Biodiversity Informatics, 15, 2020, pp. 11-54

CAN ECOLOGICAL INTERACTIONS BE INFERRED FROM SPATIAL DATA?

Christopher R. Stephens[1,2,*], Constantino González-Salazar[1,3], María del Carmen Villalobos-Segura[4] and Pablo A. Marquet[1,5]

[1]*C3—Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico;* [2]*Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico;** [3]*Departamento de Ciencias Ambientales, CBS Universidad Autónoma Metropolitana, Unidad Lerma; Estado de México, Mexico;* [4]*Laboratorio Ecología de Enfermedades y Una Salud, Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Mexico City, Mexico;* [5]*Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile; Instituto de Ecología y Biodiversidad (IEB), Santiago, Chile; and The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 8731, USA*

Abstract. The characterisation and quantification of ecological interactions, and the construction of species distributions and their associated ecological niches, is of fundamental theoretical and practical importance. In this paper we give an overview of a Bayesian inference framework, developed over the last 10 years, which, using spatial data, offers a general formalism within which ecological interactions may be characterised and quantified. Interactions are identified through deviations of the spatial distribution of co-occurrences of spatial variables relative to a benchmark for the non-interacting system and based on a statistical ensemble of spatial cells. The formalism allows for the integration of both biotic and abiotic factors of arbitrary resolution. We concentrate on the conceptual and mathematical underpinnings of the formalism, showing how, using the Naive Bayes approximation, it can be used to not only compare and contrast the relative contribution from each variable, but also to construct species distributions and niches based on arbitrary variable type. We show how the formalism can be used to quantify confounding and therefore help disentangle the complex causal chains that are present in ecosystems. We also show species distributions and their associated niches can be used to infer standard "micro" ecological interactions, such as predation and parasitism. We present several representative use cases that validate our framework, both in terms of being consistent with present knowledge of a set of known interactions, as well as making and validating predictions about new, previously unknown interactions in the case of zoonoses.

Keywords— Ecology, Naive Bayes, Spatial data mining, Inference, Interaction, Biotic interactions, Distribution modelling.

Introduction

Darwin's entangled bank analogy is an adequate pictorial representation of the complexity of interactions that occur in ecological systems. Their inference and characterization have been a recurrent theme and a vexing problem in ecology, and one where theory has usually been ahead of empiricism. The seminal work of Alfred Lotka and Vito Volterra provided a theory for competition and predation that has been experimentally tested (Gause, 1934), refined (Arditi and Ginzburg, 1989; Gilpin and Ayala, 1973; Holling, 1959) and expanded in many directions, including multi-species communities (Case, 1990; Gilpin, 1975; Wilson et al., 2003). Empirical analyses of species interactions have progressed at a slower pace. Earlier work tried to estimate interaction strengths using simple measures of resource overlap as a proxy for competition coefficients (MacArthur and Levins, 1967), or estimate them using simple regression of the abundance of pairs of species across space or time (Crowell and Pimm, 1976; Schoener, 1974). These methods, however, have been largely abandoned as they make strong assumptions (Abramsky et al., 1986; Dayton, 1973; Rosenzweig et al., 1985).

**stephens@nucleares.unam.mx*

Another method, inferring interactions from species co-occurrence data, which is the focus of this contribution, has survived the test of history and provides, as we aim to show, a valid alternative. The question of whether or not, or to what degree, ecological interactions can be identified and characterised using spatial data and, in particular, co-occurrence data, has had a controversial history (Connor and Simberloff, 1979; Connor et al., 2013; Freilich et al. 2018; Morales-Castilla, 2015; Pollock et al., 2014; Royan et al., 2016). There have been multiple perspectives and opinions with respect to the hypothesis and multiple methodologies used to investigate it (Araújo et al., 2011; Araújo and Rozenfeld, 2014; Borthagaray et al., 2014; Cazelles et al., 2016; Clark et al. 2017; González-Salazar et al., 2013; Mohd et al. 2017; Pollock et al., 2014; Stephens et al., 2009)[1]. However, after more than 40 years there is still no consensus. Over the last ten years a methodology has been developed, both related to and distinct from others, that has been used to identify, characterise and quantify ecological interactions (González-Salazar et al., 2013; Stephens et al., 2009; Stephens et al., 2019), both in terms of being consistent with known interactions as well as predicting previously unknown ones (Berzunza-Cruz et al., 2015; Rengifo-Correa et al., 2017; Stephens et al., 2016). In this paper we will present an overview of the conceptual and theoretical underpinnings of this methodology and illustrate its utility using several representative use cases.

That the framework has not been more widely seen or adopted in the ecology community is perhaps linked to the belief that point collection data, in particular, is not capable of identifying, characterising and quantifying ecological interactions (Morales-Castilla, 2015). However, it has been widely accepted that such data are sufficient to characterise the spatial distributions and corresponding niches of taxa when the niche variables are restricted to abiotic variables, linked to the fundamental niche (Peterson et al., 2011). It has not been generally accepted, however, that such data can be used to model the effects of biotic factors as niche variables (Peterson et al., 2018). From a modelling point of view this is, of course, somewhat jarring—that it is fine to represent a class variable (the species you want to model) using a certain data type, but not to represent the predictors (the species that are potential niche variables) with that data type. Specially so when we know that no species exists in isolation of others, and that the occurrence of a species in a given place is the result of the interaction between physiological tolerances, interactions with other species, historical effects and dispersal limitation (Soberón and Peterson, 2005). Thus, it is questionable that niche models are really able to obtain a representation of the fundamental niche of a species without having a way of assessing the relative importance of biotic and abiotic factors in accounting for the presence of species across sites (Soberón and Nakamura, 2009).

Historically, species co-occurrence analysis has been central to community ecology theory (Diamond, 1975; Ovaskainen et al., 2010), where it was used to test whether a set of species co-occur more or less than would be expected at "random", where the question of what is random has also had a controversial history (Colwell and Winkler 1984; Gotelli, 2000). Thus, if co-occurrence patterns over the whole set deviate from the random benchmark, it has been interpreted as evidence that a structural aspect of the community is driven by biotic interactions (Brown et al., 2002; Diamond, 1975). Although this conclusion has been challenged (Connor and Simberloff, 1979; Gotelli and McCabe, 2002), a generally accepted idea is that signals of species interactions can be inferred from survey data at local scales (Gotelli et al., 2010). However, currently, an emerging issue in ecology and biogeography is to understand the interplay between the geographic distribution of species and their interactions at macro-scale levels (Aragón and Sánchez-Fernández., 2013; Gotelli et al., 2010). A commonly accepted idea is that climatic variables (Grinellian niche) are the main determinant of the geographical distribution of species, whereas biotic variables (Eltonian niches) operate at local scales, and their influence at large scales can be disregarded (Eltonian Noise Hypothesis) (Soberón and Nakamura, 2009). Consequently, biotic interactions are often neglected in spatial modelling.

Despite growing evidence that biotic interactions may determine the distribution of species (Alvarez-Martínez et al., 2015; Godsoe and Harmon, 2012; González-Salazar et al., 2013; Heikkinen et al., 2007), the debate remains open as to whether they should be considered in ecological niche modelling, and, if so, how should they be quantified? Of course, including biotic variables in spatial modelling opens up several important theoretical and methodological

[1]Some, such as Pollock et al., (2014) and Clark et al., (2017), use an approach whereby biotic factors are modeled jointly using abiotic factors as niche variables rather than as predictors themselves as in our approach.

issues. For instance, which biotic variables should be included? In general, the number of potential interactions for a single species will be much more than the known ones. Therefore, we need a framework that allows one to include different types of data (e.g., collection points, environmental layers) in order to infer, compare and contrast potential interactions.

We believe that an important barrier to making further progress is that of developing and agreeing on a deeper and more quantitative understanding of what an "interaction" is, at least in the context of patterns of co-occurrence, and how interactions can be manifested at different spatio-temporal scales. In our methodology, an interaction is *defined* by quantifying the degree of co-occurrence of variables—biotic or abiotic—relative to that expected in the absence of the interaction. In these simple terms the underlying *modus operandi* is no different than the original motivations of Diamond (1975), where it was hoped that co-occurrence data could be used to reflect inter-specific competition. However, the logic is quite universal across all areas of science—physics, chemistry, linguistics, genetics, epidemiology—interactions always affect the positions of the objects that interact. The chief difference between the different disciplines is what objects are interacting and how should co-occurrences be defined so as to characterise the interactions?

In the case of ecology, the use of point collection data for determining co-occurrence, and the interpretation of the associated analysis to infer biotic interactions, has been controversial and, especially recently, has generated many papers—see, for instance, (Wisz et al., 2013) for a recent discussion. Although our characterisation of interaction is definitional, it is important to determine to what degree such a characterisation captures the intuition associated with the standard classification of ecological interactions—such as predation, mutualism, commensalism etc. The latter are associated with the relative impact of the interaction on each participant—positive for the predator, negative for the prey for example—and are linked to a specific set of "labels" that mark each participant, such as predator = yes/no, prey = yes/no, with each interaction being associated with a particular label. These ecological interactions are "micro" interactions, in that they are most manifest at the level of individuals, such as predation events, where a bobcat kills and eats a rabbit for example. As all of these interactions are local and direct, they should all be amenable to an analysis in terms of a statistical ensemble of suitably defined co-occurrences. However, they offer a rather poor representation in terms of predicting the "macro" distributions that are an emergent property of the micro interactions, where by macro we mean how the relation between the spatio-temporal distributions of the bobcat and the rabbit as species are affected by these micro interactions. One reason why they offer a poor representation is because the macro distribution of a species depends on many labels, the majority of which are unknown. For instance, just how many species are hosts of a given zoonosis but are unknown as such? Is the label of "prey" sufficient to characterise a predator-prey interaction at the macro level? What about the potential relevance of other labels, such as adult/young, male/female, large/small, strong/weak, fast/slow? All of which may be relevant to quantifying the relative success of the predator and therefore its spatio-temporal distribution.

Effectively, as in many other areas of science, we are pointed in the direction of attempting to deduce and relate interactions at a micro scale to those at a macro scale, in the knowledge that the macro scale emerges at the collective level from the combined effect of very many micro events. A vital link between the two scales is the concept of a niche. If a species is an important niche variable for another, then, by definition, it favours the presence of the species and thereby affects its spatial distribution (Giannini et al., 2013). However, a niche dimension also captures an intuition as to why at the micro level it is a niche dimension. Thus, we can accept that a predator-prey interaction is the underlying cause of the fact that presence of the prey species is a niche dimension for the predator and affects its distribution. However, it does not have to be so. Individual predation events may, in fact, be due to completely random encounters between predator and prey. This, in turn, would then leave no imprint at the macro level. We could not then speak of the prey as being an important niche variable of the predator, in spite of the fact that there existed micro-level interactions between the two. We have natural selection to thank for the fact that this type of situation would be the exception rather than the rule. A predator that captures prey randomly would soon be out-competed by a predator that is better adapted.

Although micro-scale interactions are potentially easier to measure than macro-scale ones, there are just too many to measure. For n species we may

imagine that there are $n(n-1)$ inter-specific interactions. However, it is much worse than that, as for any pair of taxa there are potentially as many interactions as they have relevant labels. It is because of this that macro-level interactions are potentially more amenable to analysis as, using a proxy such as point collection data, we can compute the deviations from a given null hypothesis for any pair of taxa from among a vast number. The relevant question then goes in reverse: instead of trying to characterise macro-level interactions as emerging from underlying micro interactions we can try to deduce properties of the micro interactions from the macro level data. Once again, it is the concept of a niche that gives hope to this endeavour as it links the two levels.

Of course, a macro-level interaction may be a result of many different types of micro interaction, thus leading to confounding. This is no different than in many other areas of science. The spatial distribution of disease is, epidemiologically speaking, a result of many underlying micro interactions. However, many of these micro interactions may not be manifest and our understanding is first pointed to relate the spatial distribution of disease with the spatial distribution of risk factors (niche variables). Subsequently, one may then attempt to give a micro explanation to the macro relations or vice versa. Confounding potentially plays an important role in ecology, where it has been suggested that apparent biotic interactions may be confounded by abiotic factors (Purse and Golding, 2015), which can lead to the right prediction for the wrong reason (Dayton, 1973). The Bayesian formalism we have developed allows for a detailed investigation of this phenomenon.

Of course, there is a fundamental question as to whether a set of spatial data gives explicit information about an interaction versus information from which an interaction is to be *inferred* and potentially characterised. The vast majority of spatial data that is available has not been generated with the specific intention of analysing a particular interaction. Rather, the data is used to *infer* the existence and nature of an interaction using a suitable mathematical framework for making statistical inferences. Bayesian inference (Berger, 1985) provides an appropriate framework for this task, where Bayes' theorem is used to update the probability of a hypothesis, such as the existence and nature of an interaction, as more evidence or information becomes available. This is particularly appropriate in our cases of interest where we can deduce more information about the interaction by including in more spatial information.

Although we will try to couch much of our discussion in general terms, our main concern is to apply these ideas to ecological interactions and, particularly, in the context of niche descriptions. Of course, the use of co-occurrence data in ecology has a long history, with the particulars depending on whether we are talking about abiotic or biotic interactions. In the case of climatic data, species distribution modelling has used the co-occurrence between a point collection of a target species and the specific environmental conditions at that point, the latter being modelled as environmental layers at the pixel level (Elith and Leathwick, 2019; Peterson et al., 2011). Many different algorithms have been used to model the relation (Qiao et al., 2015).

This paper is based on a methodological framework (González-Salazar and Stephens, 2012; González-Salazar et al., 2013; Sánchez-Cordero et al., 2008, Sierra and Stephens, 2012; Stephens et al., 2009) that has been developed to determine, characterise and quantify ecological interactions of any type, abiotic or biotic, using data of arbitrary spatial resolution. It allows one to characterise the full ecological niche of a taxon, data permitting, while comparing and contrasting the contribution of each niche variable. The formalism has been applied successfully, chiefly in the area of zoonoses, where it has led to the prediction and confirmation of many previously unknown vector-host interactions in several emerging or re-emerging diseases (González-Salazar et al., 2017; Rengifo-Correa et al., 2017; Stephens et al., 2016). In spite of its success in this important application area, it is not well known in a wider context and, importantly, its conceptual underpinnings and its general applicability to the general area of identifying and characterising ecological interactions have not been exploited. As a complement to these papers, we will here concentrate on its conceptual and mathematical basis and, in particular, how it can be used to infer causal chains and identify confounding factors in any given ecological setting and to predict micro ecological interactions.

The format of the paper is as follows: In the second section, we discuss interactions in a wider context, as it is important to see that relating micro interactions to macro interactions permeates all of science. Fundamentally, there is nothing different between doing so in physics versus in ecology, other

than ecology is much more complex in terms of the number of different types of interaction and the large array of factors that characterise them. What links them all is that interactions lead to a different spatial distribution of the objects that are interacting than would be the case in the absence of the interaction. Then, we present the empirical definition of an interaction that has been at the root of our efforts—that an interaction can be identified using statistical ensembles of spatio-temporal data to show that in the presence of the interaction the spatial distribution of the members of the ensemble is different to that in the absence of the interaction. Such a notion is universal in science, the most fundamental formulation being associated with Newton's laws of motion, where forces (interactions) can be identified from the spatio-temporal trajectory of an object relative to the null hypothesis that all objects not subject to an interaction are at rest or in inertial motion. In order to not distract the reader, we include discussion of co-occurrence and interaction in several text boxes that can be read separately and independently of the main text. Next, we discuss how the notion of a co-occurrence can be used as a fundamental variable for distinguishing between interacting and non-interacting systems. We also show how co-occurrences can be compared and contrasted between variables that have radically different spatial resolutions and also different data types by making all variables, abiotic and biotic, binomial and bringing them all to the same spatial resolution. Next, we discuss the Bayesian modelling framework that is the heart of our methodology. We show how, based on the fact that we can bring all variables to the same type and resolution, we may compute the relative weight of any variable to the probability to find a given taxon, thereby determining its importance as a niche variable and simultaneously as a determinant of the spatial distribution of the species. Moreover, we show how the formalism can be used to compare and contrast the relative degree of confounding of one variable and another, showing how this permits us to begin to disentangle the complex causal chains that exist in ecological systems. In particular, we will be able to show that, generally, biotic factors are confounders for abiotic factors, not vice versa. Further, we discuss the relation between micro and macro interactions, showing how, and under what circumstances, micro interactions may be inferred from macro data. In addition, we present several use cases to give ample support to all the assertions previously made. Finally, in the last section we draw some conclusions.

What are interactions and should we be able to characterise them through spatio-temporal data?

The most general notion of interaction across the sciences is simply that one thing affects another—mutually—with the main differences being what we mean by "thing" and what we mean by "affect". In physics, for example, the presence of one electrically charged particle affects the presence of another, and vice versa. In ecology, the presence of a predator affects the presence of a prey, and vice versa. At a fundamental level all interactions are local, i.e., the interacting entities are located at the "same" place at the "same" time[2] and therefore *co-occur*. Interactions, if they can be characterised, are given names: Electromagnetism, gravity, predation, parasitism etc. What they have in common is that the presence of one element in the system affects the state of the other and, again, vice versa. However, in each case the state variables that are affected by the interaction may be quite different. For example, in the case of predation the most important state change of the prey is living to dead, while a state change of the predator is, for example, a transition from hunger to satiation. We believe there is value in understanding in more detail how co-occurrence and interaction are related in sciences other than ecology. However, so as to not distract from the main text we have separated this discussion into boxes.

An individual interaction may, in principle, be directly observable, as is often the case in ecology. This requires that the interaction is directly characterizable in terms of measurable variables. For instance, predation is an interaction that may be observed directly given the abrupt change in state variable of the prey: live → dead. Often, however, the interaction must be *inferred* using data and reasoning, as the change in state variable that best characterises the interaction may not be directly observable or difficult to observe. For instance, although one may actually be present at the act of predation, in other circumstances it may well have to be inferred indirectly, by, say, an examination of the faeces of the predator. Similarly, the feeding of an hematophagic insect on

[2]There are, of course, subtleties involved in defining what we mean by "same", such as the question of simultaneity in relativity or what looks like action at a distance.

a mammal may not be directly observed, but a blood analysis of the insect may reveal which species it was feeding from and allow for an inference of the corresponding interaction. In these cases, the interaction *per se* is known, but due to data considerations its characterisation must be inferred. However, there are many cases where the interaction is not previously known, where it is inferred first and then characterised and understood later. This is what happened with the fundamental interactions of physics. They were inferred from data first.

In the absence of data on the state variable changes that characterise an interaction, a state variable that is almost inevitably affected as a consequence of an interaction, especially in mobile organisms, is the position in space and time of the interacting entities relative to what they would have been in the absence of the interaction. Indeed, it may well be that, to a very good approximation, position is the only observable state variable that changes, or is readily observable due to the interaction. This absence of interaction then represents a "*null hypothesis*," with respect to which the interaction may be benchmarked.

As discussed in Box 1, in many sciences other than ecology, co-occurrence, though oftentimes not defined explicitly as such, has been used successfully to characterise interactions. So why is it that in ecology, as discussed above, the use of position data and, in particular, point collection data to deduce the nature of ecological interactions has been so controversial and, apparently, not sufficiently successful to be accepted by the wider community? We will provide an answer to this question in the following sections.

We believe that one problem with the notion of interaction in ecology is that there has been no adequate characterisation of how standardly accepted ecological interactions, such as predation, mutualism and parasitism, can be applied at the collective level. In other words, how do we characterise the "interaction" between two species, predator-prey, from knowing that there is an interaction at the level of each individual—a specific predation event? First, we need a notion of interaction that is applicable at all observation scales. Given the incontrovertible evidence from myriad disciplines that interactions lead to changes of state in the objects that are interacting and, in particular, the state variables associated with the objects' positions, which are distinct relative to the case—null hypothesis—where the interaction is absent, we will *define* an interaction to be present if the spatial distribution of the objects of study is different to this null hypothesis. This represents a purely empirical characterisation, dependent on the null hypothesis chosen, but which makes no *a prio-*

Box 1: Interactions and co-occurrence outside Ecology

Seen from the perspective of other sciences, the answer to the question as to whether interactions are characterizable through spatio-temporal data, is an unequivocal yes; the reason being that we have been doing it successfully in many scientific disciplines for centuries. In physics, where the notion of interaction has been most precisely quantified, all the principal fundamental interactions—gravity, electromagnetism, strong and weak nuclear forces—have been identified and characterised by observations of the relative positions, as a function of space and time, of objects, such as planets, electric charges, nucleons etc. The enormous success of this endeavour has been due, in large part, to the fact that each interaction is characterizable in terms of a very small number of parameters. Among these are "labels" for the objects, such as mass and electric charge, as well as universal constants which are measures of the strength of the interactions.

The characterisation of interactions through the observation of the positions in space and time of the interacting objects is not restricted to the fundamental interactions. In atomic and molecular physics and chemistry, for example, effective interactions that emerge from the underlying fundamental interactions can also be characterised by the positions in space and time of different types of object—atom, molecule, macromolecule, planets etc. In comparison with the fundamental interactions, which are all *direct* and *local*, all interactions in physics and chemistry are, by definition, *indirect*. Thus, chemical interactions, such as hydrogen bonding, covalent bonding, Van der Waals forces etc. are all emergent, indirect interactions. However, they are generally considered to be "direct". Firstly, because the first principles derivation of this indirect interaction from the underlying fundamental direct interactions is too complicated to carry out and, secondly and importantly, a better understanding of how the interaction was mediated would not necessarily help us to better understand atomic physics, molecular physics or chemistry. Thus, the question of to what degree it is convenient to characterise an interaction as indirect versus direct is to a large extent

a question of convenience. For instance, is it important to understand the nature of any intermediation, or are the intermediating states readily observable?

An important requisite for identifying the presence of an interaction using observations of the positions of the involved objects is that one must have, a priori, a notion of what those positions should be in the case of a non-interacting system. In the case of physics this is enshrined in Newton's first law—that an object will remain at rest or in uniform motion in a straight line unless acted upon by an external force. Thus, deviations from this null hypothesis serve as a *definition* of the presence of a force, i.e., an interaction. With this null hypothesis goes the idea that in the non-interacting state the degree of co-occurrence of two objects should be different to the case when they interact. Thus, the fact that electrons in a given atom co-occur in space and time with the nucleus of that atom, relative to the null hypothesis that they are independent of the nucleus, is an indication of the existence of an interaction—the electromagnetic attraction between negatively charged electrons and positively charged nucleus. The fact that the planets in the solar system co-occur in space and time with the sun, relative to the null hypothesis that they are independent of the sun, is also an indication of the existence of an interaction—the gravitational attraction between sun and planets.

