Tropical Medicine and Infectious Disease

MDPI

Article

"Does a Respiratory Virus Have an Ecological Niche, and If So, Can It Be Mapped?" Yes and Yes

Christopher R. Stephens [1,2,*], Constantino González-Salazar [1,3] and Pedro Romero-Martínez [1]

[1] C3—Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad de Mexico 04510, Mexico
[2] Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Ciudad de Mexico 04510, Mexico
[3] Instituto de Ciencias de la Atmósfera y Cambio Climático, Universidad Nacional Autónoma de México, Ciudad de Mexico 04510, Mexico
* Correspondence: stephens@nucleares.unam.mx

Abstract: Although the utility of Ecological Niche Models (ENM) and Species Distribution Models (SDM) has been demonstrated in many ecological applications, their suitability for modelling epidemics or pandemics, such as SARS-Cov-2, has been questioned. In this paper, contrary to this viewpoint, we show that ENMs and SDMs can be created that can describe the evolution of pandemics, both in space and time. As an illustrative use case, we create models for predicting confirmed cases of COVID-19, viewed as our target "species", in Mexico through 2020 and 2021, showing that the models are predictive in both space and time. In order to achieve this, we extend a recently developed Bayesian framework for niche modelling, to include: (i) dynamic, non-equilibrium "species" distributions; (ii) a wider set of habitat variables, including behavioural, socio-economic and socio-demographic variables, as well as standard climatic variables; (iii) distinct models and associated niches for different species characteristics, showing how the niche, as deduced through presence-absence data, can differ from that deduced from abundance data. We show that the niche associated with those places with the highest abundance of cases has been highly conserved throughout the pandemic, while the inferred niche associated with presence of cases has been changing. Finally, we show how causal chains can be inferred and confounding identified by showing that behavioural and social factors are much more predictive than climate and that, further, the latter is confounded by the former.

Keywords: ecology; epidemiology; SARS-Cov-2; COVID-19; ecological niche model; species distribution model; Bayesian analysis; causal inference

Citation: Stephens, C.R.; González-Salazar, C.; Romero-Martínez, P. "Does a Respiratory Virus Have an Ecological Niche, and If So, Can It Be Mapped?" Yes and Yes. *Trop. Med. Infect. Dis.* **2023**, *8*, 178. https://doi.org/10.3390/tropicalmed8030178

Academic Editor: John Frean

Received: 1 December 2022
Revised: 12 March 2023
Accepted: 13 March 2023
Published: 17 March 2023

1. Introduction

Recently there has been debate [1–6] as to whether Species Distribution Models (SDM) are appropriate tools in the study of the COVID-19 pandemic. Of course, such a debate raises the important question of when and under what circumstances an SDM, or an Ecological Niche Model (ENM), are likely to be valid and/or useful in the study of disease in general? Just which diseases, or aspects of diseases, can be usefully studied using SDM/ENMs? Which pathogens have ecological niches and which do not? In addition, if some do not, why do they not? Taking the broad view: If ecology is the study of the relations between organisms, both among themselves and with their environment, and an ecological niche is the full set of biotic and abiotic factors that favour the presence of an organism [7,8], then clearly all disease pathogens must have an ecological niche. The question, rather, is: how is that niche to be characterised and quantified? What are the appropriate niche dimensions and can a meaningful and useful ENM and, consequently an SDM, be constructed?

Although there is controversy as to the application of ENMs and SDMs to the case of COVID-19, there have been varying degrees of success in applying them to other human diseases. In particular, to zoonoses [9,10], where non-human species are involved in the transmission cycle. The principal applications have been to the creation of ENMs and SDMs for important agents in the transmission cycle of the zoonosis, such as known vectors [11] and hosts [12], considered as risk factors, as well as the pathogen itself [13]. In all these cases the corresponding niche variables were abiotic—climate and environmental layers—in spite of the fact that biotic factors clearly play an important role as niche variables for the disease itself. Biotic factors were successfully incorporated into ENMs and SDMs for zoonoses in [14–20] Importantly, the capacity to include biotic factors into ENMs and SDMs led not only to the creation of more predictive and explainable models but also, for example, to the prediction and confirmation of multiple new hosts of several zoonoses [16]. However, all these models, independently of the inclusion of biotic factors, were based on presence–absence data and on the implicit assumption that the system was in "equilibrium" [21].

