equation 1 shows that if we know the conditional density of the covariates at the presence sites, $f_1(\mathbf{z})$, and the marginal (i.e., unconditional) density of covariates across the study area $f(\mathbf{z})$, we then only need knowledge of the prevalence $\Pr(y = 1)$, to calculate conditional probability of occurrence. MaxEnt first makes an estimate of the ratio $f_1(\mathbf{z})/f(\mathbf{z})$, referred to as MaxEnt's "raw" output. This is the core of the MaxEnt model output, giving insight about what features are important and estimating the relative suitability of one place vs. another. Because the required information on prevalence is not available for calculating conditional probability of occurrence, a workaround has been implemented (termed MaxEnt's "logistic" output). This treats the log of the output: $\eta(\mathbf{z}) = \log(f_1(\mathbf{z})/f(\mathbf{z}))$ as a logit score, and calibrates the intercept so that the implied probability of presence at sites with "typical" conditions for the species (i.e., where $\eta(\mathbf{z})$ = the average value of $\eta(\mathbf{z})$ under f_1) is a parameter τ. Knowledge of τ would solve the non-identifiability of prevalence, and in the absence of that knowledge MaxEnt arbitrarily sets τ to equal 0.5. This logistic transformation is monotone (order preserving) with the raw output. We work through each part of the MaxEnt model in the following sections, showing how the choice of landscape, species data, and selected settings influence the results.

The landscape and species records

The landscape of interest (L) is a geographic area suggested by the problem and defined by the ecologist. It might, for instance, be limited by geographic boundaries or by an understanding of how far the focal species could have dispersed. We then define L_1 as the subset of L where the species is present.

The distribution of covariates in the landscape is conveyed by a finite sample – a collection of points from L with associated covariates, typically called a background sample. These data may be supplied in the form of grids of covariates covering a pixelation of the landscape; as a default MaxEnt randomly samples 10,000 background locations from covariate grids, but the background data points can also be specified (see Yates *et al.*, 2010 and case studies below) and grids are not essential (case study 2). Note that the background sample does not take any account of the presence locations – it is simply a

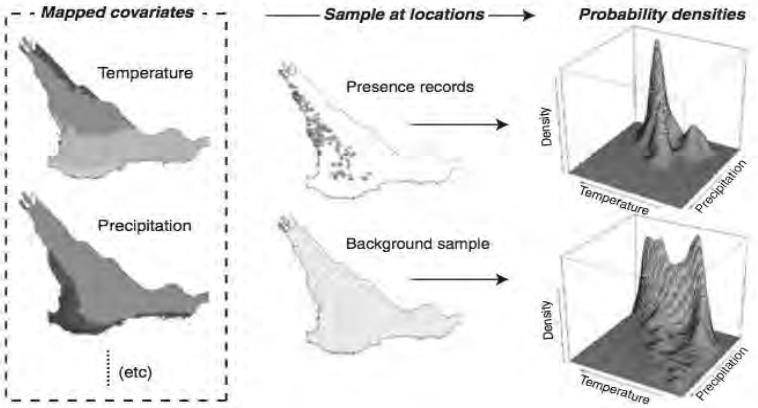

Figure 1 A diagrammatic representation of the probability densities relevant to our statistical explanation, using data presented in case study 1. The maps on the left are two example mapped covariates (temperature and precipitation). In the centre are the locations of the presence and background samples. The density estimates on the right are not in geographic (map) space, but show the distributions of values in covariate space for the presence (top right) and background (bottom right) samples. These could represent the densities $f_1(\mathbf{z})$ and $f(\mathbf{z})$ for a simple model with linear features.

sample of L, and could by chance include presence locations. Using a random background sample implies a belief that the sample of presence records is also a random sample from L_1. We deal later with the case of biased samples.

Description of the model

MaxEnt uses the covariate data from the occurrence records and the background sample to estimate the ratio $f_1(\mathbf{z})/f(\mathbf{z})$. It does this by making an estimate of $f_1(\mathbf{z})$ that is consistent with the occurrence data; many such distributions are possible, but it chooses the one that is closest to $f(\mathbf{z})$. Minimizing distance from $f(\mathbf{z})$ is sensible, because $f(\mathbf{z})$ is a null model for $f_1(\mathbf{z})$: without any occurrence data, we would have no reason to expect the species to prefer any particular environmental conditions over any others, so we could do no better than predict that the species occupies environmental conditions proportionally to their availability in the landscape. In MaxEnt, this distance from $f(\mathbf{z})$ is taken to be the relative entropy of $f_1(\mathbf{z})$ with respect to $f(\mathbf{z})$ (also known as the Kullback-Leibler divergence).