Note that any comparison between the observations of a system and the null hypothesis must be done at the statistical level, where an appropriate statistical *ensemble* of observations must be formed. One single observation is not sufficient. For example, Tycho Brahe's observations of the regularities in the dynamics of the planets led to Kepler's phenomenological laws. Kepler could not have deduced those laws from just one entry in Brahe's notebooks. An ensemble of entries as a function of time was required. Later, Newton, using the null hypothesis of Galileo that bodies not subject to an interaction (force) stay at rest or in uniform motion, deduced that planets are subject to an interaction as they do not follow that null hypothesis. Newton's laws of motion, along with Kepler's observations, allowed that interaction to be characterised, deducing that it depends on the masses of the interacting bodies and is weaker ($1/r^2$) as a function of their separation. This is probably the clearest example of the logic of identifying and deducing the nature of interactions through observations of object positions. The statistical ensemble of observations in this case was the set of positions on the celestial sphere across time of the interacting objects.

Of course, understanding interactions through examination of the relative positions of objects is not restricted to physics and chemistry. As discussed in Stephens et al., (2017a), in standard population genetics, where genes are viewed as beads on a string, the concept of interaction is associated with the notion of epistasis (Phillips, 2008), where, in this setting, the degree to which two genes are linked, i.e., they *co-occur*, can be used as a measure of such epistasis. In other words, we measure interaction by to what degree two genes actually co-occur relative to their expected distribution if they were independent—the no-interaction null hypothesis. These genetic interactions may also have differing degrees of directedness. For example, it may occur that two genes are linked, where the linkage is not direct but through the intermediation of a third gene with which the two are directly linked. Similarly, in text mining, syntactic and semantic interactions can be deduced from the co-occurrence of textual elements, such as words or phrases or other linguistic objects. It has also long been used in epidemiology. Indeed, perhaps the founding event of modern epidemiology, the analysis of John Snow of the Broad Street Cholera outbreak of 1854 (Snow, 1855) was based on a co-occurrence analysis, where the positions of disease cases and potential disease sources were mapped and an interaction—that the events were clustered around a certain water pump as opposed to being randomly distributed—was identified.

Although the concept of interaction, especially in physics, is naturally tied to an ensemble of observations in space and time this is not a prerequisite. In the absence of temporal data, we can and must infer interactions using only a spatial ensemble. In the case of genetics or language, for example, the ensemble is a specification of the positions of genetic objects—genes, exons, nucleotides etc.—or syntactic objects—nouns, articles, verbs etc.—in an ensemble of such objects, such as genomes or texts. The logic, however, is identical to that of Brahe-Kepler-Newton: that relative to a suitable null hypothesis, the spatial distribution of these objects is significantly different. Thus, articles precede nouns in English and not vice versa and an analysis of texts will identify this grammatical "interaction" relative to the null hypothesis that articles and nouns are randomly distributed. Thus, in all cases, the question of whether a given spatio-temporal distribution of objects is distinct to our null hypothesis is a question of statistical *inference*. What we may infer about an interaction is then very much related to the precise nature of the ensemble we use.

ri reference as to the nature of the interaction or its properties. We will usually think of the interaction as binary, in that it relates to two types of object. However, as we will see, we may readily extend the notion to multiple types of object and thus capture the idea of the interaction between an object and its "niche" or environment.

We might believe that we can observe an interaction using only a single observation. As discussed in Box 1, however, this is not true. The comparison between the observations of a system and the null hypothesis must be done at the statistical level, where an appropriate statistical *ensemble* of observations must be formed. For example, in the case of, say, predation, the actual act, observationally, requires an ensemble in time. Before the actual predation, both the predator and prey would be described by position coordinates changing as a function of time. However, after the act of predation, the position coordinates of only the predator would change. Thus, a characterisation of the event requires a history (an ensemble in time)—a before and after. As emphasised, this definition of interaction is scale independent, in that we can apply it to objects at multiple resolutions. In the absence of temporal data however, we must infer interactions using only a spatial ensemble. This situation also frequently occurs in other sciences, as discussed in Box 1. In all cases, however, the question of whether a given spatio-temporal distribution of objects is distinct to our null hypothesis is a question of statistical *inference*. What we may infer about an interaction is then very much related to the precise nature of the ensemble we use.

Classifying interactions

Given our empirical definition of interaction, all we may deduce is that a set of spatiotemporal data is consistent with the presence or absence of an interaction; but this tells us nothing about the nature or properties of that interaction and therefore does not necessarily help us to understand it. To characterise the interaction we must seek parameters, or state variables, on which the interaction depends and determine if there are changes in the interaction as those state variables or labels change.

Unlike physics, as discussed in Box 2, in ecology, there are many labels that are relevant for potentially characterising a given empirical interaction,

Box 2: Classifying interactions

An important element in understanding interactions is to characterise the properties of an object that give rise to an interaction in the first place. Each such property can be associated with a "label". For instance, in physics each fundamental interaction—gravity, electromagnetism etc.—is characterised by one and only one label—mass, electric charge etc. However, the nature of those labels goes a long way towards allowing us to characterise and understand the phenomenology of the interaction. Thus, although gravity is much weaker than electromagnetism, it can manifest itself at large scales because mass—the gravitational "charge"—is always positive, whereas electric charge is positive or negative and macroscopic matter is neutral. A consequence of this is that the fundamental interactions manifest themselves at very different spatial scales, a fact which lends itself to an enormous simplification when trying to disentangle their relative effects. However, the effect of a fundamental interaction, such as electromagnetism, can become much more subtle and complicated at the collective level. For example, two atoms can repel at a small scale while attracting at a larger, molecular, scale. As the complexity of the interacting objects increases so does the potential number of labels that characterise the objects involved in the interaction. So, in molecular physics we must not only specify the atomic components but must understand the three-dimensional structure of the atoms and molecules in order to understand their interactions. In the end though, physics and chemistry are relatively simple, in that, at any given scale, there is usually one dominant interaction and correspondingly, one, or a few, most relevant labels. However, which labels are relevant can change radically from one observational scale to another.

An important property of an interaction with respect to these labels is its degree of universality. Thus, the fundamental forces are fundamental because they are completely universal, depending only on one label, in all places and at all times. They are not contextual. The gravitational force between the earth and the sun depends only on their masses and is independent of any other parameter. We can deduce this fact by observing properties that are consequences of the interaction—position in space and time for instance—and noting that predictions based on our characterisation of this interaction are independent of other labels. Thus, the force of gravity is independent of the state of the gravitationally interacting body—so that labels, such as gas giant, rock, hot, cold etc. are unimportant

for this interaction. Hence, the gravitational attraction of one single 10 kg mass is the same as that of ten 1 kg masses combined, while the electrostatic force exerted by one charge of 10 Coulomb is the same as that of ten charges together of 1 Coulomb.

The taxonomy of the fundamental interactions in physics is clear and well established. At higher levels of organisation there are also established taxonomic classifications of interactions, e.g., covalent versus ionic bonding, and labels, such as from the periodic table, that allow us to characterise and quantify the interactions. At the most gross, phenomenological level, we may speak of an interaction in terms of whether it is positive or negative (attraction/repulsion), the "charges" (labels) on which it depends, its strength and its relative importance.

such as prey/predator, male/female, parent/offspring, carnivore/herbivore, mammal/reptile, old/young, fast/slow, etc. as well as the taxonomic names of the involved species, all of which may affect the spatio-temporal distribution of the organisms. Indeed, the deviation of the spatio-temporal distribution of objects from a null hypothesis, that is at the heart of our empirical characterisation of an interaction, could be the result of potentially many distinct interaction types. Furthermore, in ecology, we do not even have a complete, accepted set of labels to use. The existence of a large number of relevant labels for a given organism, that encompass the set of possible interactions with other organisms, makes the full empirical characterisation of all its interactions by direct observation completely impossible.

An important property of an interaction with respect to these labels is its degree of universality. Thus, the fundamental forces in physics are fundamental because they are completely universal, depending only on one label, in all places and at all times. They are not contextual. In ecology, however, the degree of universality is much less. For example, for a given predator there will be a label "prey = YES/NO" associated with prey species of the predator. This label characterises the interaction and would be consistent with the classification of predation as having a positive effect on the predator and a negative one on the prey. However, the label "prey = YES/NO" is only one of many that may be relevant for characterising the interaction. For example, for one prey it may be that 70% of attempted predation events are successful, while for another it is only 10%. Also, at the collective level, such as at the species level, two different prey species may form substantially different parts of the diet of the predator. Thus, at the scale of an individual predator and an individual prey, the interaction may well be "repulsive", when viewed in terms of the trajectories of the individuals, reflecting the fact that a prey may try to avoid the predator, while at the species level, the interaction can be attractive, meaning that the predators at the collective level are attracted to where the prey species are located.

Although the taxonomy of the fundamental interactions in physics is clear and well established, where at the most gross, phenomenological level, we may speak of an interaction in terms of whether it is positive or negative (attraction/repulsion), the "charges" (labels) on which it depends, its strength and its relative importance, in ecology interactions have been classified in a somewhat different way (Lidicker, 1979; Wisz et al., 2013), based on earlier work in the social sciences (Haskell, 1949), where the characterisation of the interaction is principally based on the impact it has on the interacting taxa, where the impact is considered in terms of whether it has a "positive" (benefit) or "negative" (cost) effect (Araújo and Rozenfeld, 2014; Belmaker et al., 2015). This cost/benefit, in turn, must be evaluated in terms of some measurable function, such as reproductive success. However, this classification in no way exhausts the set of labels or other parameters that are potentially relevant for classifying the interaction. There has also been work on trying to quantify the notion of interaction strength (see for example Paine, 1992; Wootton and Emmerson, 2005), where the measures are mainly linked to experimental procedures, such as removing a species from an environment, or on mathematical models, but not as measured directly from spatial distributions.

Niches and interactions: from micro to macro

We have defined interactions as being identifiable from deviations in the spatiotemporal distributions of objects from a suitable null hypothesis. We have also emphasised that a statistical ensemble of observations is necessary. Clearly, what we may deduce about an interaction depends on the nature of those observations and what data represents them. In

particular, it depends on the scale or spatio-temporal resolution of those observations. We may consider two distinct scales—the "micro" and the "macro", though we use these as relative not absolute terms. We know, particularly in physics, as discussed in Box 3, that the nature of interactions at one scale can radically change when passing to a different scale.

In ecology, in the case of a predator-prey interaction, for example, the "micro" level would naturally correspond to observations of two individuals—one predator and one prey. Following their trajectories in space and time would allow us to determine that there is an interaction present. Moreover, labels such as "predator" and "prey" would allow us to determine its biological plausibility and characterise it. However, this ensemble of observations tells us nothing about the nature of the interaction at the collective level, at another spatial resolution—the "macro" level. In general, there is no fundamental reason why an interaction at one scale should manifest itself in an obvious and analogous way at another scale. In this sense the interaction is context (environment) dependent. Similarly, as mentioned, predator and prey may "repel" at the micro-level, in that the prey tries to avoid the predator, but may be attracted at the macro level, in that the spatio-temporal distribution of the predator species is attracted to that of the prey species. In this case, the attraction between predator and

Box 3: From the micro to the macro

Physics is the discipline where the quantitative and qualitative relations between micro and macro variables is best understood, with micro and macro being relative terms. For instance, we may consider the intra-atomic scale as being micro and the inter-atomic scale as macro. Thus, as an example, at the intra-atomic scale, there is the strong interaction between electrons and nucleus in a Helium atom, while at the inter-atomic scale there are no significant interactions between the Helium atoms themselves. A relevant label for the interaction is that of electric charge. At the micro level the electrons have a label corresponding to one unit of negative charge while the nucleus has a label corresponding to two units of positive charge. On the other hand, the Helium atoms have a label corresponding to zero charge. Of course, we fully understand how the zero electromagnetic interaction at the atomic "macro" level emerges from the underlying strong interaction at the "micro" level. Similarly, electrons repel as free particles, but they can "attract" in the case of a covalent bond. Thus, the nature of interactions at one scale can radically change when passing to a different scale, with the interactions at a more macro scale being an emergent phenomenon relative to the interactions present at the micro level.

Additionally, in physics instead of talking about the interaction between two objects we may speak of the interaction between an object and its *environment* when, for instance, the environment consists of an ensemble of objects, such as atoms, where in a solid say we may consider the interaction between an atom and its environment as represented by the ensemble of other atoms in the system. A particular atom or other structure in a solid has its "niche" in the same way as a species has its niche. The environment in both cases is the net, emergent effect of a large set of individual niche variables. Thus, in principle, we may consider interactions along a spectrum, from between two individual objects of definite types, to between an object and any conglomeration of objects that represent its niche/environment.

In physics, the fundamental interactions are direct. For example, the interaction between an electron and a proton in a hydrogen atom can be thought of as being direct, being describable directly in terms of the fundamental electrostatic attraction between the positively charged proton and the negatively charged electron. Two hydrogen atoms, though, may form a hydrogen molecule where, unlike the direct electrostatic interaction, the interaction between these atoms is "indirect" and is a consequence of the presence of other intermediating elements—the electrons. In a potential abusive use of ecological terminology, we could say that the repulsive "competitive" interaction between individuals of the proton species were turned into an attractive interaction by the "facilitation" of individuals of the electron species.

An effective, phenomenological model of a chemical bond between atoms as a "direct" interaction may well be much more useful than trying to simultaneously model the multiple, underlying, more fundamental, direct interactions between the atomic constituents. However, if we examine the molecule exhibiting the chemical bond at higher energies then these more fundamental interactions will become apparent. The important point to make is that how an interaction in a system is characterised in physics and chemistry, and the degree to which we describe it as direct or indirect, is very much a function of the scale at which we make observations of that system.

prey at the species level is an emergent phenomenon relative to the discrete predation events observed at the individual level.

In ecology, it is the concept of a niche that provides a powerful and intuitive framework for understanding emergence. That a biotic variable, X_α, is considered and observed to be an important niche component of a species, C, will have the consequence that the presence of X_α favours the presence of C. The conditional probability $P(C|X_\alpha)$ is a suitable mathematical representation of this relationship, where we will specify below what ensemble of observations is suitable for calculating $P(C|X_\alpha)$ and, in particular, the dependence on the spatio-temporal scale of the observations. However, the intuition is that micro-level interactions between C and X_α can manifest themselves at a larger scale by X_α being a niche variable for C. This concept of niche is relevant for understanding the relations between objects in any area of study.

Built into the concept of both the Eltonian (Elton, 1927) and Hutchinsonian (Hutchinson, 1957) niches is the notion that biotic interactions between a target species and other species in its niche affect the spatial distribution of the target species and vice versa. Thus, to what degree a species persists in a given geographical area is affected by its interactions with other species. Here, at the niche level, in distinction to the notion of interaction between two discrete objects, we consider the interaction between an object and its *environment*. Thus, the concept of a niche links in a profound way the notion of micro interactions to macro interactions, as defined by our concept of deviations of the spatial distribution of a taxon from a null hypothesis, with the micro interactions being the supposed drivers of the spatial distribution as explained by its niche.

Direct versus indirect interactions

The concept of direct versus indirect interaction occurs in all the examples we have mentioned. As stated, labelling an interaction as direct or indirect is to some degree a question of convenience and a question of observational scale. In ecology, determining whether an interaction is direct or indirect is difficult. This can be perfectly illustrated in the context of a simple food chain, such as: carnivore ← herbivore ← plant ← sun. One would be tempted to argue that the interaction between carnivore and herbivore was more direct than between carnivore and plant. However, a perfectly acceptable predictive model for the carnivore distribution might be built using plants as niche variables. Moreover, the principle effect of climate on the carnivore distribution will be an indirect interaction, intermediated by the plant and herbivore distributions, rather than a direct one (see for example (Rebolledo et al., 2019). We must then determine how from data we may disentangle these interactions and characterise their degree of indirectedness.

As a prelude to a later discussion ("A Bayesian framework for causal inference"), we may intuit the degree of indirectedness of an interaction between two taxa, C and X_α, by trying to determine if there exists one of more other variables, X_β, that are more directly linked to C or X_α and therefore act as confounders. Thus, if C is a carnivore, with an important prey species, X_α, and X_β is the principal plant food source of X_α then the interaction between C and X_β will be intermediated by X_α.

Co-occurrence as a measure of interaction

Having defined an interaction as being associated with a spatial distribution of two or more sets of objects that differs from a null hypothesis where the interaction is absent, we must define observable, measurable parameters that allow for a comparison between an observed distribution and the corresponding null hypothesis. An extremely useful measure is that of a *co-occurrence*.

In ecology, as elsewhere, co-occurrences are a *necessary* condition for an interaction. For a predation event to occur, the predator and the prey must be in the same place at the same time. Similarly, for pollination, or any other type of ecological micro interaction. Of course, this does not deny the possibility of non-local, action-at-a-distance type interactions that are intermediated by other variables such as climate teleconnections or nutrient transports across continents or global scale phenomena such as climate change and commercial trade. (Bradley et al., 2012; Bristow et al., 2010; Wang et al. 2000). For instance, two species may interact through an abiotic intermediary, where the interaction between the species and the intermediary is direct and local but the effective interaction between the species is non-local and indirect. However, the fact remains that any direct interaction is local.

Defining co-occurrence

Co-occurrence is a notion about discrete events happening in space and time, either at the same place,

or same time, or both. Here we will consider only a two-dimensional space. To define same place in space and time we first specify some partition of a space, $\mathcal{A}$, and time interval, $\mathcal{T}$, into cells. A natural, though not obligatory, partition of $\mathcal{A}$ is that of an array of squares of a fixed linear dimension, while that of $\mathcal{T}$ would be into fixed intervals of time. A cell i thus defines an area, ΔA and a time interval Δt. To the question of what is co-occurring we may consider variables $(X_1(x_1,t_1),X_2(x_2,t_2),\ldots,X_m(x_m,t_m))$, which, in principle, may be discrete or continuous and where $\mathbf{x}_i = (x_i,y_i)$ are the two-dimensional coordinates for geo-referencing.

A co-occurrence of any subset, $\mathcal{V} = \{X_\bullet(x_\bullet,t_\bullet),X_\beta(x_\beta,t_\beta),\ldots\}$, of these variables can then be defined by an indicator function as $\mathcal{I} = 1 \Longleftrightarrow$ all $(x_j,t_j) \in \mathcal{V}$ are observed in the cell i and is zero otherwise. Thus, $\mathcal{I}$ is a Boolean function. For example, for two variables, $C(x,t)$ and $X_\alpha(x',t')$, a co-occurrence in a cell i is such that $\mathcal{I} = 1 \Longleftrightarrow C$ and X_α are observed in the cell i, with $x,x' \in i$, and is zero otherwise. For a single variable, $X_\alpha(x,t)$, we can also use the indicator function, where now $\mathcal{I} = 1 \Longleftrightarrow$ all $(x,t) \in \mathcal{V}$ are observed in a cell i.

The set, $\mathcal{S}$, of N cells represents a statistical ensemble. We may then count events on this ensemble. For any single observable X_α, such as the presence of a species, we may count the occurrences of X_α on $\mathcal{A} \times \mathcal{T}$ as

$$N_{x_\alpha} = \sum_{x\in\mathcal{A}} \sum_{t\in\mathcal{T}} \mathcal{I}\big(X_\alpha(x,t)\big) \tag{1}$$

Similarly, the number of co-occurrences of two variables X_α and X_β is given by

$$N_{x_\alpha x_\beta} = \frac{1}{2}\sum_{x\in\mathcal{A}} \sum_{t\in\mathcal{T}} \sum_{x'\in\mathcal{A}} \sum_{t'\in\mathcal{T}} \mathcal{I}\big(X_\alpha(x,t),X_\beta(x',t')\big) \tag{2}$$

In this way, we could count multiple variable pairs within the same spatio-temporal cell. We can also count by simply counting in each cell occurrences of a given type, independently of how many examples of the type there are in the cell. In this case, the number of co-occurrences is given by

$$N_{x_\alpha x_\beta} = \sum_i \mathcal{I}\big(X_\alpha(i),X_\beta(i)\big) \tag{3}$$

and the number of occurrences N_{X_α} by

$$N_{x_\alpha} = \sum_i \mathcal{I}(X_\alpha(i),) \tag{4}$$

In the case of a purely spatial ensemble, $\mathcal{S}$ represents a set of N spatial cells and (3) and (4) are calculated over this set. Note that $\mathcal{I}$, as a Boolean function, may represent any composition of variables. For instance, we may consider $\mathcal{I}(X_\alpha(x,t),X_\beta(x',t'),X_\gamma(x'',t'')) = 1 \Longleftrightarrow ((x,t)$ AND $(x',t') \in \mathcal{V})$ OR $((x,t)$ AND $(x'',t'') \in \mathcal{V})$. For example, α could represent a species while β and γ represented two species in the same genus and we were counting co-occurrences of α with that genus. In fact, we may consider the co-occurrence $N_{X_\alpha \mathbf{X}}$, where $\mathbf{X} = (X_{\beta_1}, X_{\beta_1},\ldots,X_{\beta_m})$ can represent, for example, any set of potential niche variables. $\mathbf{X}$ in this sense can be seen as one single, composite variable. With N_{X_α} and $N_{X_\alpha X_\beta}$ in hand we may also calculate any associated probability distribution over our ensemble, such as $P(X_\alpha) = N_{X_\alpha}/N$, $P(X_\alpha X_\beta) = N_{X_\alpha X_\beta}/N$ or $P(X_\alpha|X_\beta) = N_{X_\alpha X_\beta}/N_{X_\beta}$.

The null hypothesis

So, how do we infer that a given observed value of $N_{X_\alpha X_\beta}$ or $P(X_\alpha|X_\beta)$ indicates the presence of an interaction? To do so we need a no-interaction null hypothesis. The relative merits of different null hypotheses have been the subject of much study (Gotelli, 2000). In our methodology, we take it to be that, in the absence of any interaction, we expect the distribution of X_α to be governed by the probability distribution $P(X_\alpha)$. This is equivalent to the null hypothesis of type SIM2 in the classification of Gotelli (2000)[1] and corresponds in the framework of presence-absence matrices to keeping the number of observations fixed but randomising their location. Thus, in an ensemble of size N_{X_β} we would expect to see $N_{X_\beta}P(X_\alpha)$ events of type X_α. Put another way, our

[1] This null hypothesis is one that leads to lower rates of Type I errors.

null hypothesis is that $P(X_\alpha X_\beta) = P(X_\alpha)P(X_\beta)$, i.e., the events X_α and X_β are independent. In the Bayesian sense, as we will see below, the null hypothesis is associated with $P(X_\alpha)$ being identified as a prior distribution, while $P(X_\alpha|X_\beta)$ represents a posterior distribution after the addition of the information X_β.