In all of these cases of human disease there has been a component that is considered "ecological", either in the targets of the ENM/SDM, such as vectors or hosts of a zoonosis, as well as in the niche variables themselves, such as climatic and other environmental variables, as well as potential hosts and vectors. However, in the case of a disease such as COVID-19, which is based on human-to-human transmission, there is a question as to the relevance of such factors. Additionally, abundance, rather than just presence–absence is a fundamental concern. Finally, it is clear that for epidemics or pandemics of this nature in no way is an equilibrium assumption valid. It is these criticisms, among others, that have led to the questioning of an ecological ENM/SDM approach and to the conclusion that an epidemiological rather than ecological approach is to be preferred, which leads us to a further question: when is an ecological versus an epidemiological approach more appropriate? The difference between the two is stated quite clearly in the epidemiological literature [22,23]. Put simply, ecological models are naturally associated with the "where" of a disease, while epidemiological models are more concerned with the "who". This difference is aptly captured in [24] with the question: "Is it better to have a heart attack in the United States or Canada?", the emphasis being that this type of question is of an ecological nature, whereas the question: How many people out of a cohort of individuals diagnosed with COVID-19 will subsequently die? is an epidemiological question. Indeed, this ecological "where" perspective has a long history in the social sciences as a whole [25,26]. From a modelling viewpoint, the difference between the two approaches lies in what statistical ensemble will be used to draw inferences. Epidemiological models use an ensemble of "individuals", while ecological models use an ensemble of "places", where each place is associated, either explicitly or implicitly, with a population. The "places" can vary depending on context: from countries and other political units, to eco-regions, any fixed-area spatial cells, and to pixels on a raster.

Additionally, there is a subsequent question of "when", that is relevant for both the ecological and epidemiological points of view. Standard SIR-type modelling [27], for instance, is associated with the "who" and the "when", with the "who" being associated with a fixed-number of states—susceptible, infected, etc. that are associated with individuals [28]. It is the "when" and, relatedly, "how many" that has dominated the modelling of the COVID-19 pandemic, with standard SIR models, and variations thereof [29–31], playing an important role [32,33]. However, more sophisticated techniques, such as deep learning, have also been used [34]. All these methods however, are based on modelling the time series of the data to be predicted, such as cases or deaths. In no way can they account for the high degree of multi-factoriality involved in the evolution in space and time of the pandemic. In other words, they cannot account for the "whys" that accrue from the direct and indirect causes of this evolution. Thus, although such dynamical models can be extended to consider "place" by constructing a SIR-type model for a particular place [35,36]; such models, however, do not account for the distinguishing features of that particular "place", as does an SDM/ENM. On the other hand, currently, SDM/ENMs do

not account for the dynamics of a disease and are also unnecessarily restrictive in the set of niche variables considered. Finally, there is, behind the "who", "where", and "when", the "why" that explains them. This, we argue, is the most important role for the use of ENMs in the study of human disease.

In this paper, we show that an ecological approach, using SDM/ENMs, can be usefully applied to any human or non-human disease, transmissible or not, taking as a specific example the spatio-temporal distribution of COVID-19 in Mexico to answer in the affirmative that a respiratory virus does, indeed, have an ecological niche, and that it can be mapped. We show that a wide variety of habitat variables, both environmental, behavioural, and social, and of different types and spatio-temporal resolutions, can be included, to yield a deeper understanding of the factors that drive the COVID-19 pandemic. Moreover, we show how to generalise SDM/ENMs to incorporate and predict the dynamics of the disease. Finally, we show how the formalism can be used to disentangle the complex causal chains that are a fundamental part of a complex adaptive system.

2. Materials and Methods

Several of the methodological elements used in this paper have been used previously for creating ENM/SDMs in multiple contexts [15–18]. Further details can be found in these papers and in the Supplementary Material.