Using background data informs the model about $f(\mathbf{z})$, the density of covariates in the region, and provides the basis for comparison with the density of covariates occupied by the species – i.e., $f_1(\mathbf{z})$ (Fig. 1). Constraints are imposed so that the solution is one that reflects information from the presence records. For example, if one covariate is summer rainfall, then constraints ensure that the mean summer rainfall for the estimate of $f_1(\mathbf{z})$ is close to its mean across the locations with observed presences. The species' distribution is thus estimated by minimizing the distance between $f_1(\mathbf{z})$ and $f(\mathbf{z})$ subject to constraining the mean summer rainfall estimated by f_1 (and the means of other covariates) to be close to the mean across presence locations.

We note that previous papers describing MaxEnt focused on a location-based definition over a finite landscape (typically a grid of pixels). We will call this a definition based in geographic space and compare it with our new description, which focuses on environmental (covariate) space. Note, though, that we are not implying by this wording that in either definition there is any consideration of the geographic proximity of locations unless geographic predictors are used. In the original definition (Phillips *et al.*, 2006), the target was $\pi(x) = \Pr(x|y = 1)$, which was a probability distribution over pixels (or locations) x. This was called the "raw" distribution (Phillips *et al.*, 2006), and gave the probability, given the species is present, that it is found at pixel x. Maximizing the entropy of the raw distribution is equivalent to minimizing the relative entropy of $f_1(\mathbf{z})$ relative to $f(\mathbf{z})$, so the two formulations are equivalent (see Appendix S2 for equations showing the transition from the geographic to environmental definitions). The null model for the raw distribution was the uniform distribution over the landscape, since without any data we would have no reason to think the species would prefer any location to any other. As mentioned at the start of this section, in environmental space, the equivalent null model for $\mathbf{z}$ is $f(\mathbf{z})$.

Constraints were described earlier in reference to covariates, but – as explained in the section on covariates and features – MaxEnt actually fits the model on features that are transformations of the covariates. These allow potentially complex relationships to be modelled. The constraints are extended from being constraints on the means of covariates to being constraints on the means of the features. We will call the vector of features $h(\mathbf{z})$ and the vector of coefficients β (note, this notation is different to previous papers: Table 2). As explained

Table 2 Terminology used in this paper.

Item/concept	Definition	Notation
Background	A sample of points from the landscape	
Entropy	A measure of dispersedness. Previous papers* described the model as maximizing entropy in geographic space; this paper focuses on minimizing relative entropy in covariate space.	
Features	An expanded set of transformations of the original covariates	
Mask	A gridded layer of 1 / no data used to indicate areas to be included in background sampling (=1) and those to be excluded (=no data). To be included as a predictor. For projecting to the whole region, a grid called mask, but containing any values – say, 1 across the whole region of interest – should be supplied along with all other covariate grids.	
MESS map	Multivariate Environmental Similarity Surface –measures the similarity of any given point to a reference set of points, with respect to the chosen predictor variables. It reports the closeness of the point to the distribution of reference points, gives negative values for dissimilar points and maps these values across the whole prediction region (Elith *et al.*, 2010)	
Prevalence is not identifiable	Prevalence cannot be exactly determined from presence-only data in isolation, regardless of the sample size. This is a fundamental limitation of presence-only data.	
Probability density functions	Describe the relative likelihood of random variables over their range; can be univariate or multivariate.	
Regularization (tuning) parameters	Regularization refers to smoothing the model, making it more regular, so as to avoid fitting too complex a model. In MaxEnt the regularization parameters can be changed if required.	β in previous papers*, λ in this paper
Sampling bias	Some areas in the landscape are sampled more intensively than others. Usually occurs in geographic space but could be environmentally based.	$s(\mathbf{z})$
Weights or coefficients	These are the parameters of the model that weight the contribution of each feature.	λ in previous papers*, β in this paper

*Phillips *et al.* (2006), Phillips & Dudík (2008)

in Phillips *et al.* (2006), minimizing relative entropy results in a Gibbs distribution (Della Pietra *et al.*, 1997) which is an exponential-family model:

$$f_1(\mathbf{z}) = f(\mathbf{z})e^{\eta(\mathbf{z})} \quad (2)$$

where $\eta(\mathbf{z}) = \alpha + \beta \cdot h(\mathbf{z})$

and α is a normalizing constant that ensures that $f_1(\mathbf{z})$ integrates (sums) to 1.

From this, it is clear that the target of a MaxEnt model is $e^{\eta(\mathbf{z})}$, which estimates the ratio $f_1(\mathbf{z})/f(\mathbf{z})$. It is a log-linear model, similar in form to a GLM, and depends on both the presence samples and the background samples that are used in forming the estimate. Hence, the definition of the landscape is intimately linked to the solution that is given.