Making the world binary

In the above we have implicitly had in mind that a set of variables of special interest in ecology are the binomial variables that represent presence/absence or presence/no presence of a taxon. Thus, we will count co-occurrences of presence or absence/no presence of different taxa. However, there are many variables that could be relevant labels for characterising and quantifying an interaction that are not intrinsically binomial. For instance, a phenotypic variable, such as size, may be continuous, or abundances, or abiotic variables such as temperature, where we could potentially speak of interactions, as defined by deviations in spatial distributions relative to a null hypothesis. Thus, we may speak of the interaction between the presence of a species and a given climatic condition—temperature or precipitation—or, indeed, any abiotic variable, noting that by so doing we are emphasising the pragmatic, empirical nature of our characterisation of interaction rather than forcing it to accord with the standard taxonomy of ecological interactions. The problem with a continuous variable, X_β, relative to that of a binomial variable, X_α, such as presence, is that the number of co-occurrences $N_{X_\alpha X_\beta}$ will be very small, its magnitude depending on the resolution of the measurement of the continuous variable. Indeed, with arbitrary resolution, there will be no repetitions of the same value and $N_{X_\alpha X_\beta} = 0, 1$. In this case, direct statistical inference by counting is impossible, although by assuming a particular functional form of the distribution of the continuous variable as a function of position, it may be possible to proceed.

An alternative approach is to coarse grain any continuous variable into a set of discrete values and consider each discrete range as a new binomial dummy variable. Thus, for example, we can divide temperature into 10 bins with intervals $T_{min} + n(T_{max} - T_{min})/10$, $n = [1,10]$. Temperature in a given range is representable by a variable $X_{\beta n}(\mathrm{x}) = 0, 1$, as in cell we may ask if there is a "presence" of a temperature in any given range. Thus, the abiotic variable can now be treated as 10 presence/absence variables[2]. By doing this we avoid any model bias associated with an assumption about the relationship between independent and dependent variable, as would be the case in a regression analysis. There is, of course, a question of how many bins to choose and what should be their ranges? Too few bins risks missing potentially relevant information about variation of the variable within the bin, while too many bins risks losing statistical significance by having too few data points in a bin. The number of bins should also be motivated by underlying biological or ecological factors. For instance, we may ask if the spatial distribution of a biota is sensitive to variations in average annual temperature of 0.1°C? If not, then there is no need to use such resolution.

By making every variable binary we may represent any state of a given cell by a vector of binary variables $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, where $X_i = 0, 1$. An equivalent coding is to consider a single, composite variable of cardinality 2^m. Either can be used to represent any set of niche variables—abiotic and/or biotic.

Cell size and the problem of variables of different resolutions

We have defined co-occurrences with respect to a spatial cell of a given size. Besides the problem of the dependence of our results on this cell size[3] we must also ask how variables of quite different resolutions can be compared given a fixed cell size? The natural resolution of an environmental raster is at the pixel level, where there may be hundreds of thousands or even millions of pixels associated with a geographic area of interest. However, if we are to proxy biotic data by, say, point collection data, then for a given species we may have only tens or hundreds or data points. If we choose a pixel level resolution, then the interactions associated with this variable will be calculated from a sample size which is enormously greater than that for the biotic variable. More importantly, in any predictive model the contribution of the low-resolution variable will be very small relative to that of the high resolution variable in terms of its coverage, i.e., how many pixels are affected by the biotic variable. We may then be led to conclude that the abiotic interaction was much more important than the biotic one. This is a pure

[2] In this case we can truly say absence as we may identify those cells where the temperature definitely isn't in a chosen range.

[3] A problem (Gehlke and Biehl, 1934) in spatial data mining known as the "modifiable areal unit problem" (MAUP) (Openshaw, 1983).

effect of sample size associated with the resolution of the data however, without reflecting any underlying real difference. It is not the same as comparing the effect of a geographically restricted prey species versus a geographically disperse one. In this case, the disperse prey species would be associated with more cells than the restricted one. However, this reflects an underlying reality. The fact that reflects a $N_{X_{disperse}} \gg N_{X_{restricted}}$ difference in the importance of their interactions for a predator.

Thus, an environmental raster is equivalent to having corresponding presence/absence variables at every pixel. There are two alternatives to bringing variables to the same scale: i) extrapolate variables at a lower resolution to a higher one, or; ii) coarse grain variables of a higher resolution to a lower one. In the former, we must make assumptions associated with how this interpolation is done. For instance, one may use the proposed distributions for mammals of Hall (1981). A more convoluted way is to convert a discrete distribution into a raster by using a species distribution model that was built using abiotic rasters (Araújo et al., 2014; Atauchi et al., 2018). In option ii), however, there is no interpolation or assumption. The coarse graining is done using the dummy variables defined elsewhere ("Making the world binary"). Thus, at a given cell resolution, for any pixel-level raster we ask in that cell how many presence-absence dummy variables are represented. For instance, if in the cell there are pixels that lie in two of the ten temperature ranges then those two ranges are present in the cell and the other eight are absent.

The more general effect of cell size can be appreciated if we consider the limits of very large or very small cells. There are two general considerations: one is the effect of cell size on the effective size of the statistical sample to be analysed, and the other relates to its effect on the number of co-occurrences. For the former, as we are carrying out a statistical analysis and a corresponding hypothesis testing, it is natural to take advantage of the samples at our disposal as much as possible. If we have N events, then the maximum size of the sample of cells is also N. However, if the cell size is such that multiple, assuming they are independent, events occur in a given cell then the effective sample size is reduced. For instance, for a random distribution of events, if the cell size is such that, on average, the number of events per cell is 4, then a reduction in the cell size by a factor of 2 will probably lead to cells where the expected number of events per cell is closer to one. In other words, all else being equal, the number of cells should naturally scale as the number of events. For the case of co-occurrences, for a finite set of events, if we go to the limit of very small cells, it is clear that eventually we will end up with zero co-occurrences. On the other hand, in the limit of very large cells we will end up with only one co-occurrence, as all the events will be in one cell. The choice of cell size has been investigated empirically (Sierra and Stephens, 2012), where it has been determined that although an optimal resolution exists that maximises the number of co-occurrences, our results are robust to the precise cell size.

Testing the null hypothesis

We now have a means to quantify the difference between the spatial distribution of two taxa, C and X_α, and the distribution of either one of them in the absence of the interaction using either $I_1(CX_\alpha) = (P(C|X_\alpha) - P(C))$ or, equivalently, $I_2(CX_\alpha) = (P(CX_\alpha) - P(C)P(X_\alpha))$, with $I_2 = P(X_\alpha)I_1$. As I_1 and I_2 represent deviations from the null hypothesis of no interaction, any non-zero value might be interpreted as evidence of an interaction. However, as this question is being answered with respect to a statistical ensemble, we must determine its degree of statistical validity. Various diagnostics may in principle we used. However, here, given our emphasis on converting everything to binomial variables, we will use a simple binomial test based on our null hypothesis. Specifically, we will use

$$\varepsilon(C|X_\alpha) = \frac{N_{X_\alpha}(P(C|X_\alpha) - P(C))}{\sqrt{(N_{X_\alpha}P(C)(1-P(C)))}} \qquad (5)$$

In the case where the binomial distribution may be approximated by a normal distribution, then $|\varepsilon(C|X_\alpha)| > 1.96$ corresponds to the 95% confidence interval for consistency with the null hypothesis. When a normal approximation is inadequate then a more sophisticated approximation may be used, such as the Wilson intervals (Wilson, 1927). Note that $\varepsilon(C|X_\alpha) \neq \varepsilon(X_\alpha|C)$, i.e., it is asymmetric in its arguments, thus modelling the fact that the effect of X_α on C may not be the same as that of C on X_α.

If we had used I_2 instead of I_1 to quantify deviations from the null hypothesis then the corresponding diagnostic is

$$\varepsilon(CX_\beta) = \frac{N(P(CX_\alpha) - P(C)P(X_\alpha))}{\sqrt{(NP(C)P(X_\alpha)(1 - P(C)P(X_\alpha)))}} \quad (6)$$

which is a measure of the deviations from the null hypothesis that $P(CX_\alpha) = P(C)P(X_\alpha)$. Of course, when there are sufficiently many X_α, then some $\varepsilon(C|X_\alpha)$ will be statistically significant for any chosen p-value. This is nothing new and there are many methods that may be used to ameliorate this effect, such as applying a Bonferroni correction, or an ANOVA analysis. Importantly, one should also have some intuitive notion of why it should be significant in the first place. Note that the principal role of $\varepsilon(C|X_\alpha)$ is to inform us that the spatial distributions of C and X_α are such that they are not consistent with the null hypothesis that the distribution of C is independent of the distribution of X_α and that, therefore, by our definition, there is an interaction between them.

What else can it tell us? Note that it has both an intensive and an extensive character. I_1, or equivalently I_2, can be used to begin to characterise the intrinsic strength of the interaction in that the larger is I_1 the more the presence of species α is correlated with the presence of species C. On the other hand, by virtue of taking into account the sample size, N_{X_α}, it allows us to infer a greater importance for the interaction as the sample size increases, if the sample size is representative of the underlying relative geographic distribution of the taxon. As an example, consider the relation between a vector, C, and a potential host species, α. The greater is I_1, the more likely it is that the vector occurs when the host species is present relative to a random benchmark. On the other hand, for a fixed I_1, the greater the value of N_{X_α}, the higher the expected number of observed instances of C relative to the null hypothesis.

So, $\varepsilon(C|X_\alpha)$ captures two different facets of the interaction, the first associated with the strength of the interaction via I_1 and the second with the coverage of the interaction via $N_{X\alpha}$. It can be used to compare and contrast multiple interactions. For example, we may have three distributions C, $X_\alpha(\mathrm{x})$ and $X_\beta(\mathrm{x})$, such that $I_1(CX_\alpha) \gg I_1(CX_\beta)$ but that $\varepsilon(C|X_\alpha) \approx \varepsilon(C|X_\beta)$, due to the fact that $N_{X_\beta} \gg N_{X_\alpha}$. In this case, we would conclude that the interaction between C and α is stronger than that between C and β.

Note that (5) and (6) generalise to the case where X_α represents a composite variable, equivalent to a set of variables. Thus,

$$\varepsilon(C|\boldsymbol{X}) = \frac{N_{\boldsymbol{X}}(P(C|\boldsymbol{X}) - P(C))}{\sqrt{N_{\boldsymbol{X}}P(C)(1 - P(C)))}} \quad (7)$$

represents a measure of the interaction between a taxon C and its niche variables $\mathbf{X}$. Indeed, we may use equation (7) to quantify the niche. The greater is the difference $I_1 = (P(C|\mathbf{X}) - P(C))$, for a given configuration of the niche variables $\mathbf{X}$, the more that configuration represents more favourable niche conditions for the presence of C. Similarly, if $I_1 < 0$, the more negative it is, the more the corresponding configuration of variables represents "anti-niche" conditions, i.e., conditions that are unfavourable for the presence of C. However, the problem with considering multiple niche variables within this simple diagnostic directly is that $N_{\mathbf{X}} = 0, 1$ when many niche variables are included.

Note that as a measure of non-random co-occurrence relative to the null hypothesis, $\varepsilon(C|X_\alpha)$ is quite different to the checkerboard matrix elements (Roberts, 1990) $C_{CX_\alpha} = (N_C - N_{CX_\alpha})(N_{X_\alpha} - N_{CX_\alpha})$, as the latter is independent of the size of the statistical ensemble chosen. Additionally, in the case of the checkerboard, the null hypothesis is applied to the entire matrix and a single index—the checkerboard score—is evaluated relative to that null hypothesis.

What can we infer from co-occurrences?

The interpretation of co-occurrences is very much related to the characterisation of our cells and our objects. To fix intuition: in the case of a space-time ensemble, a predation event could be characterised by the dynamics of an individual of a predator species, represented by a presence variable, $X_\alpha(\mathbf{x},t) = 0, 1$, and an individual of a prey species represented by a presence variable, $X_\beta(\mathrm{x}',t') = 0, 1$, which represent the trajectories of predator and prey in space and time. In this case, predation would be associated with the fact that $N_{X_\alpha X_\beta} = 1$, where $X_\alpha(\mathrm{x},t) = X_\beta(\mathrm{x},t) = 1$ and after the predation event $X_\beta(\mathrm{x}',t') = 0$ for any $t' > t$. In other words, the predator and prey have trajectories that intersect, such that after the co-occurrence the prey species individual disappears. The statistical ensemble of events here is as-

sociated with the positions, x(t), of the two individuals. This ensemble, however, would tell us nothing about how typical this particular interaction was nor, indeed, whether or not it was just a chance encounter. In other words, it would tell us nothing beyond the interaction of those two individuals. At the population level, we must account for the fact that a prey species may be easy to catch but be relatively rare, or difficult to catch and widely distributed, or any combination of these and other characteristics. Thus, an ensemble of co-occurrences of predator-prey would potentially yield much more information about their interaction beyond the individual level. From such ensembles we may deduce the likelihood of success of an attempted predation for instance, or the importance in the overall diet of the predator of a given prey species, all of which are relevant for characterising the interaction and for considering the prey species as a niche variable of the predator.

There is also the possibility that an interaction is more, or less, manifest at the level of one spatial resolution/ensemble versus another. For instance, it may be that the predator-prey interaction is manifest at the level of the trajectories of the individuals, but that there is no correlation between these events. In other words, that the distributions of predator and prey are random, and that any predation event is the result of a chance encounter. This is a common occurrence in physics. For instance, a gas of Helium atoms will exhibit a strong interaction between the positively charged nucleus and the two electrons that make up the atom but there will be no resultant interaction between the neutral atoms themselves.

As well as depending on the ensemble chosen, the inference that may be drawn from any event, or set thereof, depends on the spatio-temporal resolution of the cells. If ΔA is 1 m^2 and ΔT = 1 second then we might infer, knowing that one individual was a predator species and one was a prey species, that disappeared after the co-occurrence, that a predation event had occurred. The trajectories before the co-occurrence might also give extra information that supported the hypothesis. Was the prey species trying to avoid the predator species for example? There are multiple null hypotheses that could be used for comparison associated with the expectation of the trajectories of the individuals in the absence of the interaction. For instance, that the degree of correlation between the trajectories is zero. We could also construct an ensemble of co-occurrences of the two species with the same spatial resolution and just count the number of cells in which predator and prey co-occurred where subsequently the prey disappeared. However, if the spatial resolution were 1 km^2 and the temporal resolution were 1 year then the co-occurrence only tells us that the predator and prey were in the same 1 km^2 area at some point in the last year. This, of course, is not sufficient to infer a particular predation event.

How might the relation between predator and prey be inferred now? In this case, we must infer any relation from a different statistical ensemble. If we consider a spatial ensemble of N cells, we have stated that ε can be used to determine the existence of an empirically defined interaction between two taxa. What else can it tell us? As noted, it has both an intensive and an extensive character. First, I_1 can be used to begin to characterise both the sign and the intrinsic strength of the interaction, in that the larger is I_1 the more the presence of species C is predictive of the presence of species α. If $I_1 > 0$ we will say that the interaction is *attractive*, in that the probability to find C and α co-occurring is higher than our null hypothesis, and *repulsive* in the contrary case. Furthermore, the larger the magnitude of I_1, the stronger the interaction. Finally, considering N_{X_α} allows us to infer a greater importance for the interaction as N_{X_α} increases. These interpretations naturally depend on the fact that the underlying data of the ensemble is representative of the underlying relative geographic distribution of the species.

As an example, consider the relation between a predator, C, and two potential prey species, α and β. The greater is I_1, the more likely it is that the predator occurs when the prey species is present relative to the null hypothesis. Thus, if $I_1(C|X_\alpha) = I_1(C|X_\beta)$ we say that the two interactions have the same strength. This is a measure of the interaction between any given individual predator and any individual prey taken from the ensemble. However, if $N_{X_\alpha} > N_{X_\beta}$ we will say that the interaction between C and α is more important than that between C and β. This is a measure of the interaction at the population level.

So, the process by which we may infer an interaction is the following: i) construct a statistical ensemble of cells associated with a given spatial and/or temporal resolution; ii) compute co-occurrences of presence variables on that ensemble, where the presence variables may have an associ-

ated set of labels—taxonomic, phenotypic, behavioural etc.; iii) compute one or more statistics based on the distribution of co-occurrences; iv) compare that statistic with a suitable null hypothesis; v) if the statistic and the null hypothesis are statistically significantly different conclude that there is a macro interaction; vi) use the labels and other data to deduce the nature of the interaction and relate it to micro interactions.

What spatial data?

Essentially, we have highlighted the predation example to emphasise that, as in physics, spatio-temporal data can be used to identify and characterise ecological interactions. The real question is: what spatio-temporal data is necessary or sufficient? Must we have micro-data associated with each, individual event across space and time? How are interactions manifest at different resolutions? Clearly, we are not in the position of being able to track in space and time the positions of representative sets of all taxa. There are just too many potential interactions to be identified and quantified individually. Thus, in the absence of true, detailed observational data on biotic interactions at a macro-scale, as derived aggregated micro data, one must resort to proxy data. This is similar to the dilemma faced when contrasting epidemiological considerations against clinical or physiological variables, where the causal distance between a disease, say, and its symptoms, may be much less than that between the same disease and its risk factors.

An important consideration then is what data will represent the spatio-temporal distributions of biota? In standard niche modelling, where the output variable is the spatial distribution of a taxon, the proxy of choice has been a database of point collection data. Such data may be bespoke, in that it represents a dedicated, controlled study that tries to be as unbiased as possible, versus museum collection data, where the data, although ample and widespread, is potentially biased and unrepresentative. However, in spite of all its shortcomings, which have been amply discussed (Hortal et al., 2008; Soberón and Peterson, 2004), such data has been used ubiquitously to calculate species distributions (Peterson et al., 2011). Moreover, as mentioned, other sets of potentially biased biotic data have then been used as dependent variables to produce models using abiotic independent variables that are then input as independent variables as potential proxies of biotic interactions in the calculation of the distribution of a species of interest (Araújo et al., 2014; Atauchi et al., 2018).

The ultimate test has to be to validate point-collection data by the predictions of models that use it as input. There are two possibilities: i) use it to calculate species distributions, using abiotic and/or biotic data, and infer interactions as defined herein, and then determine to what extent those distributions and interactions are consistent with *known* ecological interactions and, furthermore, if they lead to more precise species distribution models and a better ecological understanding of the niche in such models; ii) use it to predict *unknown* ecological interactions and use experimental protocols incorporating field work and potentially laboratory work to validate those predictions.

A Bayesian framework for analysing ecological interactions

Our definition of an interaction is probabilistic, based on analysing the difference between $P(C|\mathbf{X}(t))$ and a null hypothesis $P(C)$ for a taxon C. This is equally applicable in the case that $\mathbf{X}$ represents just one niche variable, X_α, versus many. In either case, with a statistically significant deviation between them we infer the existence of an interaction between C and $\mathbf{X}$ that we must then further understand.

We take it as an axiom that the spatial distribution of a taxon is a result of *all* the interactions, both abiotic and biotic, that impact on that distribution and that there exists an underlying, potentially dynamic, probability distribution that predicts and explains the distribution of the species. Of course, there are many reasons why a predicted distribution may not be a good representation of the real distribution. First, it may be that the underlying distribution is dynamic and approximating it by a static ("equilibrium") distribution is inadequate. Secondly, it may be that important variables have been omitted from the model, such as biotic variables. Thirdly, it may be that the data representation of the variables is inadequate, because of data bias, and, finally, it may be that the mathematical relation between those variables is not being modelled correctly. Thus, any model and hypothesis about interactions must be validated both through its predictions and how it increases our understanding. In this

context we would like to know how different variables $X_\alpha \in \mathbf{X}$ contribute to $P(C|\mathbf{X})$. In other words, from overall information about the interaction between a taxon and its niche, how may we deduce and characterise the relative effects of interactions with individual niche factors? In other words, how does an overall interaction taxon-niche, as proxied by $P(C|\mathbf{X})$, emerge from its component interactions $P(C|X_\alpha)$?

The theoretical framework we use to answer these questions is probabilistic and Bayesian based. The Bayesian formulation of probability theory and statistical inference has had an enormous impact (Berger, 1985). One of its most important advantages is the natural way in which qualitative information about beliefs can be incorporated as Bayesian priors, as well as the way in which quantitative information may then be naturally incorporated into a posterior probability using Bayes' theorem. A second advantage is how it can naturally incorporate new information and beliefs and update posterior probabilities in the light of this new information. A third important advantage is that it gives a very natural setting for considering issues of causality (Pearl, 2000).

The fundamental basis for the Bayesian formulation of probability is Bayes Theorem

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})} \tag{8}$$

where $P(C)$ is the prior probability for the event C in the absence of the information $\mathbf{X}$, which can be represented by a vector of variables $\mathbf{X} = (X_1, X_2, \ldots, X_m)$. $P(\mathbf{X}|C)$ represents the likelihood of observing the information $\mathbf{X}$ given the event C, and $P(C|\mathbf{X})$ is the posterior probability that takes into account how the data $\mathbf{X}$ allows one to adjust the expectation of C relative to its prior. The evidence, $P(\mathbf{X})$, is a normalisation factor independent of C. From a frequentist viewpoint all these probabilities may, in principal, be calculated directly from data. For example, $P(C) = N_C/N$, where N_C is the number of events of type C and N is the total number of events, i.e., the size of the statistical ensemble. However, $P(C)$ may also represent our belief about the probability of the event C. This is relevant when the concept of an ensemble, measurable in frequentist terms, is not readily available.

Our diagnostic, equation (5), for characterising the presence of an interaction has a very natural Bayesian interpretation, as I_1 measures the deviation of the posterior probability $P(C|X_\alpha)$ in the presence of the information X_α from the prior distribution $P(C)$. This is equally true if X_α represents a single or a composite variable. From a hypothesis testing perspective, equation (5) provides us with an estimate of the degree of confidence we may have as to whether the information X_α leads to an improvement in our estimation of the probability of C.

Often, to get rid of the C-independent evidence function, $P(\mathbf{X})$, the following "score" function is used

$$S(C|\boldsymbol{X}) = ln\left(\frac{P(C|\boldsymbol{X})}{P(\bar{C}|\boldsymbol{X})}\right) = ln\left(\frac{P(\boldsymbol{X}|C)}{P(\boldsymbol{X}|\bar{C})}\right) + ln\left(\frac{P(C)}{P(\bar{C})}\right) \tag{9}$$

where $\bar{C}$ is the set complement of C (absences or no presences of C) and hence $P(\bar{C}) = 1 - P(C)$. If $S(C|\mathbf{X}) > 0$ it is more likely that the information indicates the presence of the event C, and vice versa for $S(C|\mathbf{X}) < 0$. The $\mathbf{X}$ independent constant term on the right-hand side just accounts for the different class weightings. So, if the class is a small percentage then S naturally leans towards classifying instances into $\bar{C}$ rather than C.

Seen as a rigid classifier, $S(C|\mathbf{X}) > 0$ indicates that the instance $\mathbf{X}$ should be assigned to the class C. In the context of ecological interactions, $S(C|\mathbf{X}) > 0$ indicates that the conditions $\mathbf{X}$ are favourable for the presence of C, with the higher the value of S the more favourable the conditions and vice versa for $S(C|\mathbf{X}) < 0$. In the context of species distribution modelling, the estimation of $P(C|\mathbf{X})$ or $S(C|\mathbf{X})$, or an equivalent, such as $P(C,\mathbf{X})$, can be done using different algorithms. The well-known Maxent procedure (Phillips et al., 2004, Phillips et al., 2006) is one. Even a more "black-box" procedure, such as GARP (Stockwell, 1999), is effectively doing the same thing. Of course, $S(C|\mathbf{X})$, as representing the niche in the Hutchinsonian sense, can be used to map back into geographic space and thus provides a species distribution model whose performance can then be measured using any one of many metrics, such as the area under the Receiver Operating Curve

(AUC), or confusion matrix statistics, such as sensitivity and specificity etc. Each has its pros and cons.

Unfortunately, when **X** is of high dimension then neither $P(C|\mathbf{X})$ nor $P(\mathbf{X}|C)$ may be estimated reliably from data as $N_{C\mathbf{X}}$ is either 0 or 1. In other words, if we define an environment with sufficient precision then there is either no co-occurrence with C or only 1 due to the fact that we will not find exactly the same environment in two different places. So, how to proceed? A general approximation, that maintains the Bayesian philosophy is to approximate the likelihood function by assuming that the variables $(X_1,X_2,...,X_m)$ are independent. Hence, $P(\mathbf{X}|C) \rightarrow \prod_{\alpha=1}^{m} P(X_\alpha|C)$. In this case equation (9) becomes

$$S(C|\boldsymbol{X}) = \sum_{\propto=1}^{m} s(X_\propto) + ln\left(\frac{P(C)}{P(\bar{C})}\right) \tag{10}$$

where is $s(X_\alpha) = \ln(\frac{P(X_\alpha|C)}{P(X_\alpha|\bar{C})})$ the contribution ("score") to the overall $S(C|\mathbf{X})$ from the variable X_α. If $s(X_\alpha) > 0$, < 0 then the factor X_α contributes positively/negatively to the occurrence of the event C. This is just the well-known Naive Bayes approximation (NBA) that is still extensively used in many, many applications as it is easy to implement, computationally very efficient and very transparent (Broos et al., 2011; Burak and Ayse, 2009; Wang et al., 2007; Wei et al., 2011). The maximum entropy algorithm and the NBA are closely related (Alfonso and Vilar., 2007), as the latter can be written in maximum entropy form, but with a Gibbs distribution that is factorizable.

Within this approximation we determine precisely how a measure of the overall interaction between taxon and niche variables, $S(C|\mathbf{X})$, is composed of measures, $s(X_\alpha)$, of the individual interactions between C and any given variable X_α. Thus, the higher/lower the value $s(X_\alpha)$ the more it determines favourable/unfavourable conditions for the presence of C. If $S(C|\mathbf{X}) = 0$ then the taxon C is effectively distributed randomly, the variables **X** having no overall effect. However, it does not follow that each $s(X_\alpha) = 0$. The composition of a set of potential niche variables can be such that their net effect is neutral by cancellation between positive and negative contributions. Note that both $S(C|\mathbf{X})$, are $s(X_\alpha)$ are measures of interaction strength, in that they make no reference to how widely distributed are **X** or X_α. In Machine Learning terms the latter is more related to the coverage of the features **X** or X_α, meaning how many data instances, in this case cells, are represented by them. Thus, $s(X_\alpha)$ may be very positive for a given X_α but, if this represents a rare species, N_{X_α} may be very small relative to N. In other words, although it is an important variable, it is not widespread. Similarly, one may have a widespread variable that has only a weak interaction. Our interaction diagnostic (5) considers both aspects of the interaction: strength and coverage.

The chief criticism of the NBA is its strong assumption of feature independence, where in the case of ecology one certainly knows that many niche variables will be highly correlated. In spite of this it works, almost unreasonably, well. As has been pointed out and quantified (Stephens et al., 2017b), one reason the NBA works better than expected is that the correlations between features can be both positive or negative across different feature combinations. In fact, it can be generalised, so that different features may be combined. This leads to an improved approximation, both in terms of predictive accuracy as well as enhanced understanding of which variables are correlated (Stephens et al., 2017b). We will present some basic elements of this below ("Beyond the Naive Bayes approximation"), as it is very relevant for the important question of confounding and causality.

In practice, we must decide what data to use to estimate $\varepsilon(C|X_\alpha)$, $s(X_\alpha)$ or their niche counterparts with $X_\alpha \rightarrow \mathbf{X}$. The fundamental components are counts: N_{CX_α}, N_{X_α} and N_C, that are taken from a statistical ensemble. The ensemble of most interest is that of N cells within which we count presence or no presence/absence. In the case of small samples, we may have that $N_{CX_\alpha} = N_{X_\alpha}$ or that $N_{CX_\alpha} = 0$. In these cases, $s(X_\alpha) = \infty$ or $-\infty$ respectively. To avoid this the probabilities may be smoothed using a correction factor, such as the Laplace correction (Chen, 1996), whereupon $N_{CX_\alpha} \rightarrow N_{CX_\alpha} + A$ and $N_C \rightarrow N_C + B$, where A and B are constants. A common choice is $A = 1$ and $B = 2$.

Model Selection

Prior probabilities and model selection.—An important question is: which variables should be included in the model? A niche model that only accounts for climatic data, using a feature vector $\mathbf{X}^a$ to

describe the distribution of a species C, would, relative to a uniform prior $P(C)$, calculate the posterior probability to be

$$P(C|\boldsymbol{X}^a) = \frac{P(\boldsymbol{X^a}|C)P(C)}{P(\boldsymbol{X}^a)} \tag{11}$$

We may then ask how biotic factors, $\mathbf{X}^b$, may be added. In the Bayesian formulation, we may take the posterior probability, $P(C|\mathbf{X}^a)$, after including in abiotic variables and use it as a new prior probability for which we then compute the likelihood associated with the new biotic information, $\mathbf{X}^b$, and subsequently calculate the new posterior distribution, $P(C|\mathbf{X}^b\mathbf{X}^a)$, that includes both biotic and abiotic factors. Thus,

$$P(C|\boldsymbol{X}^a\boldsymbol{X}^b) = \frac{P(\boldsymbol{X^b}|C,\boldsymbol{X^a})P(C|\boldsymbol{X^a})}{P(\boldsymbol{X}^a|\boldsymbol{X}^b)} \tag{12}$$

Similarly, for the score function we have

$$\begin{aligned} S\left(C\middle|\boldsymbol{X}^a\boldsymbol{X}^b\right) &= ln\left(\frac{P\left(C\middle|\boldsymbol{X}^b\boldsymbol{X}^a\right)}{P\left(\bar{C}\middle|\boldsymbol{X}^b\boldsymbol{X}^a\right)}\right) \\ &= ln\left(\frac{P\left(\boldsymbol{X}^b\middle|C,\boldsymbol{X}^a\right)}{P(\boldsymbol{X}^b|\bar{C},\boldsymbol{X}^a)}\right) \\ &\quad + ln\left(\frac{P(C|\boldsymbol{X}^a)}{P(\bar{C}|\boldsymbol{X}^a)}\right) \end{aligned} \tag{13}$$

In the NBA, $P(\mathbf{X}^b|C,\mathbf{X}^a) = P(\mathbf{X}^b|C) = \prod_{\alpha=1}^{m} P(X_\alpha^b|C)$, where the first equality uses the fact that in this approximation the abiotic and biotic factors act independently. Hence,

$$\begin{aligned} &S\left(C\middle|\boldsymbol{X}^a\boldsymbol{X}^b\right) \\ &= \sum_{\propto=1}^{N_b} s\left(X_\propto^b\right) + \sum_{\propto=1}^{N_a} s(X_\propto^a) + ln\left(\frac{P(C)}{P(\bar{C})}\right) \end{aligned} \tag{14}$$

where N_a and N_b are the numbers of abiotic and biotic variables respectively. We may now better understand what an approximation where only abiotic variables are used implies. Essentially, it is equivalent to $P(C|\mathbf{X}^a\mathbf{X}^b) = P(C|\mathbf{X}^a)$ and hence $P(C|\mathbf{X}^b) = P(C)$. In the NBA, we hence have $\sum_{\propto=1}^{N_b} s\left(X_\propto^b\right) = 0$, which is most naturally interpreted as $s\left(X_\propto^b\right) = 0$ for each $X_\propto^b$. If this is not true, then the approximation of omitting biotic variables will not be a good one. The same would be true if we omitted abiotic variables and considered only biotic ones. Our methodology of bringing everything to the same spatial resolution and making every variable binomial allows us to make a direct comparison of the contributions from every single variable. Furthermore, as we will see, we may use the predictions of the model to validate the inclusion or exclusion of certain variables using the pragmatic criterion of whether or not they improve model performance. A further subtlety is that although we leave out a certain set of variables that does not mean their influence is absent, due to the fact that there may be correlations between omitted and included variables such that the latter include effects from the former. This is just the effect of confounding.

Evaluating predictability = ranking interactions.—Independently of the approximation used to calculate the likelihood, an important task is to compare and contrast the contributions of the different features, and/or feature combinations, in order to obtain a better understanding of their relative importance. Different criteria may be used, both data based and "belief" based. The belief-based component is as discussed above—selection is based on supposed knowledge of what are important variables to include. Thus, a modeller of the niche and species distribution of a predator may decide to include in as biotic factors only its two "known" principle preys. Of course, this biased model should be compared with others to determine if it leads to better model performance and better understanding. For instance, if, in fact, the predator has several other preys that are unknown then their inclusion would be expected to improve model performance.

To choose variables for inclusion in an overall model, an appropriate measure of the relative importance and associated statistical significance of a niche variable X_α is just the binomial test (5). Similarly, for a combination of features $\mathbf{X}$ we may use (7). We may use the same statistical criteria on ε, such as $|\varepsilon| > 1.96$, to decide whether or not to include the variable in our prediction model. Thus, we link a machine learning-based concept—feature selection—with our definition of interaction, so that

features (niche variables) are included only if they represent significant interactions.

Beyond the Naive Bayes approximation.—In spite of the robust nature of the NBA, it is important to try and determine which niche variables are correlated and whether a prediction model may be improved by considering them together. Even more importantly, such correlations will help us study aspects of causality in order to help distinguish between causation and correlation. We consider first the impact of a pair of features, X_α and X_β, following the same procedure as for one variable. In this case, for the joint probability distribution, $P(C, X_\alpha,X_\beta) = P(X_\alpha X_\beta|C)P(C)$, we define (Stephens et al., 2017b)

$$\Delta(X_\alpha X_\beta|C) = (P(X_\alpha X_\beta|C) - P(X_\alpha|C)P(X_\beta|C)) \tag{15}$$

as a measure of correlation between niche variables. Significant deviations from the null hypothesis $\Delta(X_\alpha X_\beta|C) = 0$ indicate the presence of significant correlations which we interpret as the fact that the niche variables X_α and X_β are not independent with respect to the taxon C. In this case the interaction is ternary from the point of view of X_α, X_β and C, though one can also consider it to be binary if we consider $X_\alpha X_\beta$ as a combination variable. In analogy with $\varepsilon(C|X_\alpha)$, we can use a binomial test to determine the degree of consistency with the null hypothesis, considering (Stephens et al., 2017b)

$$\varepsilon(X_\alpha X_\beta|C) = \frac{N(C)(P(X_\alpha X_\beta|C) - P(X_\alpha|C)P(X_\beta|C))}{\sqrt{(N_C P(X_\alpha|C)P(X_\beta|C)(1 - P(X_\alpha|C)P(X_\beta|C)))}} \tag{16}$$

Equation (16) determines when there is a statistically significant correlation between the features X_α and X_β. Instead of considering the likelihoods, we may also consider the posterior probability $P(C|X_\alpha X_\beta)$ directly. To determine the impact of a pair of features, X_α and X_β, we may follow the same procedure as for one variable, considering

$$\varepsilon(C|X_\alpha X_\beta) = \frac{N_{X_\alpha X_\beta}(P(C|X_\alpha X_\beta) - P(C))}{\sqrt{(N_{X_\alpha X_\beta} P(C)(1 - P(C)))}} \tag{17}$$

Once again, in the case where the binomial distribution may be approximated by a normal distribution, then $|\varepsilon(C|X_\alpha X_\beta)| > 1.96$, which corresponds to the 95% confidence interval for testing consistency with the null hypothesis. In this case however, (17) cannot distinguish between the relative contributions of X_α versus X_β. This can be done, though, by considering alternative null hypotheses. Considering as null hypothesis $P(C|X_\alpha)$ and $P(C|X_\beta)$ in turn, we have

$$\varepsilon(C|X_\alpha X_\beta; X_\beta) = \frac{N_{X_\alpha X_\beta}(P(C|X_\alpha X_\beta) - P(C|X_\beta))}{\sqrt{(N_{X_\alpha X_\beta} P(C|X_\beta)(1 - P(C|X_\beta)))}} \tag{18}$$

and, similarly,

$$\varepsilon(C|X_\alpha X_\beta; X_\beta) = \frac{N_{X_\alpha X_\beta}(P(C|X_\alpha X_\beta) - P(C|X_\alpha))}{\sqrt{(N_{X_\alpha X_\beta} P(C|X_\alpha)(1 - P(C|X_\alpha)))}} \tag{19}$$

Equations (18) and (19) can be used as measures of the relative impact of one variable versus another. For instance, (18) expresses the contribution of the variable X_β to the posterior probability $P(C|X_\alpha X_\beta)$, i.e., in the presence of the feature X_α, but relative to the contribution marginalised over X_α. These equations, in principle, allow us to determine which, of a combination of two variables, is the most important in terms of prediction of the class variable. Moreover, they facilitate the determination and analysis of confounding variables. As an example, imagine we determine that $\varepsilon(C|X_\alpha)$ is significant, but hypothesise that there exists a confounding variable X_β. In this case, we may consider $\varepsilon(C|X_\alpha X_\beta;X_\alpha)$ and $\varepsilon(C|X_\alpha X_\beta;X_\beta)$. If X_β is, indeed, a confounding variable, then we should find that $\varepsilon(C|X_\alpha X_\beta;X_\alpha) > \varepsilon(C|X_\alpha X_\beta;X_\beta)$ as the real influence on C is from X_β and so, if we take $P(C|X_\beta)$ as null hypothesis, we should find little residual predictability. Note that (16) can be used to identify those feature combinations that should be considered together, and this information used to generate a Generalized Bayes approximation (Stephens et al., 2017b), where the factorization of the likelihoods is not maximal, and which leads to an improved predictive model.

Inferring causality

A criticism that has been levelled against our methodology is that it does not allow one to infer causality, in that an apparently important contribution from a biotic niche variable may, for example, just be reflecting the existence of an underlying abiotic factor that acts as a confounder (Purse and Golding, 2015). Of course, the question of distinguishing correlation from causation is a topic of

great import across many branches of science and has a large literature (Pearl, 2000), especially in the social and medical sciences, where there is a "classical" approach (Hill, 1965) and a "modern" approach (Rosenbaum and Rubin, 1983; Rubin, 1974; Rubin, 1978). In the classical approach (Hill, 1965), a set of criteria were introduced for judging epidemiological evidence of a causal relationship between a presumed cause and an observed effect. They are:

1. *Strength (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.* We will use this in the context of comparing Bayesian score contributions from different variable types.
2. *Consistency (reproducibility): Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.* This could be checked by considering different sample populations for a given hypothesis.
3. *Specificity: Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.* We will consider this in the context of the "modern" approach by inserting sets of potential confounders.
4. *Temporality: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).* This requires time ordered data and is, in principle, possible with ecological data if it has such time ordering.
5. *Biological gradient: Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.*
6. *Plausibility: A plausible mechanism between cause and effect is helpful.* That there is some sound ecological/biological underpinning.
7. *Coherence: Coherence between epidemiological and laboratory findings increases the likelihood of an effect.* In other words, that a laboratory experiment, such as a determination of the positivity of a species with respect to infection by a pathogen is consistent with an inference of a relation between host and vector from point collection data.
8. *Experiment: "Occasionally it is possible to appeal to experimental evidence".*
9. *Analogy: The effect of similar factors may be considered.*

For the purposes of illustration, an example where we will apply the above framework is that of the relations between climate, vegetation, herbivore and carnivore. In particular, below ("Some representative results"), we will consider as a specific example of a causal chain: *Lynx rufus* as a carnivore; *Sylvilagus floridanus* as a known, important prey of the *L. rufus*; *Microchloa kunthii* as a known, important food source of *Sylvilagus floridanus* (Hudson, 2005); and, finally, climate as a known, important factor in the presence and abundance of the plant species. By the nature of this causal chain, we are led to hypothesise that the presence/no presence of *Sylvilagus floridanus* is more important to the presence of *Lynx rufus* than the presence/no presence of *Microchloa kunthii*, which in turn is more important than the presence/no presence of a particular climatic configuration. The question will be if the nature of this causal chain may be deduced from spatial data and to what degree we can characterise confounding?

A Bayesian framework for causal inference

With the Bradford-Hill criteria in mind, we may use the formalism presented elsewhere ("Beyond the Naive Bayes approximation"), to see how to infer causality. Consider two factors, biotic or abiotic, X_α and X_β, and their impact on the distribution of taxon C, as modelled by the conditional probability $P(C|X_\alpha X_\beta)$. We would like to determine the combined impact of X_α and X_β relative to some null hypothesis and use this analysis to determine the relative importance of X_α and X_β in predicting the presence of C. The most natural starting point is to consider X_α and X_β together as one composite variable and calculate

$$\varepsilon(C|X_\alpha X_\beta) = \frac{N_{X_\alpha X_\beta}(P(C|X_\alpha X_\beta) - P(C))}{\sqrt{(N_{X_\alpha X_\beta} P(C)(1-P(C)))}} \tag{20}$$

This can be compared with $\varepsilon(C|X_\alpha)$ or $\varepsilon(C|X_\beta)$ separately. What conclusions could we draw? If $I_1(C|X_\alpha X_\beta) > I_1(C|X_\alpha)$ or $I_1(C|X_\beta)$ we would infer that the interaction of species C with α and β together is

stronger than that with either α or β separately. Moreover, we may consider different combinations of $X_\alpha = 0, 1$ and $X_\beta = 0, 1$. Thus, $I_1(C|X_\alpha = 1\ X_\beta = 1)$ is the strength of the interaction in the presence of both factors, while $I_1(C|X_\alpha = 1\ X_\beta = 0)$ and $I_1(C|X_\alpha = 0\ X_\beta = 1)$ are the strengths in the presence/absence of α and absence/presence of β. Finally, $I_2(C|X_\alpha = 0\ X_\beta = 0)$ is the strength of the interaction in the absence of either factor. If $I_1(C|X_\alpha = 1\ X_\beta = 0) > I_1(C|X_\alpha = 0\ X_\beta = 1)$ this tells us that the interaction between C and α is stronger than that between C and β. Using Bradford-Hill criterion number 1 (Hill, 1965] in the section entitled "Inferring causality and identifying cofounders," we will take that as an indication that α is causally closer to C than β is. We will provide some specific examples of this in the section entitled "Some representative results."

Note that it is not appropriate to compare the values of $\varepsilon(C|X_\alpha X_\beta)$ itself directly with those of $\varepsilon(C|X_\alpha)$ or $\varepsilon(C|X_\beta)$, as, all else being equal, $\varepsilon(C|X_\alpha X_\beta)$ will be less than $\varepsilon(C|X_\alpha)$ or $\varepsilon(C|X_\beta)$ simply because $N_{X_\alpha X_\beta} < N_{X_\alpha}$ or N_{X_β}. Thus, $\varepsilon(C|X_\alpha X_\beta)$ is naturally commensurate with $\varepsilon(C|X_\alpha X_\gamma)$ but not with $\varepsilon(C|X_\alpha)$ or $\varepsilon(C|X_\beta)$.

Predicting interactions

We have defined an interaction as a deviation from an appropriate null hypothesis of the spatial distribution of a taxon conditioned on one or more abiotic and/or biotic variables. We have argued that the statistical diagnostic ε allows us to intuit a measure of the strength of this interaction as well as its importance (coverage), while $s(x_\alpha)$ is more a direct measure of its strength. We have also argued that its interpretation depends on the statistical ensemble of data used to compute it, emphasising that point collection data is a set of special interest given its wide availability. Although this definition of interaction is not scale dependent when we speak of using point collection data and corresponding species data to compute ε then we are naturally speaking of "macro" level interactions.

As noted, we may view interactions between individual taxa or between a taxon and its niche. The score functions $s(X_\alpha)$ provide a means for determining the relative contribution of X_α to $S(C|\mathbf{X})$ and we may consider ranking all included potential niche variables X_α, $\alpha = [1,n]$, from highest, s_{max}, to lowest, s_{min}, as a means of comparing their relative strengths, with s_{max} being the strongest positive interaction—most favourable niche variable—and s_{min} the strongest negative interaction—most unfavourable niche variable. We may do the same using $\varepsilon(C|X_\alpha)$, ranking them in descending order, from ε_{max} to ε_{min}, with the most positive values of $\varepsilon(C|X_\alpha)$ representing the strongest and most important attractive interactions and the most negative values representing the strongest and most important repulsive interactions. The extra component in $\varepsilon(C|X_\alpha)$ relative to $s(X_\alpha)$, as has been amply discussed above, is its dependence on N_{X_α} which is a measure of the geographic coverage of the niche variable.