Mechanics of the solution

In coming to a solution, MaxEnt needs to find coefficients (betas) that will result in the constraints being satisfied but not match them so closely that it overfits and produces a model with limited generalization. MaxEnt handles the issue by setting an error bound, or maximum allowed deviation from the sample (empirical) feature means. MaxEnt first automatically rescales all features to have the range 0–1. Then, an error bound (λ_j in equation 3) is calculated for each feature (again note the change in notation from previous papers, Table 2). It will reflect the variation in sample values for that feature, adjusted by a tuned (pre-set) parameter for the feature class (Phillips & Dudík, 2008; and equation 3). MaxEnt *could* estimate feature error bounds only from the data, for example using cross-validation, but to simplify model fitting and because the data are often biased, it uses feature class-specific tuned parameters based on a large international dataset (Phillips & Dudík, 2008). That dataset covers 226 species, 6 regions of the world, sample sizes ranging from 2 to 5822, and 11–13 predictors per region (Elith *et al.*, 2006). It is possible that the tuning may not work well for very different datasets – e.g., if there are many more predictors. The tuned parameters can be changed by the user if desired. The pre-tuning also includes restrictions to the set of feature classes that will be considered for small samples.

$$\lambda_j = \lambda\sqrt{\frac{s^2[h_j]}{m}} \quad (3)$$

where λ_j is the regularization parameter for feature h_j. This feature's variance is $s^2[h_j]$ over the m presence sites, and its feature class has a

tuning parameter λ. Conceptually, λ_j corresponds to the width of the confidence interval, and therefore it takes the form of the standard error (the square root expression) multiplied by the parameter λ according to the desired confidence level.

The lambdas in equation 3 allow regularization – i.e., smoothing the distribution, making it more regular. These error bounds are a specific form of regularization called L1-regularization (Tibshirani, 1996) that gives sparse solutions (ones with many zeros, i.e., many features removed). Regularization is not specific to MaxEnt; it is a common modern approach to model selection. It can be thought of as a way of shrinking the coefficients (the betas) – i.e., penalizing them – to values that balance fit and complexity, allowing both accurate prediction and generality. In MaxEnt, the fit of the model is measured at the occurrence sites, using a log likelihood (Box 1). A highly complex model will have a high log likelihood, but may not generalize well. The aim of regularization is to trade off model fit (the first term in equation 4 below) and model complexity (the second term in equation 4). In this sense, MaxEnt fits a penalized maximum likelihood model (Phillips & Dudík, 2008; equation 4) closely related to other penalties for complexity such as Akaike's Information Criterion (AIC, Akaike, 1974). Maximizing the penalized log likelihood is equivalent to minimizing the relative entropy subject to the error-bound constraints.

$$\max_{\alpha,\beta} \frac{1}{m}\sum_{i=1}^{m} \ln(f(\mathbf{z}_i)e^{\eta(\mathbf{z}_i)}) - \sum_{j=1}^{n} \lambda_j|\beta_j| \qquad (4)$$

subject to $\int_L f(\mathbf{z})e^{\eta(\mathbf{z})}d\mathbf{z} = 1$

where $\mathbf{z}$ is the feature vector for occurrence point i of m sites, and for $j = 1\ldots$ n features.

MaxEnt's logistic output

MaxEnt (from version 3 onwards) gives a logistic output as its default. It is an attempt to get as close as we can to an estimate of the probability that the species is present, given the environment, $\Pr(y = 1|\mathbf{z})$. This is a post-transformation of the MaxEnt raw output that makes certain assumptions about prevalence and sampling effort (Box 2 and Appendix S3). These two output types of MaxEnt (raw and logistic) are monotonically related, so if the purpose of a study is to rank sites according to suitability, it does not matter which type is used – both will yield identical ranking and hence identical rank-based measures (e.g., AUC values). MaxEnt's logistic transformation is not a commonly used statistical procedure, so here we explain the background and the issues.

From equation 1, we see that a simple approach to estimate $\Pr(y = 1|\mathbf{z})$ would be to simply multiply $e^{\eta(\mathbf{z})}$ by a constant that estimates prevalence; this approach has the disadvantage that $e^{\eta(\mathbf{z})}$ can be arbitrarily large, which implies that we may get an estimate of $\Pr(y = 1|\mathbf{z})$ that exceeds 1 (Keating & Cherry, 2004; Ward, 2007b). Exponential models can be especially badly behaved when applied to new data, for instance, when extrapolating to new environments. To avoid these problems, and to side-step the non-identifiability of the species prevalence, $\Pr(y = 1)$, MaxEnt's logistic output transforms the model from an exponential family model (equation 2) to a logistic model:

$$\Pr(y = 1|\mathbf{z}) = \tau e^{\eta(\mathbf{z})-r}/(1-\tau+\tau e^{\eta(\mathbf{z})-r}) \qquad (5)$$

where $\eta(\mathbf{z})$ is the linear score from equation 2, r is the relative entropy of MaxEnt's estimate of $f_1(\mathbf{z})$ from $f(\mathbf{z})$, and τ is the

Box 1 Log likelihood

In statistics, a log likelihood describes the log of the probability of an observed outcome. It varies from 0 [ln(1)] to negative infinity [ln(0)]. If the space of outcomes is continuous, we measure the probability density at the observed outcome, rather than probability. With presence-only data the only known outcomes are presences, so when measuring likelihoods, the calculation is simply done at presence sites (compared to logistic regression where they are calculated at presence and absence sites). For a set of observations the average log likelihood is estimated. When fitting a MaxEnt model from the software interface, a gain bar is shown reporting the improvement in penalized average log likelihood compared to a null model.