Our characterisation of interaction is such that if there is a deviation in the spatial distribution of a taxon relative to one or more niche variables at a given spatial resolution then we define that as representing an interaction. We must now ask—do these interactions represent ecological interactions in the standard sense? By ecological interaction here we refer to the standard micro interactions, such as predation, mutualism, commensalism etc. The first task is to determine if this empirical interaction has a natural biological interpretation. This is related to criterion 6 of Bradford-Hill. The naturalness of the interpretation depends on the labels associated with C and the X_α. With a hypothesis in hand as to the potential ecological interaction we may then determine the degree to which the ranked list based on macro data represents known results about micro interactions, and also to what degree it offers new predictions that may subsequently be checked with a suitable experimental protocol.

An illustrative example, that has been used frequently in applications of the methodology (Stephens et al., 2009, Stephens et al., 2016), is that of predicting hosts of a zoonosis, host range being an important parameter in understanding its biology, its transmission cycle and what are appropriate potential public health interventions. Although the interaction of interest here is that of host-vector, there are multiple facets to this complex interaction which can be validated in different ways, depending on which part of the transmission cycle is used. Thus, the micro interaction where the potential host is a blood meal for the vector is a *necessary* co-occurrence condition that there is an interaction pathogen-host/pathogen-vector. Similarly, if a potential host is positive for a pathogen then it must have co-occurred at some point with the vector of that pathogen. Co-occurrence is then a *necessary* condition for the pathogen to be passed from vector to host and vice versa. Infection events are the analogues here of predation events—

a "micro" interaction. The question is then: what imprint, if any, do these micro interactions leave on the macro distributions of the involved species?

We hypothesise that the host-vector micro interaction will result in any host being a relevant, favourable niche dimension for the vector. This implies that, as a niche dimension, there will be more "attraction" between hosts and the vector versus nonhosts and the vector, with the intuition that more co-occurrence will, all else being equal, lead to more potential micro interactions and, as a consequence, a systematic sampling of different potential hosts will yield the result that the probability of finding an infected individual of one potential host versus another is higher for a species with a higher value of $\varepsilon(C|X_\alpha)$ or of $s(X_\alpha)$. This presumes that sampling can be sufficiently intensive so that a sufficient number of individuals of each potential host species are collected. However, to increase our statistical power, we may go further and consider collections of species. Thus, with the same logic, if we divide the ranked list into groups, we would expect more species to be found to be positive in the group with the highest values of ε versus the lowest. This just accounts for the fact that a species that co-occurs significantly with the vector and is widely distributed should lead to more positives than a species that does not co-occur significantly and is restricted in its range. Note that this analysis is not concerned with whether or not a potential host could be physiologically able to be a host. Potential hosts may be infected by a vector in the laboratory even though they never encounter one another in nature. Note also that this does not simply imply that the most widespread species will be most likely to be the most highly ranked. This has been checked explicitly in several examples where the ranked list by ε and the ranked list in terms of pure geographic coverage are radically different, with the former leading to a much more precise prediction model.

Note, at this level, we are considering only two labels: vector and potential host = YES/NO. However, there are many other potentially relevant labels, such as the competence of the potential hosts, or if certain potential hosts are blood meals only at a certain time of the year etc. that would be relevant for determining their role in the transmission cycle. It is for these reasons that we added the caveat "all else being equal" above.

We can think analogously about the question of predation. If we take a predator, C, we can rank all potential prey species, X_α, by $\varepsilon(C|X_\alpha)$ or $s(X_\alpha)$. Co-occurrence is a necessary condition for predation, and we hypothesise that prey species will be an important niche dimension for the predator in that there will be more "attraction" between prey species and the predator versus non-prey species and the predator. Again, we apply the intuition that more co-occurrence will, all else being equal, lead to more potential micro interactions and, as a consequence, a systematic sampling of different potential preys will yield the result that the probability of finding a confirmed prey from one potential prey versus another is higher for a species with a higher value of $\varepsilon(C|X_\alpha)$ or of $s(X_\alpha)$.

In the case of both vector-host and predator-prey, the hypothesis is that species, X_α, that represent important niche dimensions for the target species C should be correlated with more micro interaction events. In other words, that the micro interactions leave an imprint at the macro level due to the fact that they are associated with important niche dimensions. Of course, it is not required that in all circumstances micro interactions should leave an imprint at the macro level.

An important element to emphasise here is that the macro level characterisation of micro level interactions is a statistical inference. This occurs ubiquitously in epidemiology, where we may infer a link, for instance, between diabetes and obesity, or, with an even more indirect link, between diabetes and socio-economic status. The macro link between these is an imprint of an underlying set of micro physiological events. However, being a statistical inference, the correlation is not 100%. Not all diabetics are obese and not all obese are diabetics. Not all poor people are diabetics and not all diabetics are poor. Similarly, not all species, X_α, that have a large value of $\varepsilon(C|X_\alpha)$ or $s(X_\alpha)$ with a predator are by definition prey species. A principle reason for this is the existence in ecology of multiple micro interactions, and hence multiple labels, that influence their spatial distributions and therefore contribute to the overall macro interaction. Thus, a predator distribution may also be affected by potential competitors, or climate, or many other factors. Both considerations of the ecological significance of the labels for the species involved, as well as the inclusion of other potentially correlated niche variables, can help us further refine a list of micro interaction candidates, as we will see below.

One last point concerns how we would identify micro interactions in the first place. From what data? Are the micro interactions to be exhaustively tested? If so, how are we sure we have identified all interactions both direct and indirect? For any given micro interaction between a taxon C and a set of other taxa X_α, $\alpha \in [1, m]$, all m potential interactions must be checked observationally. If we consider n taxa, C, then we might think that the number of potential micro interactions is nm. However, this too is potentially a grave underestimate, as any given taxon has multiple labels and each label can represent a different micro interaction. If we cannot exhaustively check all interactions how may we predict them using a model? Independently of the type of model used its performance must be evaluated. To do so we must always start with some benchmark. What should that benchmark be and how do we make a precise specification of it? A natural benchmark is to make a hypothesis that the micro interaction only manifests itself uniformly within a particular set of candidate taxa, as identified using one or more labels that we believe to be important for the presence of the interaction. Such a restriction is equivalent to a Bayesian prior. For instance, for prey species of the bobcat, one may restrict to mammals, or lagomorphs. In the first case we would miss prey species that are not mammals, such as some birds. In the second, we would also miss mammal preys other than lagomorphs, such as rodents. In modelling terms, such a strategy introduces, by definition, false negatives. The safest bet of course is to consider all species, as then all possible candidates are included and there is no underlying bias from a chosen prior. By uniformly, we mean that the prior probability that the micro interaction exists is equal for every candidate taxon. For this probability we may have pre-existing information about the species that interact, such as a set of known preys of the bobcat or known hosts of a zoonosis. However, the quality of this benchmark depends on how well the set of observed interactions represents the full set of underlying interactions.

Supervised versus unsupervised learning models

In statistical modelling terms, the above methodology using ε to predict interactions is a form of unsupervised learning, as no information whatsoever about the class we wish to predict—those with a particular micro interaction, such as parasitism or predation—enters into the model creation. The model is based only on the logic: micro interactions between taxa lead to a niche association between them which in its turn leaves an imprint in the spatial distributions of the taxa.

If we are to use a supervised learning model however, then we must have some data that the model can learn from. In this case any list of potential candidates for a given micro interaction that enter into the model have to be labelled as such if they are already known cases. Thus, for a predator-prey interaction, we use the label *PREY = YES* for the known prey species on our list of candidates. We then need other data to be used as predictors. Naturally, we can use the already defined parameters, $\varepsilon(C|X_\alpha)$ and $s(X_\alpha)$, but we may also appeal to other labels than PREY if they are available. These labels could indeed already have been used in terms of a model selection, where sets of labels were omitted, as discussed in the section entitled "Prior probabilities and model selection." Which labels to use depends on several factors—first and foremost, are they available? Potentially relevant labels, such as those associated with phenotypic characteristics, or trophic guild, are not widely available, at least not at the level of covering large numbers of species. However, a set of labels that is always available is that corresponding to Linneaean taxonomy. A species name, or indeed any higher order taxon, is a shorthand notation for a host of characteristics that make that species, or higher taxon, distinguishable from other species, or higher taxa. We would argue that implicit in these taxonomic names are associated labels that are relevant for many micro interactions, as well as identifiers as niche dimensions.

With a set of labels in hand we may take a list of N species, already labelled with respect to the micro interaction of interest, I, and create a supervised learning model. The predictors are labels, X_i, other than the label, C_I, associated with the micro interaction. We may then use the Bayesian framework of the section entitled "A Bayesian framework for analysing ecological interactions," and, in particular, the NBA. Now, however, we are using no spatial information whatsoever, just the labels of the species. For a set of predictor labels $\mathbf{X}$, we may determine the overall score where

$$S(C_I|\boldsymbol{X}) = \sum_{i=1}^{m} s(X_i) + ln\left(\frac{P(C_I)}{P(\overline{C_I})}\right) \tag{21}$$

$s(X_i) = \ln(\frac{P(X_i|C_I)}{P(X_i|\bar{C}_I)})$ is the contribution ("score") to the overall $S(C_I|\mathbf{X})$ from the label X_i. If $s(X_i) > 0$, < 0 then the label X_i contributes positively/negatively to the class C_I. $P(C_I) = N_{C_I}/N$, with N_{C_I} being the number of species in the list labelled with the micro interaction label. $P(\overline{C_I}) = 1 - P(C_I)$ is the fraction of species on the list that do not have the micro interaction label. $P(X_i|C_I) = N_{C_I X_i}/N_{C_I}$, where $N_{C_I X_i}$ is the number on the list that have the micro interaction label and the label X_i. The model may be trained on a subset of data and then performance measured on a test set. However, the true out-of-sample test would be to use the model to predict from among non-identified species which are most likely to have micro interactions with the target and then make a systematic and representative geographic sampling to determine which species are correctly identified.

An example, that we will consider in more detail in the section entitled "Identifying specific ecological micro interactions," is where the label C_I represents the class of known prey species of a given predator, such as the bobcat, the specific interaction being predation of the bobcat on other species. What about the labels X_i for the candidate preys? As mentioned, there are myriad labels of possible relevance. The set we will use in our examples are Linneaean classifications at the level of kingdom, phylum, class, order, family and genus. Thus, *Sylvilagus*, *Lutzomyia* and *Macadamia* are potential labels at the genus level, while *Mammalia* and *Aves* would be two labels at the class level. Thus, the total score for a given candidate prey represented by a classification $\mathbf{X} = (X_{kingdom}, X_{phylum}, X_{class}, X_{order}, X_{family}, X_{genus})$, where X_{class} denotes in which taxonomic group is the candidate prey species, would be

$$S(C_I|\mathbf{X}) = s(X_{kingdom}) + s(X_{phylum}) + s(X_{class}) + s(X_{order}) + s(X_{family}) + s(X_{genus}) + \ln\left(\frac{P(C_I)}{P(\bar{C}_I)}\right) \quad (22)$$

The N species may be ranked with respect to this score function and a suitable performance measure, such as a confusion matrix or a ROC curve calculated.

In using supervised learning, we are subject to biases that are not necessarily present in the unsupervised models. For instance, the representativity of the set of species labelled with C_I could be a cause for concern. If observations of the micro interaction I have been biased towards a certain subset of the N species, say mammalian preys of the bobcat have been much more studied than non-mammalian preys, then the model results will be subject to this bias. However, as this bias would potentially be present in both training and test sets model performance might not be affected. For this reason, we emphasised using the model in a completely out-of-sample set wherein the most likely preys that have not been observed as such are studied.

Naturally, the performance of the supervised model may be compared and contrasted with the unsupervised model that uses only spatial information. However, we may also adopt a meta-model viewpoint, constructing a model that is a mix of supervised and unsupervised Bayesian models, using labels, such as the taxonomic labels mentioned above, as well as spatial information through $s(X_\alpha)$ or $\varepsilon(C|X_\alpha)$ and compare its performance to either the supervised or unsupervised model. We will do this explicitly in the section entitled "Identifying specific ecological micro interactions." The advantage of this is that the unsupervised model, considering only single species as niche variables and without further information on their labels, identifies only an overall interaction that may be due to the superposition of several micro interactions. On the other hand, the supervised model identifies relevant labels but does not account for the fact that a micro interaction can only take place if there is a co-occurrence. A mixed model can compensate for the individual defects of each separate model type.

Interpretation issues

Our proposal is that biotic interactions between organisms may be statistically inferred from their relative positions in space and time, where by statistical inference we mean to reach a conclusion based on evidence (a statistical ensemble of observations) and reasoning. A possible objection to this thesis is that: there are many other factors that affect the relative positions of biota in space and time, other than biotic interactions, which can act as confounders. A second objection, in the case where we use point collection data to model the positions and distributions of species, is that such data is biased and therefore not reliable. Our answer to this objection is two-fold: first, and most importantly, does the model make predictions that can be tested and what is the model performance? Secondly, the same

biased data is already being used ubiquitously in fundamental niche modelling. In other words, if the data bias is sufficient so as to invalidate its use for modelling a biotic niche variable Xα then it is also inadequate to model a target taxon C.

In reference to the first objection, we believe our methodology in the sections entitled "A Bayesian framework for analysing ecological interactions" and "A Bayesian framework for casual inference" provides a framework for iteratively adding the presence/no presence or absence of different variables in order to test which is playing the dominant role, the intuition being that the more causal is a factor, the more important it is likely to be as a predictor. Of course, confounding is an element that may be mentioned in the context of the analysis of any complex adaptive system, be it ecological, social, physiological etc. It is impossible to even list the full set of factors that may influence the presence of a given species. It is also impossible to determine unambiguously, from a statistical inference, that there does not exist a factor that has not been accounted for that is relevant or a confounder. To do so we would first have to have a consensus on the full set of potential confounders. We would then have to have a data representation of them in order to include them in the models. We would then have to check them one by one using our Bayesian inference framework, or other, to determine their relative importance and to see if one variable con-founds another. An alternative and more sensible approach is to make a hypothesis about a specific, potential confounder for the appear-ance of a biotic interaction in macro distribu-tions, such as "shared history, geography, mi-gratory patterns, climate preferences...," pro-vide the data that allows us to include that fac-tor in our model, and then test the hypothesis. To show the feasibility of this approach, we explicitly show in the section entitled "Inferring causality and identifying confounders" an example where we can prove that rather than climatic factors being a confounder for apparent biotic interactions, on the contrary, it is biotic factors that are confounders for abiotic effects.

One factor that can cause problems however, is the restriction to time independent data. Not only because micro interactions themselves may be time dependent, but because we are taking the distributions of biota as "snapshots." Thus, two taxa might interact according to our definition, but this may be a result of the fact that, although the time trajectories of the distributions show no interaction across time, the snapshot shows a state that gives the appearance of an interaction. Without time dependent data this cannot be analysed. In that sense, by using point collection data without a time dependence, we are assuming that species are in "equilibrium." Of course, in the case of multiple niche variables it is very, very unlikely that all snapshots of all species pairs exhibit interactions by accident.

A Niche versus a network perspective

Up to now, we have emphasised macro interactions as being fundamentally related to the concept of a niche. In other words, if two taxa C and X_α exhibit an interaction in terms of their co-occurrences it is because X_α is an important niche dimension for C. We have thus framed everything in terms of "binary" interactions—between one taxon and one or more taxa, i.e., between C and X_α, or C and $\mathbf{X}$. However, our methodology has also been extensively applied at the community level, where we consider multiple target taxa, C, such that we have a set of target taxa and a set of niche taxa $\mathbf{X}$. Each C and X_α can be represented as nodes of a network and $\varepsilon(C|X_\alpha)$ used to weight links between these nodes. The resultant network we term a Complex Inference Network (CIN) (González-Salazar and Stephens, 2012; Stephens et al., 2009). The term "inference" is used as it does not represent specific, previously identified micro interactions between taxa, such as in a food web, but, rather, represents the set of inferred interactions between the taxa as defined by our ε diagnostic in equation (5). To what degree any given interaction represents an underlying micro interaction is discussed amply in previous sections. The advantage of CINs is that they allow for an analysis of interactions at the community level. For instance, in Figure 1 we see the network that results from using the Ecological Community functionality of the SPECIES platform[3] to analyse the relation between two predators—the bobcat and the coyote—as target species, and species of the order Lagomorpha, Artiodactyla and Rodentia as potential prey species. This example will be of relevance for the section "Some representative results." For the network links, only links corresponding to values of $\varepsilon > 8$ are considered thus representing the most important possible positive interactions.

[3] http://species.conabio.gob.mx

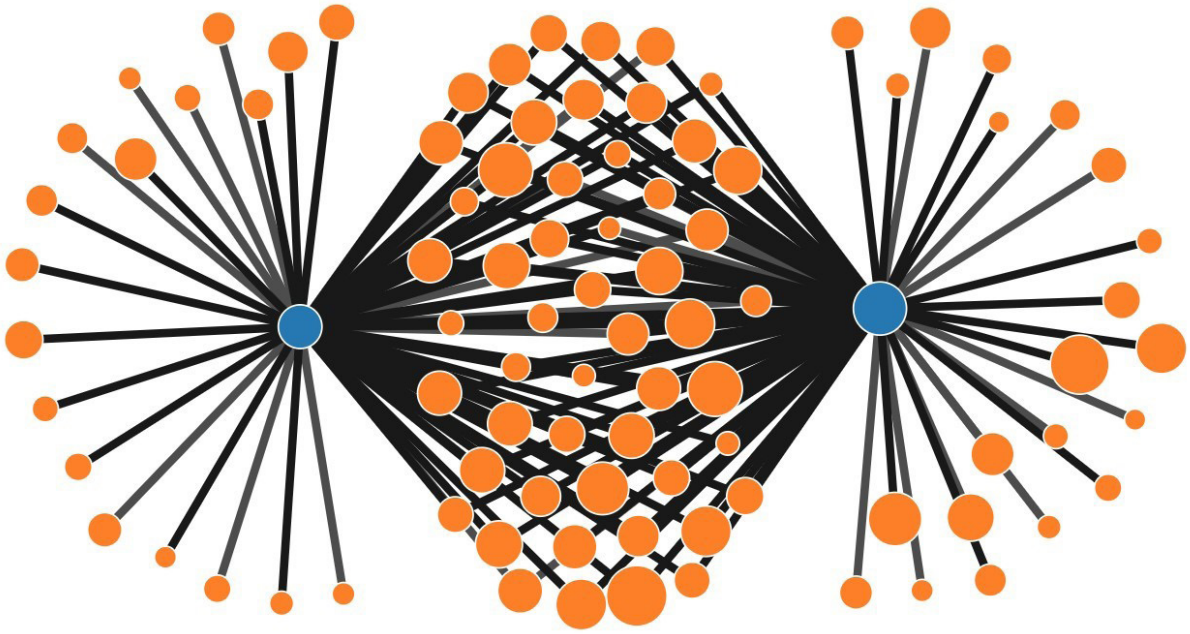


Figure 1. Complex Inference Network between the bobcat (blue circle to the left) and the coyote (blue circle to the right) and the set of potential prey species (orange circles) from the orders Lagomorpha, Artiodactyla and Rodentia. Only the most important interactions corresponding to $\varepsilon > 8$ are shown.

Some representative results

The success of any scientific theory must be judged on its ability to explain and predict phenomena. Theory can be used to explain and further understanding of already existing data or may be used to propose hypotheses and associated predictions that require new data to be collected. Here, we will present a small, representative sample of the results that have been obtained using the present formalism that comply with all these requirements. We will show that: i) biotic factors can be generally and successfully introduced into species distribution and niche modelling; ii) confounding can be analysed by comparing the simultaneous presence/absence of multiple niche factors; iii) specific ecological micro interactions can be inferred from macro (point collection) data. We could exhibit the advantages of our approach in myriad examples, several of which are already in the literature, trying in each example type to show what links it to other examples and what makes it different. Here, we can only try to show a pair of representative examples that show the power of our methodology hoping that they show the applicability to any other possible use case. We also hope that the reader will test these ideas in their own use case of interest using the powerful SPECIES platform (Stephens et al., 2019) that implements our methodology in an open, easy-to-use, on-line environment that uses data from the Sistema Nacional de Información de la Biodiversidad (SNIB[4]) and, more recently, North American data from GBIF.[5]

Including biotic factors into niche and species distribution predictions

Perhaps the least controversial application of our methodology is to include biotic factors into the characterisation of species' niches and their associated geographic distributions, without discussing any subsequent interpretation in terms of micro ecological interactions. In this case we use the Bayesian models of 4 to determine $P(C|\mathbf{X})$, where $\mathbf{X}$ can be any combination of abiotic or biotic variables. To create models, we use the SPECIES platform,[6] which incorporates North American point collection data from both the SNIB and GBIF databases. The SNIB itself contains data on over 57,000 species.

As a first example, we will consider the construction of the niche of the bobcat *Lynx rufus* and use it to illustrate many of the points we have discussed

[4] http://www.snib.mx/
[5] https://www.gbif.org/
[6] http://species.conabio.gob.mx

above. The data used for this example are from the SNIB and consider data for Mexico only. The first question that arises is: which niche dimensions should be included in **X**? In principal, in SPECIES we may include in as covariates/niche variables for the bobcat all species other than the bobcat, as well as WorldClim as a set of abiotic variables, and other environmental layers. As discussed above, by bringing all variables to the same spatial resolution and making all variables binomial, all potential niche dimensions can be fairly compared and contrasted. In Figure 2 we see the relative performance of three models: one including only mammal species as biotic niche variables, one including only abiotic variables (WorldClim) and another including both. What is shown there, for the two groups of variables chosen, is the total score from equation (10) in subsets (deciles) of spatial cells. All cells are ranked by their total score and then divided into deciles. Thus, decile 10 represents the 10% of spatial cells, wherein the total score is highest by category, corresponding to those conditions that are most favourable to the presence of the bobcat. Decile 9 is the next 10% of highest score cells—more favourable than decile 8 but less favourable than decile 10. Decile 1 is the 10% of cells with lowest scores, the least favourable cells, or most "anti"-niche.