Box 2 Consider the jaguar: reconciling logistic output and sampling effort

The jaguar (*Panthera onca*) and the collared peccary (*Pecari tajacu*) have very similar ranges in South and Central America, and MaxEnt models for the two species would therefore be similar using the default τ. However, the jaguar is much rarer than the peccary, so how can the outputs be compared? The answer is that probability of presence is only defined relative to a given definition of presence/absence (i.e., the temporal and spatial scale of a sample; see Preamble). For instance, for a rare species like the jaguar a presence record is likely to derive from sampling over a longer time and/or larger area (e.g., using camera traps over months) than it would for the peccary, which is fairly common and easier to observe. Since with presence-only data there is usually no information on sampling effort, this elasticity in definition is largely conceptual – it explains how to think about the meaning of the probabilities across species. When τ is 0.5 typical presence sites will have a logistic output near 0.5. This is reasonable as long as we can interpret logistic output as corresponding to a temporal and spatial scale of sampling that results in a 50% chance of the species being present in suitable areas. See Appendix S3 for more information.

Alternatively, if the value of τ is available for a given level of sampling effort, it could be used instead of the default and then the predictions for the two species would be directly comparable. Tau measures a form of rarity (Rabinowitz *et al.*, 1986). The jaguar has very low local abundance even in suitable areas within its range, so a very small value τ is appropriate for all but the most intensive sampling schemes. The estimate of τ could come from expert knowledge or targeted surveys. While τ is determined by prevalence, and vice versa, τ is arguably more ecologically intuitive, as it is a characteristic property of the species while prevalence strongly depends on the choice of study area.

probability of presence at sites with "typical" conditions for the species (i.e., where $\eta(\mathbf{z})$ = the average value of $\eta(\mathbf{z})$ under f_1). The default value for τ is arbitrarily set at 0.5. Equation 5 is derived using a "minimax" or robust Bayes approach (details in Appendix S3). In unsuitable areas, the logistic output's denominator is close to 1-τ, so the result is just a linear scaling of raw output. For more suitable areas, the effect of the denominator is mainly to bound model output below 1. The logistic output with $\tau = 0.5$ empirically gives a better calibrated estimate of $\Pr(y = 1|\mathbf{z})$ than the untransformed raw values (Phillips & Dudík, 2008).

Because the species prevalence, $\Pr(y = 1)$, is not identifiable from occurrence data, the prevalence $\Pr(y = 1)$ implied by the logistic output (with the default value of τ) will not converge to the true prevalence, even given ample occurrence data. On the other hand, the true prevalence depends on the definition of the response variable y, which itself depends on the sampling method - often unknown for presence-only data (see Preamble). Further, if additional information is available that could be used to estimate τ, prevalence will be identifiable. We therefore offer guidance for interpretation of MaxEnt's logistic output in relation to sampling effort and τ (Box 2).

Implications for modelling

These properties of the MaxEnt model have several implications for how it should be used.

MaxEnt relies on an unbiased sample (as do all species modelling methods), so efforts in collecting a comprehensive set of presence records (cleaned for duplicates and errors) and dealing with biases are critical (Newbold, 2010). Methods are implemented for dealing with biased species data (see case study 1, and Dudík *et al.*, 2006; Phillips *et al.*, 2009; Elith *et al.*, 2010). The main alternatives are to provide background data with similar biases to those in the presence data (e.g., by using sites surveyed for other species in the same biological group) or to use a bias grid that indicates the biases in the survey data (see tutorial provided with MaxEnt for an example). All the values in this grid should be positive (or specified as no data) and should be scaled to represent relative survey effort across the landscape *L*. There is one additional important consideration. If the covariate grids are unprojected (i.e., latitude and longitude in degrees, for instance WorldClim data - http://www.worldclim.org/), any region covering a non-trivial range in latitude (say, more than 200 km, especially away from the equator) will have grid cells of varying area. For instance, in Australia, cells in the north are approximately 1.3 times the area of cells in the south. MaxEnt randomly samples cells, implicitly assuming equal area cells. Solutions are to project the grids to an equal area projection, create a grid showing the variations in cell area that can then be used as a bias grid, or create your own background sample with appropriate sampling weights (case study 1).

The MaxEnt solution is affected by the landscape (region) used for the background sample, as demonstrated by VanDerWal *et al.* (2009). Conceptually, that landscape should include the full environmental range of the species and exclude areas that definitely have not been searched (unless the reason for no searching is that there is unambiguous knowledge that the species does not occur there). A local endemic that is, for instance, likely to be geographically restricted because of barriers to dispersal, should be modelled with background selected from areas into which it might have dispersed. Cleared areas that would not be surveyed because there is no remaining habitat for the species should be excluded. Excluding areas from the background sample can be achieved through use of masks, as explained in the online tutorial for MaxEnt (and see Table 2). Predictions can still be made to excluded areas, if required, by using the projection facilities. We will discuss some caveats to these general concepts for background selection in the first case study.

MaxEnt includes a range of feature types, and subsets of these can be used to simplify the solution. By default, the program restricts the model to simple features if few samples are available (linear is always used; quadratic with at least 10 samples; hinge with at least 15; threshold and product with at least 80) because – as for any modelling method – few samples provide limited information for determining the relationships between the species and its environment (Barry & Elith, 2006; Pearson *et al.*, 2007). In such cases, it is also a good idea to first reduce the candidate predictor set using ecological understanding of the species (Elith & Leathwick, 2009b). Hinge features tend to make linear and threshold features redundant, and one way to form a model with relatively smooth fitted functions, more like a GAM, is to use only hinge features (e.g., Elith *et al.*, 2010 and case study 1). Excluding product features creates an additive model that is easier to interpret, although less able to model complex interactions.

MaxEnt has an inbuilt method for regularization (L1-regularization) that is reliable and known to perform well (Hastie *et al.*, 2009). It implicitly deals with feature selection (relegating some coefficients to zero) and is unlikely to be improved - and more likely, degraded - by procedures that use other modelling methods to pre-select variables (e.g., Wollan *et al.*, 2008). In particular, it is more stable in the face of correlated variables than stepwise regression, so there is less need to remove correlated variables (unless some of them are known to be ecologically irrelevant), or preprocess covariates by using PCA and selecting a few dominant axes. Note, though, that since there are often many variables available, some expert pre-selection of a candidate set is often a good idea (Elith & Leathwick, 2009b). Selecting proximal variables is likely to be particularly important when models are to be used in different regions or climates. If smoother models are required, regularization parameters can be increased by the user (e.g., see Elith *et al.*, 2010).

If comparing models for different species some care is needed in use of the logistic outputs because probability of presence is only defined relative to a given level of sampling effort, which as a default is assumed to be one that results in a 50% chance of observing the species in suitable areas (Box 2). The implied sampling effort therefore depends on the species.

This presents some challenges for cross-species comparisons of habitable areas, but these are a direct result of using presence-only data, and are not unique problems to MaxEnt. Some users may in fact see the species-specific scaling as an opportunity, since the literature on favourability functions (e.g., Real *et al.*, 2006) claims that probability of presence is itself hard to work with.

USING MAXENT

Case study 1: Modelling current and future distributions of a plant

This analysis predicts the current distribution of *Banksia prionotes*, then uses the model to identify where suitable environments for the species are likely to occur under climate change. In it, we highlight the importance of choice of landscape and dealing with survey bias, debiasing background samples from unprojected covariate grids, use of a reduced set of feature types for a smoother model, and tools for assessing the environments in new times or places.

Banksia prionotes is a woody shrub to small tree native to south-west Western Australia (WA). It is widely distributed across its range and shows a preference for deep sandy soils. Often a dominant plant in scrubland and low woodlands, it is an important nectar source for honeyeaters, and an outstanding ornamental species for cut flowers.

Methods

Here, we use species data from the Banksia Atlas (Taylor & Hopper, 1988; Yates *et al.*, 2010), with 361 records for *B. prionotes* from the 4631 sites across the South West Australia Floristic Region (SWAFR) that were surveyed for *Banksia* and for which we had complete environmental data. The atlas is the result of a community science project, and records could either be interpreted as presence-only or presence-absence data, depending on what assumptions are made about the search patterns of contributors. Here we treat them as presence-only data, but use the full set of locations as one "background" treatment. To demonstrate the effect of this choice, two alternative backgrounds (i.e., landscape definitions) were evaluated: a sample of 10,000 sites within the SWAFR (Yates *et al.*, 2010; and Fig. 2) and a sample of 20,000 sites across the whole of Australia. The larger number of sites across Australia was used to ensure good representation of all environments, based on previous tests of the effects of background sample size on model structure for these predictors (J. Elith, unpubl. data). Because the covariate data for this study are unprojected, these samples were weighted according to cell area (see methods in Appendix S4) but otherwise random.

Using random sites within the floristic region implies that the presence records are a random sample from all locations where the species is present in the region, which is unlikely because records were from extant vegetation patches in likely suitable environments (the region has been extensively cleared for agriculture, and some of the more inland areas are too arid for many *Banksia* species). Using random sites across Australia implies the species could have dispersed anywhere across the continent, and the whole continent considered available for sampling. This is questionable because the desert areas to the north and east of the inhabited area are likely barriers to dispersal. We will come back to implications of this later.