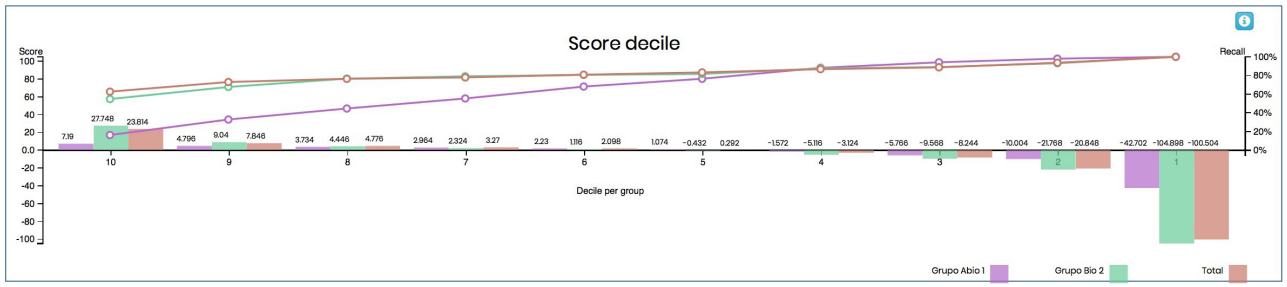


Figure 2. Performance of predicted distribution models for the bobcat based on abiotic variables only (WorldClim), biotic variables only (mammals) and a combination.

The relative performance of the models is evaluated using: Recall = True positive/(true positives + false negatives). In SPECIES, model performance is gleaned from a 70%-30% training-test split which is repeated five times. Thus, any performance metric is an average over five such iterations. Clearly the performance of the biotic model is far superior to that of the abiotic model, with the combination of the two being even better. Simply put, mammals as niche variables are much more predictive than climatic data for presence of the bobcat. We should emphasise here that this exercise, as implemented in the SPECIES platform can, quickly and efficiently, be used to validate or invalidate the Eltonian Noise Hypothesis (Soberón and Nakamura, 2009) for any of the many thousands of species present as target taxa. Indeed, one will quickly convince oneself that the Eltonian Noise Hypothesis is rarely valid when all potential biotic interactions are considered and that biotic factors in general are more important niche dimensions than climatic factors, with the latter characterising the anti-niche more than the niche, i.e., they affect much more the unfavourability of a place than its favourability.

Of course, WorldClim data is only a representative subset of all abiotic factors, not an exhaustive set. Similarly, mammals only represent a subset of biotic variables. However, using our methodology, in the SPECIES platform we may consider any arbitrary subset of potential niche variables and in any combination. Thus, we may compare a particular abiotic factor with a particular biotic one, or any group thereof, or compare all 57,000 biotic variables with all abiotic variables. Even the mere form of the distribution of scores as a function of decile is informative. Note that the relative contribution of climate is much more significant in the first deciles than in the tenth. This confirms that climate affects only weakly the niche of the bobcat but is much more influential in determining the anti-niche. A subsequent analysis of the highest ranked versus lowest ranked mammals shows that known prey species of the bobcat appear with high values of ε (González-Salazar et al., 2013).

Besides comparing biotic and abiotic contributions, we can also compare different types of biotic variable. For instance, in Figure 3 we see the performance of a model that includes mammals as biotic factors with another that includes the order Magnoliales—a class of flowering plant—as biotic factors. These are chosen as an example of a group of biotic

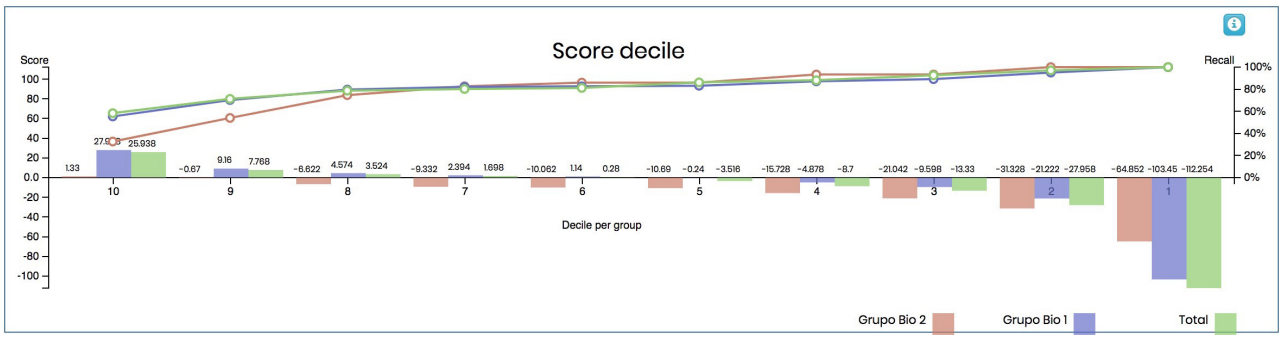


Figure 3. Performance of predicted distribution models for the bobcat based on two classes of biotic variables: Group Bio 1 = Mammalia and Group Bio 2 = Magnoliales, and their combination.

factors where there is no a priori reason to expect any relevant direct micro interaction with the bobcat. As we can see, the performance of the model using mammals is much superior to that using Magnoliales, as might be expected for flowering plants as predictors of the bobcat distribution. From the form of the curves we may see that the plants play no role as niche variables for the bobcat, although their effect as anti-niche variables is apparent. Of the 120 species of Magnoliales as potential niche factors, only two have a positive score and have $\varepsilon > 1.96$. How should we interpret the fact that the presence of Magnoliales is negatively correlated with the presence of the bobcat? That there is a repulsive interaction between the bobcat and these flowering plants? That they are, for some reason, in competition? Of course not. First, we must interpret the apparent macro interaction from the point of view of biological plausibility. Second, we must check for confounding and causality to determine the extent to which climate confounds the negative interaction of these flowering plants with the bobcat. This can be done using the methodology presented in the section "A Bayesian framework for causal inference."

Inferring causality and identifying confounders

We will show that our methodology is capable of untangling the underlying causal relationships between niche variables in the context of a specific example: the characterisation of the niche of the bobcat; considering particularly the relations between a known prey species of the bobcat—*Sylvilagus floridanus*, *Microchloa kunthii* a known food source of *S. floridanus* and, finally, climate, as proxied by WorldClim. An important reason for doing this arises from the criticism that a particular biotic niche factor may be confounded by climatic or other variables. In other words, that the reason why two species co-occur is because they share the same habitat preferences rather than that there is a particular biotic interaction between them (Ovaskainen et al., 2010; Royal et al., 2016).

In this case we consider $P(C|X_\alpha X_\beta X_\gamma)$ and analogously for $\varepsilon(C|X_\alpha X_\beta X_\gamma)$, where $X_\alpha = 0, 1$ represents presence/no presence of *Sylvilagus floridanus*, $X_\beta = 0, 1$ represents presence/no presence of *Microchloa kunthii* and $X_\gamma = 0, 1$ will range over the decile ranges for two significant climatic variables from WorldClim: Mean Annual Temperature (Tmp R) and Mean Annual Precipitation (prec R), with decile 10 representing the highest temperature/precipitation and decile 1 the lowest. Thus, for a given cell, $X_\gamma = 1$, if the temperature range denoted by X_γ is present in the cell and zero otherwise. For each of the ten temperature and precipitation variables there is one presence/absence variable.

In Table 1 we analyse the combined effects of the presence/no presence/absence of the two biotic factors and two abiotic factors in terms of heatmaps. The colour with the corresponding scales corresponds to $P(C|X_\alpha X_\beta X_\gamma)$, the probability of a cell having a presence of the bobcat given the corresponding configuration of niche variables. Remember that this is not an absolute probability to detect a bobcat in a given cell, but a relative measure based on point collection data. The numbers in each cell of the graph are the corresponding $\varepsilon(C|X_\alpha X_\beta X_\gamma)$ values, which allow us to determine if a given value of $P(C|X_\alpha X_\beta X_\gamma)$ is statistically significant or not. They also allow us to infer the coverage of the corresponding niche variable combination. The sign of ε also allows us to infer if the interaction is positive or negative with a positive/negative value indicating that $P(C|X_\alpha X_\beta X_\gamma) >, < P(C)$—the null hypothesis. By reading vertically from top to bottom we can see the effect of decreasing temperature or precipitation, concentrating on the presence variables

Table 1. Probability $P(C|X_\alpha X_\beta X_\gamma)$ and $\varepsilon(C|X_\alpha X_\beta X_\gamma)$ for the bobcat with respect to a prey species, a food source of that prey species and climate.

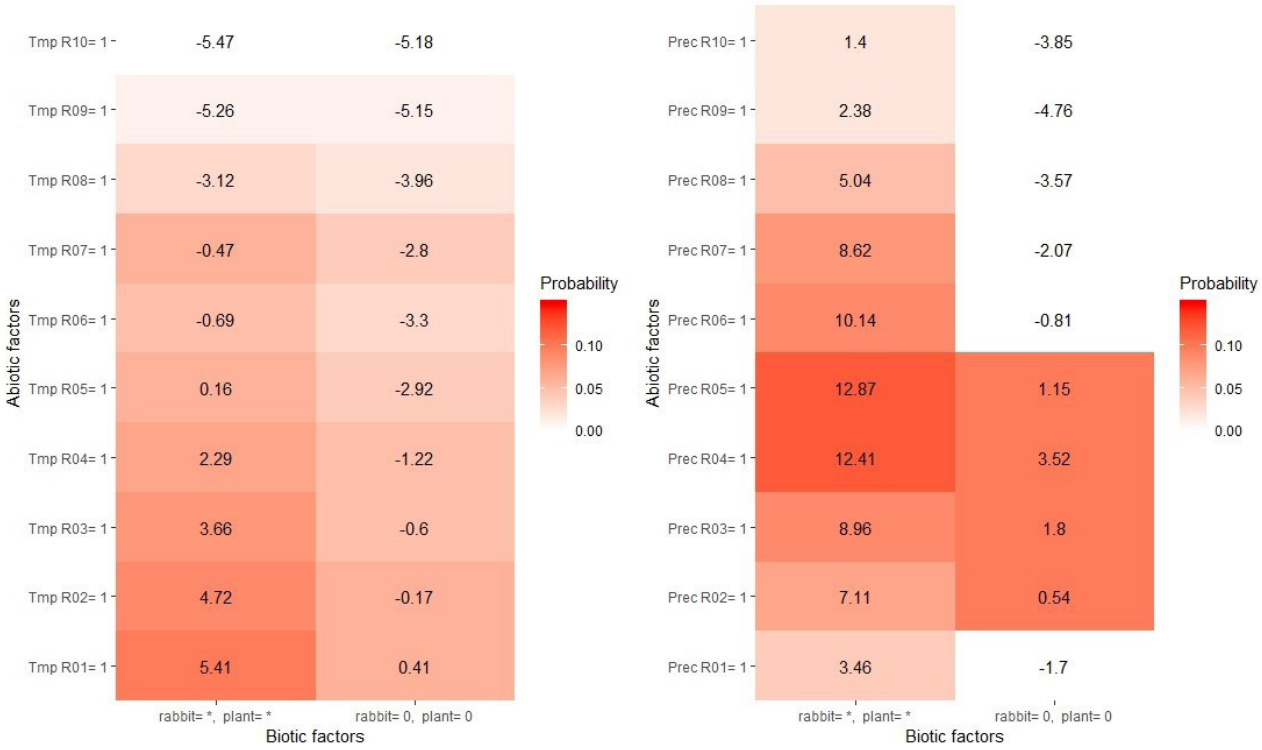
for temperature or precipitation as denoted by $R_i = 1$. The values $R_i = 0$ correspond to the absence of the corresponding climatic range. Similarly, by reading horizontally from left to right, we may see the effect of increasing the overall presence of the biotic factors from both not present to both present.

What is immediately clear is that the biotic factors play a much more important role than climate in determining what are favourable conditions, i.e., they play a preponderant role in determining the niche of the bobcat. This is fully consistent with our findings in the previous section, where we found that biotic factors were more generally associated with determining the niche, while abiotic factors were more relevant for the anti-niche. This is equally true here. The gradient in the probabilities left to right is much greater than the gradient top to bottom. In the absence of both biotic factors, corresponding to cells without a presence of either species, there is no temperature range that corresponds to a statistically significant positive niche factor, where the probability to find the bobcat is greater than the null hypothesis. On the other hand, the temperature ranges R10-R6 are associated with probabilities to find the bobcat that are less than the null hypothesis and correspond to a statistically significant negative interaction between the bobcat and these higher temperatures. In fact, in ranges R10 and R9 there is very little probability of finding the bobcat, independent of whether the biotic factors are present or not.

Turning to conditions where the plant is present, but the prey is not, we see that once again the probabilities to find the bobcat are consistent with the null hypothesis, as no temperature range is associated with favourable conditions. However, when the prey is present and the plant not, we see that the bobcat is present at a frequency statistically significantly greater than the null hypothesis for almost all temperature ranges. Only the highest temperature ranges, R10=1 and R9=1, correspond to conditions such that, despite the presence of the prey, the bobcat is present less than would be expected by the null hypothesis. The same is true when both biotic factors are present. We can also note that the presence of the plant in the absence of the prey is less significant in indicating the presence of the bobcat than when the prey is present and the plant not. So, we may note that lower/higher temperatures are associated with more niche/anti-niche conditions, while presence of a food source of the prey of the bobcat is a positive niche factor but less significant than the presence of the prey itself. However, the presence of both together leads to even more favourable niche conditions. The same considerations apply for precipitation. The presence of the prey is more important than the presence of the plant, which in turn is more important than precipitation. We can see that both high precipitation (R10=1, R9=1) and very low precipitation (R1=1) are anti-niche conditions.

That we can isolate the effects of confounding in this case can be even more plainly seen in Table 2, where we compare $P(C|X_\alpha X_\beta X_\gamma)$ for $X_\alpha = 0$ (rabbit not present) and $X_\beta = 0$ (plant not present) against $P(C|X_\alpha X_\beta X_\gamma)$ for $X_\alpha = 0$ or 1 (rabbit present or not present) and $X_\beta = 0$ or 1 (plant present or not present). In this latter case $P(C|X_\alpha X_\beta X_\gamma)$ represents the marginalised probability $P(C|X_\gamma)$, where X_γ is a purely climatic factor. Similarly, for $\varepsilon(C|X_\alpha X_\beta X_\gamma)$ and $\varepsilon(C|X_\gamma)$. By

Table 2: Probability $P(C|X_\alpha X_\beta X_\gamma)$ and $\varepsilon(C|X_\alpha X_\beta X_\gamma)$ for the bobcat in the absence of biotic factors, and $P(C|X_\gamma)$ and $\varepsilon(C|X_\gamma)$, where X_γ represents just climate.



marginalising we are omitting the direct influence of the biotic factors. This is equivalent, in the Bayesian sense, of a model selection where only abiotic variables are chosen, as is done in standard niche and species distribution modelling. Far from it being the case that abiotic factors are confounders for biotic ones, here we see that, on the contrary, the abiotic factors are confounded by the biotic factors in terms of determining the niche. As can be seen comparing the two cases, the same tendency is observed in the case of conditioning on the no presence of rabbit and plant versus with no conditioning, in that colder temperatures and moderate precipitation are more favourable for the presence of the bobcat. However, the strength of the interaction is very different. If we consider only abiotic variables, the apparent probabilities for presence of the bobcat and their corresponding statistical significance are greatly enhanced, due to the fact that climate is correlated with the biotic niche factors. In summary, we can characterise the niche of the bobcat, quantify the contribution of each niche factor and deduce their correlations. Presence of the prey species and/or the presence of one of the prey species food sources are positive niche factors, as well as non-extreme temperatures and precipitation. We can deduce the causal chain of factors, noting that the factor closest causally to the bobcat—its prey—plays a much more important role than the factor which is indirectly linked—the prey's food source, which, in turn, is more important in determining the niche than the climate. This ranking of the relative importance of the niche factors: prey > prey food source > climate, is as you would expect for a vagile mammal such as the bobcat. Thus, the direct interaction between bobcat and prey, is stronger than the indirect interaction between bobcat and food source of the prey which, in its turn, is stronger than the even more indirect interaction between bobcat and climate. Moreover, we can see that climate is confounded by the underlying presence of relevant biotic factors. This is *prima facie* evidence that standard fundamental niche modelling, based only on abiotic variables, does not represent the true statistical relationship between climate and species distributions. Rather, the relationship between climate and bobcat reflects the presence of important biotic confounders such as the bobcat's prey species. Thus, fundamental niche models need to include biotic factors with subsequent analysis of the potential confounding between one type of factor and another.

Identifying specific ecological micro interactions

As we have used throughout an empirical characterisation of an interaction, defined via a deviation in the spatial distribution of a taxon from some "non-interaction" null hypothesis, it is not clear to what degree this definition of an interaction accords with the ecological definition in those cases where an ecological micro interaction has been identified and characterised. We will consider a test case where a verified ecological interaction is known[7] and show that our empirical characterisation accords with the ecological one, using the unsupervised learning technique presented in the section "Predicting interactions." Specifically, we consider the prediction of prey species of the bobcat[8]

We have emphasised the importance, from a Bayesian perspective, of model selection: which variables are to be included in as possible prey species in the first place? We will consider three sets, in order of increasing bias: all species (53,722 species), mammals (496 species) and lagomorphs (14 species). By bias here, we mean that in the second and third groups we include the assumption (Bayesian prior) that preys of the bobcat are only to be found among mammals, or among lagomorphs, respectively. Thus, for each group we rank the included species by ε, with the hypothesis that those species most likely to have been identified as preys of the bobcat will have higher values of ε, as they have a higher rate of co-occurrence and are more disperse.

There is no information that enters in this unsupervised model other than the distribution of biota, as proxied by point collection data. In particular, we use no information about any labels that might be at hand. As discussed in the "Clarifying interactions," potentially relevant labels could be big/small, slow/fast, nocturnal/diurnal, terrestrial/aerial etc. Such labels are not widely available in point collection databases for large numbers of species. However, one set of labels that is universally available are the standard Linneaean taxonomic labels. Indeed, these were used implicitly in the choice of the three proposed models with *class = mammalia* being used in one case and *family = lagomorpha* in the other. We will use those taxonomic labels here.

Our examples used georeferenced point collection data from the SNIB with corresponding taxonomic labels. A bibliographic search was conducted to determine the set of known (identified) prey species of the bobcat in Mexico. Sixty-seven prey species were identified. The data was divided into 70%/30% training/test and for each training set scores for each taxonomic label calculated using equation (22). These scores were applied to the test set and model performance was then calculated using the area under the ROC curve (AUC). One hundred iterations of this process were carried out and an average performance calculated along with its standard error. Finally, combined models were created—one that used the score from the taxonomic labels as well as $\varepsilon(C|X_\alpha)$ and $s(X_\alpha)$ and another that used the score from the taxonomic labels and $\varepsilon(C|X_\alpha)$ only. The results can be seen in Table 3, where the correct class is *Prey* = 1.

What these results clearly show is that a statistical inference of this particular micro interaction, using only ε as a measure of interaction, is very accurate, with AUCs of 0.98, 0.91 and 0.95 for the *all*, *mammal* and *lagomorph* groups. As can be seen, it is actually much better than the supervised model in the case of mammals and lagomorphs. Of the 67 identified prey species, which corresponds to 0.12% of the total number of *all* species, 22.4% can be found in the top 0.1% (54 species) of the ranked list by ε, while 70% are found in the top 1% and 95.5% in the top 10%. Similarly, for the mammals only model that incorporates 496 mammals: 40% of known prey species can be found in the top 10% by ε. In Table 4 we see the list of the top 0.1% of species from the *all* list. The statistical ensemble here is composed of 26,944 cells of dimension 16 × 16 km. $n_i = 238$ is the total number of cells with presence of the bobcat, n_j is the total number of cells with a presence of the potential prey species, j, and n_{ij} is the number of cells with a co-occurrence.

This list gives good insight into why it is perhaps difficult to accept that ecological interactions can be identified using point collection data. In terms of statistical inference, the model performance is out-

[7]By "known" here we mean that we know it exists with respect to a given label and have a set of examples. This does not imply, however, that those examples necessarily form a complete set, or we have a complete set of relevant labels.

[8]We have also carried out a similar analysis for other examples: i) pollination—*Leptonycteris curasoae*, a bat species that pollinates agaves; ii) pollination—*Dalechampia scandens*, a twining vine that rewards insect pollinators; iii) mutualism—*Aechmea bracteacta*, a tank bromeliad that provides an ideal habitat for the development and refuge of aquatic and terrestrial organisms; iv) facilitation—*Neobuxbaumia mezcalaensis*, a plant that depends on nurse plants to have a favourable microhabitat and avoid humidity loss due to direct contact with the sun. Detailed results will be presented in another publication. Note that the SPECIES platform can be used to consider and validate any other example where appropriate date exists.

Table 3: Model performance for 4 model types: i) ε (unsupervised); ii) taxonomic labels (supervised); iii) ε, s and taxonomic labels (meta-model); and iv) ε and taxonomic labels (meta-model).

Lynx/prey	Total candidate prey species	Total true positives	AUC (ε)	AUC (Taxonomic labels score)	AUC (ε, s(X) and taxonomic labels score)	AUC (ε and taxonomic labels score)
Total	53722	67	0.98	0.99 Std error 0.00	0.99 Std error 0.00	0.99 Std error 0.00
Mammalia	496	50	0.91	0.70 Std error 0.00	0.71 Std error 0.00	0.71 Std error 0.00
Lagomorpha	14	6	0.95	0.78 Std error 0.02	0.94 Std error 0.01	0.94 Std error 0.01

Table 4: The top 57 highest ranked species by ε corresponding to those species with the most important interaction with the bobcat. The true positive rate in this group is 22.4% compared to the null (random) benchmark of 0.1%.