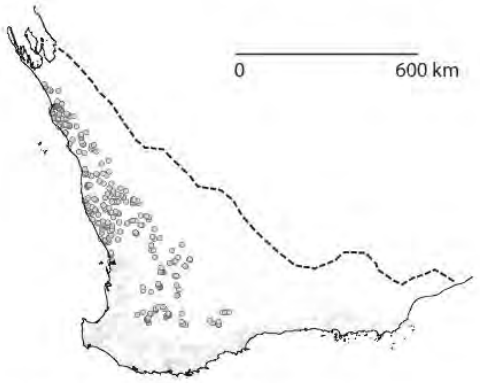

Figure 2 All Banksia Atlas sites (grey) with occurrences of *Banksia prionotes* in grey circles.

Yates *et al.* (2010) identified important climatic drivers for plants of southwest Western Australia. We base our candidate set of predictors on their study, but use a different data source so we can train and predict over the whole of Australia. Described in Appendix S4, our covariates (all unprojected, at 0.01 degree or approximately 1 km grid resolution) included five climate variables: isothermality (ISOTHERM), mean temperature of the wettest quarter (TEMPWETQ), mean temperature of the warmest quarter (TEMPWARMQ), annual precipitation (RAIN) and precipitation of the driest quarter (RAINDRYQ), and an estimate of the solum plant-available water-holding capacity (SOLWHC). We present this as a demonstration study only, and recognize that for rigorous application in this region, better soils data and predictors representing land transformation are needed for more precise predictions (Yates *et al.*, 2010). The future environment was represented by changes predicted under the A1FI scenario for 2070 estimated over the ensemble of 23 GCMs in IPCC AR4 (Solomon *et al.*, 2007); the SOLWHC was assumed to remain as it is now.

Models were fitted and projected to both current and future climates (Fig. 3) using only hinge features, with default regularization parameters (see Appendix S5 for model details, and for a comparison with models fitted with all feature types). We fitted all models on the full data sets but also used 10-fold cross-validation to estimate errors around fitted functions and predictive performance on held-out data. The latter is a good test for each model but – given the different backgrounds – not comparable across models. Note also that the AUC in this case is calculated on presence vs. background data (Phillips *et al.*, 2006). To compare the models on consistent data, we also divided the atlas data into training and testing sets for a

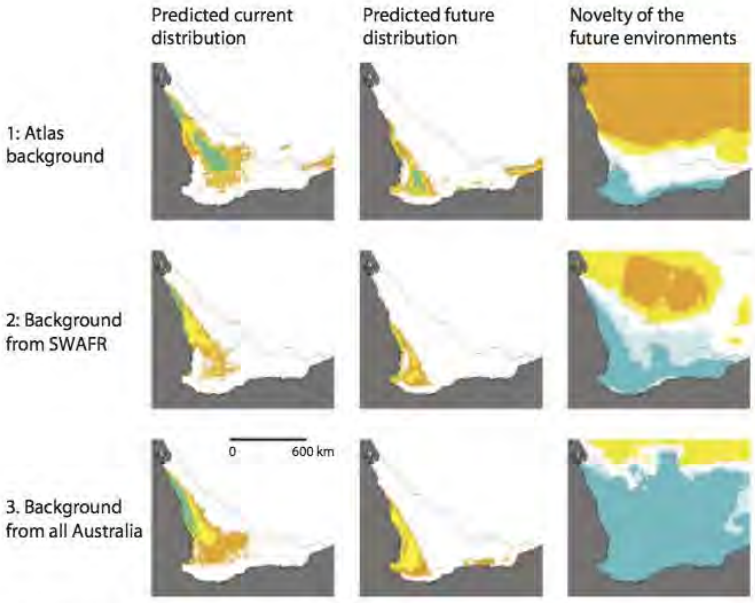

Figure 3 Model results for case study 1, showing for the three data sets (in rows): predicted current and future distributions, and extent of extrapolation compared with the training data. Predicted distributions are logistic outputs, from low values (white, 0–0.2) through orange, yellow, green to blue (0.8–1.0). For extrapolation maps, warm colours indicate extrapolation is occurring, with orange the most extreme. Grey indicates the ocean.

manual 5-fold cross-validation, testing each model on identical withheld data via two test statistics (area under the receiver operating characteristic curve (AUC), and correlation, COR; details in Appendix S4). Example code for running such analyses are available online (Appendix S4).

Results

Atlas background (model 1) produced a mapped distribution in the inhabited region with more of an eastward emphasis compared with other background treatments (Fig. 3). The coastward (westerly) bias in the distribution of survey sites (Fig. 2) affected the distributions predicted by models 2 and 3 (random background across SWAFR or Australia) but was factored out by using atlas background (model 1). The more easterly distribution is more consistent with the known ecology of the species and with the observed distribution (Taylor & Hopper, 1988). Variable importance varies with data set, with TEMPWETQ being much more prominent when using an all-Australia background than when restricted to the south-west. Similarly, shapes of fitted functions vary across data sets (Appendix S5). This is to be expected, because each data set implies a different modelling question (e.g., the all-Australia background asks: why is this species only in environments occurring in the southwest?).