Species	nij	nj	ni	n	Epsilon	Score	Class	Order	Prey
Canis latrans	106	400	238	26944	54.75	3.7	Mammalia	Carnivora	0
Urocyon cinereoargenteus	85	535	238	26944	37.09	3.05	Mammalia	Carnivora	0
Taxidea taxus	32	87	238	26944	35.79	4.18	Mammalia	Carnivora	0
Lepus californicus	64	383	238	26944	33.1	3.11	Mammalia	Lagomorpha	1
Peromyscus maniculatus	99	871	238	26944	33.06	2.67	Mammalia	Rodentia	1
Otospermophilus variegatus	54	339	238	26944	29.61	3.06	Mammalia	Rodentia	1
Procyon lotor	56	371	238	26944	29.25	2.99	Mammalia	Carnivora	1
Tadarida brasiliensis	66	520	238	26944	28.78	2.79	Mammalia	Chiroptera	0
Sylvilagus audubonii	58	417	238	26944	28.43	2.9	Mammalia	Lagomorpha	1
Puma concolor	33	143	238	26944	28.36	3.52	Mammalia	Carnivora	0
Mephitis macroura	46	279	238	26944	27.86	3.1	Mammalia	Carnivora	1
Odocoileus virginianus	71	633	238	26944	27.78	2.65	Mammalia	Artiodactyla	0
Bassariscus astutus	45	270	238	26944	27.72	3.11	Mammalia	Carnivora	0
Sayornis saya	92	1045	238	26944	27.36	2.38	Aves	Passeri-formes	0
Thomomys bottae	51	351	238	26944	27.32	2.95	Mammalia	Rodentia	0
Haemorhous mexicanus	118	1648	238	26944	27.23	2.16	Aves	Passeri-formes	0
Conepatus leuconotus	41	236	238	26944	27.07	3.16	Mammalia	Carnivora	1
Bubo virginianus	62	519	238	26944	26.93	2.72	Aves	Strigiformes	0
Dipodomys merriami	79	814	238	26944	26.9	2.49	Mammalia	Rodentia	1
Corvus corax	110	1504	238	26944	26.65	2.18	Aves	Passeri-formes	0
Spizella passerina	96	1197	238	26944	26.39	2.28	Aves	Passeri-formes	0
Regulus calendula	89	1044	238	26944	26.39	2.35	Aves	Passeri-formes	0
Icterus parisorum	65	590	238	26944	26.31	2.63	Aves	Passeri-formes	0
Reithrodontomys megalotis	58	488	238	26944	25.97	2.72	Mammalia	Rodentia	1
Sylvilagus floridanus	62	564	238	26944	25.66	2.63	Mammalia	Lagomorpha	1
Ursus americanus	16	43	238	26944	25.46	4.2	Mammalia	Carnivora	0
Lanius ludovicianus	108	1573	238	26944	25.36	2.11	Aves	Passeri-formes	0
Accipiter cooperii	81	939	238	26944	25.36	2.36	Aves	Accipitri-formes	0

Table 4: (continued from previous page)

Euphagus cyanocephalus	59	532	238	26944	25.16	2.64	Aves	Passeri-formes	0
Peromyscus eremicus	63	604	238	26944	25.08	2.57	Mammalia	Rodentia	0
Buteo jamaicensis	119	1922	238	26944	24.87	2	Aves	Accipitri-formes	0
Colaptes auratus	81	971	238	26944	24.84	2.32	Aves	Piciformes	1
Pipilo maculatus	61	594	238	26944	24.45	2.55	Aves	Passeri-formes	0
Phainopepla nitens	67	709	238	26944	24.38	2.46	Aves	Passeri-formes	0
Tyrannus vociferans	91	1237	238	26944	24.33	2.19	Aves	Passeri-formes	0
Sayornis nigricans	92	1279	238	26944	24.12	2.16	Aves	Passeri-formes	0
Passer domesticus	107	1684	238	26944	23.99	2.03	Aves	Passeri-formes	0
Tyto alba	54	490	238	26944	23.98	2.63	Aves	Strigiformes	0
Myotis californicus	37	242	238	26944	23.95	3.01	Mammalia	Chiroptera	0
Zonotrichia leucophrys	65	692	238	26944	23.92	2.45	Aves	Passeri-formes	0
Neotoma mexicana	49	411	238	26944	23.92	2.72	Mammalia	Rodentia	1
Turdus migratorius	70	799	238	26944	23.8	2.38	Aves	Passeri-formes	0
Geococcyx californianus	73	865	238	26944	23.75	2.34	Aves	Cuculi-formes	0
Perognathus flavus	41	299	238	26944	23.71	2.88	Mammalia	Rodentia	0
Setophaga coronata	108	1751	238	26944	23.63	2	Aves	Passeri-formes	1
Auriparus flaviceps	70	809	238	26944	23.62	2.36	Aves	Passeri-formes	0
Aegolius acadicus	17	56	238	26944	23.57	3.89	Aves	Strigiformes	0

standing. However, let us examine what, apparently, are false positives on the list. First and foremost, we must be careful about judging that a false positive is just that—that the corresponding species is *not* a prey species, versus it has not yet been identified as such. For instance, *Taxidea taxus* may well be a prey species (Skinner, 1990) that has not been so identified in Mexico. There are also several bird species that are very highly ranked and therefore posited to be potential prey species but that have not been identified as such. In this case, our list represents a set of predictions for new potential preys that have not been previously discovered.

Secondly, our methodology is based on the fact that there is an interaction by virtue of the "attraction" between these species as niche dimensions and the bobcat. This does not mean, however, that the interaction has perforce to represent a predator-prey interaction with the bobcat as predator. It does not even mean that the interaction has to be direct. For instance, the strong interaction between the coyote (*Canis latrans*) and the bobcat may principally be through shared preys. This hypothesis is fully consistent with the CIN shown in Figure 1, where we see that there are many more strong interactions (potential preys) of both the coyote and the bobcat (orange nodes connecting the blue nodes of the bobcat (left) and coyote (right)) than are available to only one or the other. So, the bobcat and coyote have a substantial niche overlap in terms of their prey species. However, there are studies that show that, despite this overlap, there is little competition between them (Major, 1987), except in conditions where food resources are scarce. In principle, the nature of the interaction between the bobcat and the coyote can be analysed further using the formalism of the "A Bayesian framework for causal inference." In other words, we may consider $P(bobcat = present|coyote = present, prey = present)$ versus $P(bobcat = present|coyote = present, prey = absent)$, just as was done for the case of the bobcat in "A Bayesian framework for causal inference."

The above considered only spatial information. We can also adjoin labels and potentially improve the prediction model based only on ε. In Table 5 we see the top five species of the *all* group as ranked by ε. The ranking, using the score function as deduced by a supervised learning model, is quite different,

Table 5: Impact of taxonomic labels on ranking by ε for most important macro interactions with the bobcat.

LYNX/PREY	Rank by ($\mathcal{E}$)	Rank by taxonomic labels	Rank by Epsilon, score and taxonomic labels	Rank by Epsilon and taxonomic labels
Canis latrans	1	174	170	172
Urocyon cinereoargenteus	2	145	80	81
Taxidea taxus	3	76	102	102
Lepus californicus (prey)	4	2	2	2
Peromyscus maniculatus (prey)	5	34	39	3

with the coyote and the grey fox in particular being substantially downgraded in the list. Although these species have an important macro interaction with the bobcat, principally indirect through shared prey species, as is consistent with Table 5, they do not have taxonomic labels that are as highly correlated to known prey species, such as Lagomorpha. However, from Table 3 we see that the supervised model with taxonomic labels is much inferior to the unsupervised model for the *mammal* and *lagomorph* groups. So why does this supervised learning model performance decay so significantly? When *all* species are considered, there are taxonomic labels that are very predictive. For example, as the bobcat is a carnivore, any plant species will clearly receive a very negative score. However, when we get to the *mammal* group, the taxonomic labels within this set lose predictive power, as the bobcat has preys in multiple mammal genera, families and orders and, obviously, when we get to the *lagomorph* group the taxonomic labels lose relevance even further.

Interestingly, the spatial models alone using ε contain implicit information about the diet of the bobcat's diet! The highest ranked plants in the list of *all* species are *Muhlenbergia wrightii* (rank 61) and *Bromus carinatus* (rank 99). As both are grass species that are potential food sources for many of the main prey species of the bobcat, we see that the interaction in terms of these plants as niche dimensions of the bobcat is indirect, with the bobcat's interaction being intermediated by its prey species as confounders. Once again, using the methods of the section entitled "A Bayesian framework for causal inference," this confounding can be analysed to better understand the true causal nature of the interactions.

Note also the extremely non-random nature of the ranking of different taxonomic groups in the *all* list. As plants represent about 50% of the overall set of species the odds of not hitting a plant until the 61st place in the list if the species were distributed randomly would be astronomically small. So, an unsupervised model based only on macro interactions, as measured by ε, and without reference to any characteristics of the potential prey species, yields an extremely predictive model for identifying the micro interaction between the bobcat and its preys. This is because an important prey species will be an important niche dimension, and this will manifest itself in the species' distributions. However, ε alone is not equipped to distinguish between the different potential micro interactions that give rise to the observed macro interactions. On the other hand, suitable labels, such as the taxonomic labels used here, can identify characteristics of the known prey species, but then cannot distinguish between those that are niche dimensions that affect the distribution of the bobcat and those that do not. In other words, a lagomorph, such as *Sylvilagus brasiliensis*, has the right characteristics to be a prey but does not share niche with the bobcat, as can be seen in Table 6, where there are zero co-occurrences with the bobcat. A combination of unsupervised and supervised models is a way of including both macro level interactions and useful labels for distinguishing between different micro level interactions as contributors to the macro distributions. So, in Table 5 we see that a model that uses both ε and the scores from the taxonomic labels enhances the rank of those species—*Lepus californicus* and *Peromyscus maniculatus*—that have both an important micro interaction with the bobcat and taxonomic characteristics that are identified with known prey species, while, at the same time, suppressing those species—*Canis latrans* and *Urocyon cinereoargenteus*—that have an important macro interaction with the bobcat but do not have taxonomic labels that are consistent with that interaction having the micro interaction predator-prey as an important contributing source. Finally, considering only lagomorphs as potential prey, we see the list of all Mexican lagomorphs ranked by ε in Table 6. Note that once again the model is extremely good, with a sensitivity of 83.3% and a specificity of 87.5%.

Identifying disease hosts

We presented the case of predation above as a test case as it is distinct to what has been, up to now, the main area of application of the methodology—zoonoses. As mentioned, the transmission cycle of

Table 6: Ranking by ε of all lagomorphs.

Species	nij	nj	ni	n	Epsilon	Genus	Prey
Lepus californicus	64	383	238	26944	33.1	Lepus	1
Sylvilagus audubonii	58	417	238	26944	28.43	Sylvilagus	1
Sylvilagus floridanus	62	564	238	26944	25.66	Sylvilagus	1
Romerolagus diazi	9	19	238	26944	21.66	Romerola-gus	1
Sylvilagus cunicularius	17	147	238	26944	13.84	Sylvilagus	1
Sylvilagus bachmani	9	53	238	26944	12.52	Sylvilagus	0
Lepus callotis	11	100	238	26944	10.81	Lepus	1
Lepus alleni	6	67	238	26944	7.06	Lepus	0
Sylvilagus graysoni	1	5	238	26944	4.57	Sylvilagus	0
Lepus flavigularis	1	13	238	26944	2.62	Lepus	0
Sylvilagus insonus	0	3	238	26944	-0.16	Sylvilagus	0
Sylvilagus mansuetus	0	2	238	26944	-0.13	Sylvilagus	0
Sylvilagus gabbi	0	1	238	26944	-0.09	Sylvilagus	0
Sylvilagus brasiliensis	0	61	238	26944	-0.74	Sylvilagus	0

a zoonosis involves several local and direct micro interactions: host-vector, pathogen-host and pathogen-vector, that are related in a complex way. Both the host range (the number and type of hosts) and the vector range (the number and type of vectors) are important factors in the transmission cycle and highly relevant for indicating how to combat a zoonosis. It is prohibitively costly to try and test every possible host and every possible vector to determine if and how it enters in the transmission cycle of a pathogen and even more costly to determine its relative importance. We hypothesise that the micro interactions between pathogen-vector-host are such that each biotic element enters as a potential niche dimension for the others and that will leave an imprint of the interaction at the macro level when considering the relative spatial distributions of vector and host.

Although we have considered multiple zoonoses—leishmaniasis (Berzunza-Cruz et al., 2015; Stephens et al., 2009; Stephens et al., 2016), Chagas disease (Ibarra-Cerdeña et al., 2017; Rengifo-Correa et al., 2017), Zika virus (González-Salazar et al., 2017), Yellow fever, St. Louis encephalitis, Dengue and West Nile virus, we will consider here, as a representative example, only leishmaniasis. In the case of leishmaniasis, the vector-host relation is accepted to be between a hematophagous insect vector—of the genus *Lutzomyia*—and a mammalian host. In Mexico, until recently, the number of confirmed hosts was only nine, of more than 430 possible candidate mammal species (Stephens et al., 2009). Following the logic that a necessary condition for the presence of the pathogen is the presence of the host and that the disease hosts will be favourable niche dimensions for the vectors we created a list, ranked by ε, of all mammals in Mexico (Stephens et al., 2009). The list was then checked against current knowledge in terms of the nine confirmed hosts, all of which corresponded to high values of ε, indicating a strong positive interaction in terms of our empirical definition. As with the example bobcat-prey, the predictive value of this unsupervised model is very high. However, as with the predation example, we must ask whether the set of confirmed hosts is representative and if it is complete, and if so, what is the nature of the false positives or false negatives in the model? In Table 7 we see the 150 most highly ranked (most important) interactions by ε between the genus *Lutzomyia* and potential mammal hosts. The previously confirmed hosts are denoted by "Yes" in the column "Conf." Taking as an example classification criterion that any mammal in the top 5% (corresponding to rank 21 in the list) is predicted to be a host then, if we accept that the *only* mammal hosts are the nine already confirmed hosts, the sensitivity (recall) of the model is 3/21 = 14.3%, which may be compared with the null hypothesis that there is no interaction between the distributions of vectors and hosts, wherein the probability to find confirmed hosts in any group would be 9/419 = 2.1%. Thus, this simple model as a classification model for identifying known disease hosts of leishmaniasis is almost 681% better than that of a random model benchmark. If we take a larger group, the top 20%, then the sensitivity drops off, as

Table 7: List of top 150 most highly ranked potential mammal hosts of *Leishmania*.

	Mammals	Epsilon	Conf.
1	Eira barbara	10.1683	
2	Rhogeessa aeneus	9.3649	
3	Artibeus intermedius	9.1628	
4	Reithrodontomys gracilis	8.8921	Yes
5	Carollia sowelli	8.8303	
6	Heteromys gaumeri	8.8000	Yes
7	Peromyscus mexicanus	8.7859	
8	Heteromys desmarestianu	8.7164	Yes
9	Molossus rufus	8.6277	
10	Glossophaga soricina	8.5713	
11	Carollia perspicillata	8.5030	
12	Orthogeomys hispidus	8.3468	
13	Pteronotus parnellii	8.1632	
14	Desmodus rotundus	8.1519	
15	Dasyprocta mexicana	8.1128	
16	Sturnira lilium	8.0290	
17	Dermanura phaeotis	8.0055	
18	Dasyprocta punctata	7.9678	
19	Oryzomys couesi	7.7253	
20	Potos flavus	7.7246	
21	Conepatus semistriatus	7.6879	
22	Ototylomys phyllotis	7.5587	Yes
23	Ateles geoffroyi	7.4787	
24	Cryptotis magna	7.4207	
25	Cuniculus paca	7.3220	
26	Lampronycteris brachyotis	7.2852	
27	Sigmodon hispidus	7.2805	Yes
28	Peromyscus yucatanicus	7.2486	Yes
29	Oryzomys chapmani	7.1242	
30	Didelphis virginiana	7.1150	
31	Peromyscus melanocarpu	7.0260	
32	Microtus umbrosus	6.9630	
33	Thyroptera tricolor	6.9630	
34	Nasua narica	6.8953	
35	Megadontomys cryophilus	6.6830	
36	Oryzomys alfaroi	6.6816	
37	Sorex veraepacis	6.6797	
38	Carollia subrufa	6.6316	
39	Peromyscus aztecus	6.6173	
40	Didelphis marsupialis	6.4390	Yes
41	Sciurus yucatanensis	6.3865	
42	Philander opossum	6.2546	
43	Habromys ixtlani	6.1120	
44	Microtus waterhousii	6.1120	
45	Pteronotus rubiginosus	6.1120	
46	Reithrodontomys microdor	6.0967	
47	Coendou mexicanus	6.0268	
48	Centurio senex	6.0076	
49	Artibeus jamaicensis	5.9786	
50	Glossophaga morenoi	5.8847	
51	Molossus sinaloae	5.8518	
52	Artibeus lituratus	5.8422	
53	Mormoops megalophylla	5.8374	
54	Habromys lepturus	5.7848	
55	Myotis keaysi	5.6148	
56	Chiroderma villosum	5.5562	
57	Tamandua mexicana	5.4845	
58	Tylomys nudicaudus	5.4510	
59	Saccopteryx bilineata	5.2984	
60	Macrotus mexicanus	5.2472	
61	Sciurus aureogaster	5.2267	
62	Baiomys musculus	5.2092	
63	Rhogeessa tumida	5.1950	
64	Sciurus deppei	5.1414	
65	Dermanura watsoni	5.1338	
66	Otonyctomys hatti	5.1338	
67	Orthogeomys grandis	5.0556	
68	Alouatta palliata	5.0457	
69	Choeroniscus godmani	5.0457	
70	Peropteryx macrotis	5.0457	
71	Pteronotus personatus	5.0266	
72	Lontra longicaudis	4.9330	
73	Reithrodontomys mexicanu	4.9120	
74	Oryzomys rostratus	4.8681	
75	Mimon cozumelae	4.8327	
76	Pteronotus davyi	4.7943	
77	Herpailurus yagouaroundi	4.7100	
78	Glossophaga leachii	4.6849	
79	Rhogeessa gracilis	4.6317	
80	Sylvilagus brasiliensis	4.6317	
81	Hodomys alleni	4.5155	
82	Leopardus wiedii	4.4420	
83	Peromyscus simulatus	4.4195	
84	Sigmodon alleni	4.3707	
85	Bassariscus sumichrasti	4.3110	
86	Oryzomys fulvescens	4.3110	
87	Diphylla ecaudata	4.3013	
88	Oryzomys melanotis	4.2907	Yes
89	Micronycteris microtis	4.2338	
90	Mazama americana	4.2274	
91	Microtus oaxacensis	4.2061	
92	Rheomys thomasi	4.2061	
93	Oryzomys saturatior	4.2061	
94	Myotis elegans	4.2024	
95	Oligoryzomys fulvescens	4.1984	
96	Natalus stramineus	4.0626	
97	Balantiopteryx io	4.0522	
98	Nyctinomops laticaudatus	4.0522	
99	Tlacuatzin canescens	4.0119	
100	Odocoileus virginianus	3.9265	
101	Balantiopteryx plicata	3.8590	
102	Peromyscus leucopus	3.7994	
103	Sturnina ludovici	3.7888	
104	Enchisthenes hartii	3.6929	
105	Vampyrodes caraccioli	3.6929	
106	Eptesicus furinalis	3.6453	
107	Liomys pictus	3.6107	
108	Glossophaga commissaris	3.4861	
109	Lonchorhina aurita	3.4781	
110	Phyllostomus discolor	3.4781	
111	Peromyscus gymnotis	3.4516	
112	Anoura geoffroyi	3.4201	
113	Platyrrhinus helleri	3.3586	
114	Eumops bonariensis	3.3398	
115	Sciurus variegatoides	3.3398	
116	Uroderma bilobatum	3.3373	
117	Lasiurus intermedius	3.2197	
118	Lasiurus ega	3.1739	
119	Peromyscus megalops	3.1410	
120	Eumops glaucinus	3.0564	
121	Urocyon cinereoargenteus	2.9697	
122	Procyon lotor	2.9502	
123	Hylonycteris underwoodi	2.9343	
124	Rhynchonycteris naso	2.8580	
125	Eptesicus brasiliensis	2.8106	
126	Myotis albescens	2.8106	
127	Lophostoma evotis	2.8106	
128	Tapirus bairdii	2.8106	
129	Vampyrum spectrum	2.8106	
130	Marmosa mexicana	2.7731	Yes
131	Peromyscus furvus	2.7731	
132	Myotis velifera	2.5757	
133	Spilogale putorius	2.5411	
134	Microtus mexicanus	2.5268	
135	Dasypus novemcinctus	2.4725	
136	Myotis nigricans	2.4704	
137	Lophostoma brasiliense	2.4407	
138	Diclidurus albus	2.4407	
139	Sciurus niger	2.4407	
140	Leptonycteris curasoae	2.4268	
141	Nyctomys sumichrasti	2.4026	
142	Sigmodon mascotensis	2.3815	
143	Alouatta pigra	2.3374	
144	Peromyscus melanophrys	2.2204	
145	Dermanura tolteca	2.1920	
146	Trachops cirrhosus	2.1663	
147	Bauerus dubiaquercus	2.1612	
148	Spilogale pygmaea	2.1612	
149	Leptonycteris nivalis	2.1402	
150	Sylvilagus floridanus	2.1002	

it should, with a sensitivity of 7/84 = 8.3%. Thus, this model does a good job at predicting previously identified hosts. This analysis presupposes however, that no other mammal is a host beyond those already identified. If we take the list as a prediction model and a geographically systematic and random sampling is made, then from a given number of samples we would expect the highest ranked species to lead to more positives than the lower ranked species. This sampling was, indeed, carried out (Stephens et al., 2016), with the result that 922 individuals from 70 species were collected and tested for the presence of the *Leishmania* pathogen.

Of the 70 species tested, 22 that tested positive were previously unknown hosts of *Leishmania* in Mexico (shown in blue in the Table 7). Our unsupervised model now yields a sensitivity of 12/21 = 57.1%, compared to 2.1% for the random benchmark, and 14.3% if we assume that only previously confirmed species were positive. In the top 20% of the list, the corresponding sensitivity is 26.5%.

One may argue that the percentage, 42.9%, of "false" positives associated with the 5% of highest ranked candidate hosts is very high, but this would be very misleading. Take, for example, the bat species *Molossus rufus* that was collected without presenting any positive individuals. This species is very highly ranked and therefore considered to be an important niche dimension for the *Lutzomyia* genus. Is this to be considered a false positive? To do so we must have a hypothesis about the expected infection rate. The infection rate over the whole sample of 922 individuals was 6.7%. Taking this as the null hypothesis, given that only one individual of this species was collected, the probability that it represented a true negative was only 6.7%, far from a 95% confidence interval. Indeed, of the 70 collected species, none could be discarded as a potential host at this level of confidence. In other words, no species of the top 5% or top 20% can be considered as a false positive, either because it has not been collected and tested or because it has not been collected in sufficient quanti-

ty. Independently of this, from a statistical inference viewpoint, the model works extremely well.

Conclusion

We have here tried to summarise both the essential conceptual and theoretical elements that enter in our formalism for defining, characterising, quantifying and predicting ecological interactions, in the hope that it will stimulate researchers to test the methodology on their own problems of interest and judge for themselves its merits. We believe that the representative use cases we have discussed, along with others in the literature, prove its worth. As many basic elements of the methodology are now available in an online platform—SPECIES—literally thousands of test cases, each with hundreds or thousands of covariates, can be produced, each in a matter of minutes. As a concrete example, the Eltonian noise hypothesis may be validated or rejected for any species that exists in the SNIB (Mexico) or GBIF (North America) databases using any combination of biotic and abiotic variables. One can further compare and contrast the accuracy of any corresponding species distribution model associated with these combinations. This is equivalent to determining which variables are more niche-like—positively correlated with the species of interest—and which are anti-niche-like—negatively correlated with the species of interest.