An increasing number of SDM applications involve prediction to new environments (e.g., to new places or times; Elith & Leathwick, 2009a). These are contentious applications, making strong assumptions (Dormann, 2007) and usually requiring prediction to environments not sampled by the training data. MaxEnt has been extended to include new capabilities to inform users about predicting to novel environments (Elith *et al.*, 2010). MaxEnt already provides mapped information on the effect of model "clamping" – i.e., the process by which features are constrained to remain within the range of values in the training data. This identifies locations where predictions are uncertain because of the method of extrapolation, by showing where clamping substantially affects the predicted value. We feel that extreme care should be taken whenever extrapolating outside the training, so new calculations ("MESS maps", i.e., multivariate environmental similarity surfaces) display differences between the training and prediction environments (Fig. 3). In this case they show that compared with environments at the atlas sites, the northern parts of the SWAFR will experience novel climates in 2070 (Fig. 3 model 1). Models based on random background across SWAFR or the continent (models 2 and 3) require less extrapolation (because wider sampling of background points brings with it wider sampling of environments) but, given the problems with the realism of these treatments, we do not view the result as a necessary advantage for future predictions.

Appendices S5 and S6 include further information on how these models predict across the continent, for both current and future climates. They provide interesting insights into model variation across scales, regions, and datasets, and emphasize the importance of choice of background (see commentary, Appendix S5). In particular, it is interesting that model 3 restricts predictions to the correct general area and

Table 3 Variable importance and evaluation statistics for case study 1. Variable names and abbreviations for evaluation statistics are consistent with the text.

Model (background)	Variable importance						AUC (10fold CV but varying data sets)	AUC; COR (5fold CV on atlas data)
	RAIN DRYQ	RAIN	TEMP-WARMQ	TEMP-WETQ	ISO-THERM	SOL-PWHC		
1 (atlas)	57.9	30.7	7.9	0.4	1.1	2.0	0.92	0.96; 0.62
2 (southwest)	45.3	35.4	4.7	3.4	9.9	1.4	0.90	0.93; 0.52
3 (Australia)	19.7	17.7	5.3	54.0	3.0	0.3	0.99	0.91; 0.45

has the highest 10-fold cross-validated AUC (Table 3), yet has the poorest ecological justification for its choice of background and is least likely to be useful for managing the species locally. The advantage of limiting background to local, reachable areas (models 1 and 2) is that contrasts between occupied and unoccupied environments in the local area are the model focus, and – particularly with fine-scale environmental data – differentiation useful at the management scale might be achievable. It is also likely to be the most ecologically realistic choice for many locally restricted species. On the other hand, if models are to be projected well outside the local geographic area, use of local backgrounds brings with it the penalty that prediction to other areas is likely to involve considerable extrapolation. Some trade-off is clearly required.

Case study 2: Modelling the distributions of fish in rivers

This analysis predicts the current distribution of *Gadopsis bispinosus*, the two-spined blackfish, in rivers of south-eastern Australia. In the preamble, we make a case that with presence and background data, we can model the same quantity as with presence-absence data, up to the constant $\Pr(y = 1)$. One implication of that is that we should be able to use the same types of data, including fine-scale, detailed information, to model ecological relationships – i.e., we need not be restricted to coarse grid cells and basic climate variables. Here, we use detailed ecological information at the river segment scale to model the distribution of a native fish species. To our knowledge, it is the first example using MaxEnt with vector (river segment) data.

Gadopsis bispinosus is a native freshwater fish endemic to south-eastern Australia. It occurs in cool, clear upland or montane streams with abundant in-stream cover. It is most common in medium to large streams that are deep enough for reduced stream velocities and in forested catchments with relatively small sediment inputs (Lintermans, 2000).

Methods

The species data are from surveys (described further in Appendix S7) of the inland-draining rivers of northwest Victoria, Australia. In this area, there are ten major river systems grouped into four regions that start in hilly to mountainous terrain and drain northwards. *G. bispinosus* was recorded at 255 sites. We use covariate data from the 255 capture sites as our sample of L_1 and a random sample of 10,000 of the approximately 240,000 river segments for our sample of L, the background data.

The candidate predictor set comprised 20 variables summarizing information across three hierarchically nested spatial scales (segment, immediate watershed and entire upstream catchment area) and also downstream to the large river system draining to the ocean. The environmental variables estimate climate, river slope, riparian vegetation and catchment characteristics (Appendix S7). River system was also included to quantify spatial variation in land characteristics and disturbances not covered by the environmental predictor set.

These segment-based (non-gridded) data are modelled using the SWD (samples-with-data) format in MaxEnt – this involves presenting spreadsheet-like summaries of environ-

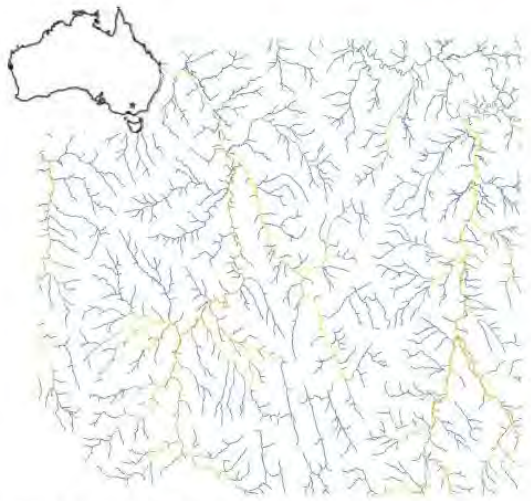

Figure 4 Predicted distribution of *Gadopsis bispinosus*, showing logistic output predictions from MaxEnt. Legend: predictions in equal intervals from 0 to 1, from blue (low) through green – yellow –orange (high). Scale: east to west the rivers map spans 45km. The star on the inset shows location.

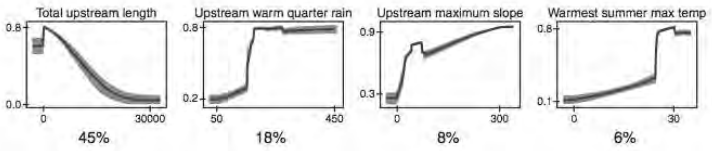

Figure 5 Partial dependence plots showing the marginal response of *Gadopsis bispinosus* to the four most important variables (i.e., for constant values of the other variables), with variable importance below each graph. The y-axes indicate logistic output.

ments at both presence and background sites. All environmental variables were continuous except the categorical river system covariate. Default settings for features and regularization were used for model training, and 10-fold cross-validation was used to obtain out-of-sample estimates of predictive performance and estimates of uncertainty around fitted functions. For mapping, the model was projected to a selected area in the Goulburn-Broken catchment. Technically, this was achieved by projecting to SWD format data, then linking the predictions to the relevant river segments in a GIS. Appendix S8 includes data and code for replicating this case study, including information on how to run MaxEnt from batch files.

Results

Consistent with ecological knowledge about the species, the model predicts *G. bispinosus* will most frequently occur in the larger streams of montane areas (Fig. 4). These locations are identified as those whose upstream catchments have relatively more precipitation in the warmest quarter and steeper maximum stream slopes. Amongst these, emphasis on segments with warmer summer maximum temperatures served to exclude the higher elevation cold streams (Fig. 5). Jackknife tests of variable importance help to identify those with important individual effects; the three most important single predictors were the summed length of all upstream links (TOTLENGTH_UCA), the upstream maximum slope (US_MAXSLOPE) and the amount of riparian tree cover upstream (UC_RIP_TRECOV); and the predictor with the most information not present in the other variables is the segment-based maximum temperature of the warmest month (MAXWARMP_TEMP). Many predictors had small to minimal impacts in the final model. The model shows strong discrimination on held out data, with a cross-validated AUC of 0.97.

Extensions/alternatives

Since records on one river system might share a more similar environment than those on different systems, an alternative approach to cross-validation would be to test the predictions iteratively on held-out rivers. We chose not to do it in this case, because presence records were concentrated in relatively few river systems, so the training sets would be substantially reduced, and the test sets, relatively few.

CONCLUSIONS

Here we have described MaxEnt from a statistical viewpoint, showing that the model minimizes the relative entropy between two probability densities defined in feature space. An understanding of the model leads naturally to recommendations for implementation, and ours included the importance of providing appropriate background samples, of dealing with sample biases, and of tuning the model – through feature type selection and regularization settings - to suit the data and application. Presence-only data are a valuable resource and potentially can be used to model the same ecological relationships as with presence-absence data, provided that biases can be dealt with and except for the non-identifiability of prevalence.

MaxEnt is regularly updated, usually to include new capabilities to suit the expanding applications, and also sometimes to change the program defaults to those most often used in practice. Recent new capabilities include the cross-validation and MESS maps (i.e., estimates of how the environmental space in predicted times and places compares with that of the training data) demonstrated in case study 1. In addition, new clickable maps allow users to interrogate predictions spatially, providing information for any grid cell on the components of the prediction (i.e., what contributes to its particular value) and where the environmental conditions "sit" on the fitted functions. Maps of limiting factors show the variable most influencing the prediction for every grid cell (Appendix S6). For further details, see Elith *et al.* (2010) and the most recent online tutorial (http://www.cs.princeton.edu/~schapire/maxent/). SDMs can provide useful information for exploring and predicting species distributions, and we are keen to see their continued development and use for learning about and conserving the world's biodiversity.

ACKNOWLEDGEMENTS

J.E. was supported by an Australian Research Council grant, FT0991640 and by an early consultancy that raised the question of how to explain MaxEnt to end-users (Jeff Tranter, Environmental Resources Information Network, Canberra, Australia). T.H. was partially supported by grant DMS-1007719 from the U.S. National Science Foundation. Simon Ferrier, John Baumgartner and Tord Snäll provided useful feedback on ideas and/or the manuscript. Robert Hijmans