We can compare and contrast the role of every niche variable from a statistically level playing field by bringing every variable to the same spatial resolution and making every variable binomial. We can compare and contrast abiotic with biotic variables, or use different groups of biotic variables, grouped by any criterion we choose (given the data is to hand), such as by taxonomic label, or phenotype or genotype or, by ecological interaction labels or, indeed, whatever we choose. We can go further and use the methodology to analyse correlations between niche variables and thereby begin to study confounding between different variable types and answer questions about who confounds who? All of this can be done, which is a benefit to species distribution and niche modelling, without ever mentioning the word "interaction".

However, without understanding the role of interaction in ecology, all of this is just building a better mousetrap. The concept of interaction is fundamental to understanding how ecology at the micro level emerges into the macro level and manifests itself in the relative distributions of species or other taxa as a function of position and time. Neutral theory (Hubell, 2001) has provided us with a null hypothesis, based on the first principles of stochasticity, that allows us to have access to how the world would be if all species were similar in their per capita rates of birth and death, and hence provided a benchmark against which to compare empirical patterns in the abundance and diversity of species (e.g. Marquet et al., 2014) and infer the relative importance of niche-related processes. Our approach makes use of a similar philosophy by comparing observed patterns to the appropriate benchmark and then making inferences about the significance of their deviation in order to understand and assess to what extent micro interactions may leave an imprint upon macro distributions of taxa.

The tension between neutral and niche also exists in physics. Some systems, such as helium atoms, have strong (intra-atomic) micro interactions and very weak (interatomic) macro interactions. On the other hand, sodium and chlorine atoms have strong (intra-atomic) micro interactions and also have strong (but less strong) inter-atomic interactions. That's how we get salt. So, is ecology more like helium or more like salt? Clearly, the answer is that some ecological systems are more like helium and some are more like salt. How can we distinguish one from the other? First, by defining, as in many other areas of science, interactions with respect to the effect they have on the constituents of the system that is interacting. In particular, on their positions in time and/or space. This is our path to defining interactions at the macro level when we cannot directly derive the macro interaction from the micro.

Thus, we defined interactions as being present when the spatio-temporal distributions of the things that are interacting is different to that in the absence of the interaction—the null hypothesis. Without data on the macro distributions however, we can go no further. Luckily, large point collection databases, notwithstanding questions about data biases, are a wonderful potential source of information about what is where and when. With this data we can determine the effect of one or any number of (niche) variables on a taxon of interest, be they abiotic or biotic. The problem is that the spatio-temporal distribution of one species is an emergent result of the micro level interactions with all potential niche variables. Thus, there is no chance of isolating the effect of abiotic

variables (the fundamental niche) on a species distribution, as the latter is a result of those variables and all the biotic ones too. This is a problem of all Complex Adaptive Systems, not just ecology. There are just too many variables involved to be able to isolate the effects of one variable by "controlling" for the rest. However, point collection data does allow us to set up a potentially infinite set of hypotheses by considering combinations of niche variables. For instance, we can determine the effect of the presence of a species of prey on the bobcat distribution by "controlling" for the effect of temperature, or precipitation, as was done in the section entitled "Inferring causality and identifying confounders." To link this to micro interactions however, we need labels for the niche variables that are ecologically relevant, and which can be used to interpret the macro interactions in micro terms. Thus, *Sylvilagus floridanus* has an important macro interaction with *Lynx rufus* and therefore is a relevant niche variable in the pragmatic sense that where the rabbit is present so is the bobcat. However, it is our understanding of its role as a prey species of the bobcat that provides the micro property that allows us to understand *why* it is an important niche dimension.

So, a macro interaction may be absent even though there is an underlying micro interaction—the helium scenario. However, there can be no macro interaction if there is no micro interaction. Macro interactions can only emerge from the collective effects of multiple micro interactions. It is this fact that allows us to attempt to infer the existence and nature of micro interactions from the macro data. We gave several concrete examples of this. The base model there (our unsupervised learning ε model) is just based on the logic that things can't interact if they don't co-occur, neither at the micro nor the macro level. This led to very predictive statistical inference models for the considered interaction. By considering the various ecological labels of the involved species we could improve those models by determining which labels are associated with false positives versus false negatives. This can be done by hand—deciding for instance that the coyote cannot be a prey of the bobcat and removing it from the model or, in a more principled way, by using a supervised learning model trained on these labels.

We believe that our methodology, and its implementation, available to all in the SPECIES platform, open up new horizons for a large set of analyses that simply were not possible before. What is more, the methodology is equally applicable to any spatio-temporal data of any resolution and of any data type, including public health data, census data, commercial data etc. The data just needs to be incorporated into the SPECIES platform. Thus, we may ask not just what the niche of a vector of a disease is, but also what is the niche of the disease itself by using geo-referenced cases and their associated labels. Given that it works in predicting new, unknown interactions it can be used to rank candidates for furthermore detailed analysis from large numbers of such candidates.

Acknowledgments

This work has benefited from many fruitful collaborations from the theoretical perspective as well as the experimental one. In particular, we thank Raúl Sierra, Victor Sánchez-Cordero, Angel Rodríguez, Ingeborg Becker, Carlos Ibarra, Laura Rengifo, Marie José Tolsá, Fabiola Nieto, Jesús Sotomayor, Gabriel García, Gerardo Suzán, Benjamin Roche. We are grateful for financial support from DGAPA-PAPIIT grant IG200217 and AFB17008 (CONICYT-Chile).

Conflicts of Interest

The authors declare no conflict of interest.

References

Abramsky, Z., Bowers, M.A. and M.L. Rosenzweig. 1986. Detecting interspecific competition in the field: testing the regression method. *Oikos* 47: 199-204.

Alfonso, J. and D. Vilar. 2007. Bridging the gap between Naive Bayes and maximum entropy. In Proceedings of the PRIS 2007, Funchal (Portugal), pp 59-65

Álvarez-Martínez, J. M., Suárez-Seoane, S., Palacín, C., Sanz, J. and J.C. Alonso. 2015. Can eltonian processes explain species distributions at large scale? A case study with great bustard (*Otis tarda*). *Div. Dis.* 21: 123-138.

Aragón, P. and D. Sánchez-Fernández. 2013. Can we disentangle predator-prey interactions from species distributions at a macro-scale? A case study with a raptor species. *Oikos* 122: 64-72.

Aranda, M., Rosas, O., Ríos, J.D.J. and N. García, N. 2002. Análisis comparativo de la alimentación del gato montés (*Lynx rufus*) en dos diferentes ambientes de México. *Acta Zool. Mex.* 87: 99-109.

Araújo, M.B., Rozenfeld, A., Rahbek, C., and P.A. Marquet. 2011. Using species co-occurrence networks to

assess the impacts of climate change. *Ecography* 34: 897-908.

Araújo, M.B. and A. Rozenfeld. 2014. The geographic scaling of biotic interactions. *Ecography* 37: 406-415.

Araújo, C.B., Marcondes-Machado, L.O. and G.C. Costa. 2014. The importance of biotic interactions in species distribution models: a test of the Eltonian noise hypothesis using parrots. *J. Biogeogr.* 41: 513-523.

Arditi, R., and L.R. Ginzburg. 1989. Coupling in predator-prey dynamics: ratio-dependence. *J. Theor. Biol*. 139: 311-326.

Atauchi, P. J., Peterson, A. T. and J. Flanagan. 2018. Species distribution models for Peruvian Plantcutter improve with consideration of biotic interactions. *J. Avian Biol.* 49: 01617.

Belmaker, J., Zarnetske, P., Tuanmu, M.N., Zonneveld, S., Record, S., Strecker, A. and L. Beaudrot. 2015. Empirical evidence for the scale dependence of biotic interactions. *Global Ecol. Biogeogr*. 24: 750761.

Bethan V.P. and N. Golding. 2015. Tracking the distribution and impacts of diseases with biological records and distribution modelling. *Biol. J. Linn. Soc.* 115: 664-677.

Berzunza-Cruz M, Rodríguez-Moreno A, Gutiérrez-Granados G, González-Salazar C, Stephens C.R., Hidalgo-Mihart M., et al. 2015. Leishmania (L.) mexicana Infected Bats in Mexico: Novel Potential Reservoirs. *PLoS Neglect. Trop. D.* 9: 1-15.

Berger, J.O. 1985. Statistical Decision Theory and Bayesian Analysis. Berlin: Springer-Verlag.

Bradley, B.A., Blumenthal, D.M., Early, R., Grosholz, E.D., Lawler, J.J., Miller, L.P., et al. 2012. Global change, global trade, and the next wave of plant invasions. *Front. Ecol. Environ* 10: 20-28.

Bristow, C.S., Hudson-Edwards, K.A. and A. Chappell. 2010. Fertilizing the Amazon and equatorial Atlantic with West African dust. *Geophys. Res. Lett*. 37(14).

Broos, P.S., Getman, K.V., Povich, M.S., Townsley, L.K., Feigelson, E.D. and G.P. Garmire. 2011. A Naive Bayes Source Classifier for X-ray Sources. *Astrophys. J. Suppl. S.* 194: 4.

Borthagaray, A. I., Arim, M. and P.A. Marquet. 2014. Inferring species roles in metacommunity structure from species cooccurrence networks. *P. Roy. Soc. B-Biol. Sci*. 281: 20141425.

Brown, J.H., Kelt, D.A. and B.J. Fox. 2002. Assembly Rules and Competition in Desert Rodents. *Am. Nat.* 160: 815-818.

Burak, T. and B. Ayse. 2009. Analysis of Naive Bayes assumptions on software fault data: An empirical study. *Data Knowl. Eng.* 68: 278-290.

Case, T.J. 1990. Invasion resistance arises in strongly interacting species-rich model competition communities. *P. Natl. Acad. Sci. USA* 87: 9610-9614.

Cazelles, K., Araújo, M.B., Mouquet, N. and D. Gravel. 2016. A theory for species co-occurrence in interaction networks. *Theor. Ecol.* 9: 39-48.

Chen, S.F. and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. Proceedings of the 34th annual meeting on Association for Computational Linguistics.

Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J., and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecol. Monogr.* 87: 3456

Colwell, R. K. and D.W. Winkler. 1984. A null model for null models in biogeography. Pp. 344-359. In: Strong, D.R., Simberloff D, Abele, L.G., Thistle (eds.) Ecological communities: conceptual issues and the evidence, Princeton University Press, Princeton, NJ.

Connor, E.F. and D. Simberloff. 1979. The assembly of species communities: chance or competition. *Ecology* 60: 1132-1140

Connor, E.F., Collins, M.D. and D. Simberloff. 2013. The checkered history of checkerboard distributions. *Ecology*, 94: 2403-2414.

Crowell, K.L., and S.L. Pimm. 1976. Competition and niche shifts of mice introduced onto islands. *Oikos* 27: 251-258.

Dayton, P.K. 1973. Two cases of resource partitioning in an intertidal community: making the right prediction for the wrong reason. *Am. Nat.* 107: 662-670.

Delibes, M., Zapata, S.C., Blázquez, M.C. and R. Rodríguez-Estrella. 1997. Seasonal food habits of bobcats (*Lynx rufus*) in subtropical Baja California Sur, Mexico. *Can. J. Zool.* 75: 478-483.

Diamond, J.M. 1975. Assembly of species communities. p. 342-444 in: Ecology and Evolution of Communities. M.L. Cody and J.M. Diamond (eds.). Harvard University Press, Cambridge

Elith J. and J.R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. S.* 40: 677-697.

Elton, C. 1927. Animal Ecology. Sidgwick and Jackson, LTD, London, 56.

Freilich, M.A., Wieters, E., Broitman, B.R., Marquet, P.A. and S.A. Navarrete. 2018. Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? *Ecology* 99: 690-699.

García, J.A.M., Martínez, G.D.M., Plata, P.F.X., Rosas, O.C.R., Arámbula, L.A.T. and L.C. Bender. 2014.

Use of prey by sympatric bobcat (*Lynx rufus*) and coyote (*Canis latrans*) in the Izta-Popo National Park, Mexico. *Southwest. Nat.* 59: 167-172.

Gause, G.F. 1934. Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence. *Science* 79: 16-17.

Gehlke, C.E. and K. Biehl. 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *J. Am. Stat. Assoc.* 29: 169-170.

Giannini, T.C., Chapman, D.S., Saraiva, A.M., Alves-dos-Santos, I. and J.C. Biesmeijer. 2013. Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography* 36: 649-656.

Gilpin, M.E. 1975. Limit cycles in competition communities. *Am. Nat*. 109: 51-60.

Gilpin, M.E., and F.J. Ayala. 1973. Global models of growth and competition. *P. Natl. Acad. Sci. USA* 70: 3590-3593.

Godsoe, W. and L.J. Harmon. 2012. How do species interactions affect species distribution models? *Ecography* 35: 811-820.

González-Salazar, C., Stephens, C.R. and P.A. Marquet. 2013. Comparing the relative contributions of biotic and abiotic factors as mediators of species-distributions, Ecol. Model. 248: 57-70.

González-Salazar, C. and C.R. Stephens. 2012. Constructing ecological networks: a tool to infer risk of transmission and dispersal of leishmaniasis. *Zoonoses Public. Hlth*. 59, s2, 179-193.

González-Salazar, C., Stephens, C.R. and V. Sánchez-Cordero. 2017. Predicting the Potential Role of Non-human Hosts in Zika Virus Maintenance. *EcoHealth* 14: 171-177.

Gotelli, N.J. 2000. Null model analysis of species co-occurrence patterns. *Ecology* 81: 2606-2621

Gotelli, N.J. and D.J McCabe,. 2002. Species cooccurrence a meta analysis of J M Diamonds assembly rules model. *Ecology* 83: 2091-2096.

Gotelli, N.J., Graves, G.R. and C. Rahbek. 2010. Macroecological signals of species interactions in the Danish avifauna. *P. Natl. Acad. Sci. USA* 107: 5030.

Hall, E.R. 1981. The Mammals of North America, Wiley, New York.

Haskell, E.F. 1949. A clarification of social science. Main Currents in Modern Thought 7, 45-51.

Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R.G. and J.H. Korber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecol. Biogeogr*. 16: 754-763.

Hill, A. B. 1965. The environment and disease: association or causation? *J. Roy. Soc. Med.* 108: 32-37

Holling, C.S. 1959. Some characteristics of simple types of predation and parasitism. *Can. Entomol*. 91: 385-398.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* 117: 847-858

Hudson, R., Rodríguez-Martínez, L., Distel, H., Cordero, C., Altbacker, V. and Martínez-Gómez. 2005. A comparison between vegetation and diet records from the wet and dry season in the cottontail rabbit *Sylvilagus floridanus* at Ixtacuixtla, central Mexico. *Acta Theriol.* 50: 377-389

Hubbell, S.P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology. 22: 415-427.

Ibarra-Cerdeña, C.N., Valiente-Banuet, L., Sánchez-Cordero, V., Stephens, C.R. and J.M. Ramsey. 2017. *Trypanosoma cruzi* reservoir-triatomine vector co-occurrence networks reveal meta-community effects by synanthropic mammals on geographic dispersal. *PeerJ* 5, e3152.

Lidicker, W.Z. 1979. A clarification of Interactions in Ecological Systems. *BioScience* 29: 475-477.

MacArthur, R., and R. Levins. 1967. The limiting similarity, convergence, and divergence of coexisting species. *Am. Nat*. 101: 377-385.

Major, J.T. and J.A. Sherburne. 1987. Interspecific relationships of Coyotes, Bobcats, and Red Foxes in western Maine. *J. Wildlife Manage*. 51: 606-616.

Marquet, P.A., Allen, A.P., Brown, J.H., Dunne, J.A., Enquist, B.J., Gillooly, J.F., Gowaty, P.A. Green, J.L., Harte, J., Hubbell, S.P., et al. 2014. On theory in ecology. *BioScience* 64: 701-710.

Mohd, M.H., Murray, R., Plank, M.J., and W. Godsoe. 2017. Effects of biotic interactions and dispersal on the presence-absence of multiple species. *Chaos Soliton Fract*. 99: 185-194.

Morales-Castilla, I., Matias, M.G., Gravel, D. and M.B. Araújo. 2015. Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30: 347-356.

Openshaw, S., 1983. The modifiable areal unit problem. Concepts and techniques in modern geography. Norfolk, UK: Geo Books.

Ovaskainen, O., Hottola, J. and J. Shtonen. 2010. Modeling species co-occurrence by multivariate logistic

regression generates new hypotheses on fungal interactions. *Ecology* 91: 2514-2521.

Paine, R.T. (1992). Food-web analysis through field measurement of per capita interaction strength. *Nature* 355: 73.

Peterson, A.T. Soberón, J., Pearson, R.G. Robert P.A., Martínez-Meyer, E., Nakamura, M. and M.B. Araújo. 2011. Ecological Niches and Geographic Distributions, Princeton University Press. (MPB-49) (Monographs in Population Biology) Princeton University Press

Peterson, A. T., Cobos, M. E. and D. Jiménez-García. 2018. Major challenges for correlational ecological niche model projections to future climate conditions. Ann. NY. Acad. Sci. 1429: 66-77.

Pearl, J. 2000. Causality, Cambridge University Press.

Phillips, P.C. 2008. Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9: 855-867.

Phillips, S.J., Dudík, M. and R.E. Schapire. 2004. A maximum entropy approach to species distribution modeling. In Proceedings of the twentyfirst international conference on Machine learning, pages 655-662.

Phillips, S.J., Anderson., R.P. and R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 190: 231-259.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O´Hara, R.B., Parris K.M. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.* 5: 397-406.

Qiao, H., Soberón, J. and A.T. Peterson. 2015. No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods Ecol. Evol.* 6: 1126-1136.

Rebolledo, R., Navarrete, S.A., Kofi, S., Rojas, S., and P.A. Marquet. 2019. An Open-System Approach to Complex Biological Networks. *SIAM J. App. Math.* 79: 619-640.

Rengifo-Correa, L., Stephens, C.R; Morrone, J.J., Téllez-Rendon, J.L. and C. González-Salazar. 2017. Understanding transmissibility patterns of Chagas disease through complex vector-host networks. *Parasitology* 144: 760-772.

Roberts, A. and L. Stone. 1990. Island-sharing by archipelago species. *Oecologia* 83: 560-567.

Rosenzweig, M. L., Abramsky, Z., and B. Kotler. 1985. Can interaction coefficients be determined from census data? *Oecologia*, 66: 194-198.

Rosenbaum P.R. and D.B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Royan, A., Reynolds, S.J., Hannah, D.M., Prudhomme, C., Noble, D.G. and J.P. Sadler. 2016. Shared environmental responses drive cooccurrence patterns in river bird communities. *Ecography* 39: 733-742.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66: 688-701.

Rubin, D.B. 1978. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* 6: 34-58.

Sánchez-Cordero, V., Stockwell, D., Sarkar, S., Liu, H., Stephens, C.R. and J. Giménez. 2008. Competitive interactions between felid species may limit the southern distribution of bobcats *Lynx rufus*. *Ecography* 31: 757-764.

Schoener, T.W. 1974. Competition and the form of habitat shift. *Theor. Popul. Biol.* 6: 265-307.

Sierra, R. and C.R. Stephens. 2012. Exploratory analysis of the interrelations between co-located boolean spatial features using network graphs. *Int. J. Geogr. Inf. Sci.* 26: 444-468.

Skinner, S. 1990. Earthmover. *Wyoming Wildlife*. 54: 4-9.

Snow, J. 1855. On the Mode of Communication of Cholera. London: John Churchill.

Soberón J. and A.T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species distributional areas. *Biodiversity Informatics* 2: 1-10.

Soberón, J. and M. Nakamura. 2009. Niches and distributional areas: Concepts, methods, and assumptions. *P. Natl. Acad. Sci. USA* 106: 19644-19650.

Soberón J. and A.T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philos. T. Roy. Soc. B*. 359: 689-698.

Stephens C.R., Heau J.G., González, C., Ibarra-Cerdeña C.N., Sánchez-Cordero V. and C. González-Salazar. 2009. Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS One* 4, e5725.

Stephens, C. R., Sierra-Alcocer, R., González-Salazar, C., Barrios, J. M., Salazar Carrillo, J.C., Robredo E.E., and E. del Callejo. 2019. SPECIES: A platform for the exploration of ecological data. *Ecol. Evol*. 9: 1638-1653.

Stephens, C.R., González-Salazar, C., Sánchez-Cordero, V., Becker, I., Rebollar-Tellez, E., Rodríguez-Moreno, A., Berzunza-Cruz, M., Balcells, C. D., Gutiérrez-Granados, G. and M. Hidalgo-Mihart. 2016. Can you judge a disease host by the company it keeps?

Predicting disease hosts and their relative importance: a case study for Leishmaniasis. *PLoS Neglect. Trop. D.* 10: e0005004.

Stephens, C.R., Sánchez-Cordero, V. and C. González-Salazar. 2017a. Bayesian inference of ecological interactions from spatial data. *Entropy* 19: 547.

Stephens, C.R, Flores, H.H, and A. Ruiz Linares. 2017b. When is the Naive Bayes approximation not so naive? *Mach. Learn.* 107: 397-441

Stockwell, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inf. Sci*. 13: 143-158.

Wang, Q., Garrity, G.M., Tiedje, J.M. and J.R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microb*. 73: 5261-5267.

Wang, B., Wu, R., and X. Fu. 2000. Pacific East Asian teleconnection: how does ENSO affect East Asian climate? *J. Climate* 13: 1517-1536.

Wei, W., Visweswaran, S. and G.F. Cooper. 2011. The application of Naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J. Am. Med. Inform. Assn*. 18: 370-375.

Wilson, E.B. 1927. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc*. 22: 209-212.

Wilson, W.G., Lundberg, P., Vazquez, D.P., Shurin, J.B., Smith, M.D., Langford, W., Gross, K.L. and G.G. Mittelbach. 2003. Biodiversity and species interactions: extending Lotka-Volterra community theory. *Ecol. Lett.* 6: 944-952.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Gryntes, J.A., Guisan,A., Heikkinen, R.K., Hoye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* 88: 15-30.

Wootton, J.T. and M. Emmerson. 2005. Measurement of interaction strength in nature. *Annu. Rev. Ecol. Evol. S.* 36: 419-444.