2.1. Defining a Spatial Grid

All SDM/ENMs are based on the notion of co-occurrence between a target, C, and, one or more, predictors/habitat variables, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$. Normally, co-occurrence is considered purely in spatial terms, although the concept can be extended to co-occurrences in time. In either case, to specify whether there is a co-occurrence or not requires a specification of a grid that divides a spatio-temporal region into cells. A spatial grid can consist of cells of arbitrary shape, as long as they form a partition; i.e., each spatial point is a member of one and only one cell. A partition may be uniform, such as formed by rectangular cells of a given area, or irregular, as is the case for political/administrative units, such as municipalities, counties, states, etc. In standard SDM/ENM modelling [37] this partition is implicit, corresponding, for example, to the pixels of environmental rasters. Such a partition, however, creates a barrier when wishing to include biotic factors, such as point collection data, that cannot be naturally represented as a raster. In short, we may wish to ask: what is the relative importance of average annual temperature, as taken from WorldClim, versus the presence of a prey species in an SDM/ENM for a vagile carnivore?

In Stephens et al. [38], a methodology has been developed for incorporating spatial data layers of arbitrary type and spatial resolution. A spatial grid is overlaid on a chosen spatial region and co-occurrences defined with respect to a given cell. Thus, if there is a co-occurrence between C and a particular variable, X_i, in a particular cell, then $N(CX_i) = 1$. In the case of a continuous variable, such as species abundance, temperature, or precipitation, the variable, X_i, is coarse grained into a set of n discrete bins, leading to n discrete "presence/absence" variables. Thus, $X_i^m(\alpha) = 0, 1$ represents the presence/absence in a cell, α, of values of X_i in the range defined by the mth bin. Any categorical variable can be left as is, or also coarse grained into a smaller number of categories, if necessary. The criterion for fixing the bin distribution of a variable are that it allows for the best discrimination—dependence of C on X_i—while maintaining an appropriate degree of statistical significance—number of cells associated with a given X_i^m. Furthermore, the target C can also be discretised if necessary in the same way, with bins C_i^m. In summary, all variables become categorical, with each category being associated with a binary variable. An advantage of this categorisation is that no relation is assumed between one bin and another, as would be the case in a regression-based approach for example. Using a non-uniform grid however, can introduce some bias, as some municipalities are bigger than others. We have, however, used spatially uniform grids without substantial changes in our results.

2.2. Constructing SDMs and ENMs

Having transformed all variables to this binomial form, counts, $N(CX_i^m)$, can be made over the region of interest, corresponding to the number of cells that contain a presence of the target C_i^m and the variable X_i^m. The number of cells associated with a given C_i^m and/or X_i^m may be fixed or varying depending on the target class. For instance, if we define the class using a relative measure, such as the 10% of cells with the highest abundance for a given species, or the highest number of confirmed cases or deaths in the case of COVID-19, then the class will always be associated with 10% of the cells, independently of time. However, if the class is based on an absolute measure, such as presence/absence of the species, or confirmed cases of a disease, the number of cells in the class may vary in time as, for example, in the case of an invasive species, or an emerging epidemic. Similar considerations hold for $X_i^m(t)$. If it represents a relative measure, such as the 10% of cells with highest average annual temperature, then it will always cover 10% of cells. However, if it represents the presence/absence of an invasive species, the number of associated cells will change. We will consider these cases in more detail below.

From the co-occurrence counts, probability distributions may be calculated, such as $P(CX_i^m)$, $P(C \mid X_i^m)$, and $P(X_i^m \mid C)$, which are related through Bayes theorem. These distributions can be compared to a null hypothesis and a binomial test used, for instance, to determine the statistical significance of the deviation from this hypothesis. For example, for the posterior, a natural null hypothesis is $P(C)$ (This is equivalent to the null hypothesis of type SIM2 in the classification of Gotelli [39]. This null hypothesis is one that leads to lower rates of Type I errors and corresponds, in the framework of presence–absence matrices, to keeping the number of observations fixed but randomising their location.) and a statistical diagnostic for determining if the habitat variable X_i^m is significant or not is

$$\varepsilon(C \mid X_i^m) = \frac{N_{X_i^m}(P(C \mid X_i^m) - P(C))}{\sqrt{(N_{X_i^m} P(C)(1 - P(C)))}} \tag{1}$$

In the case that the binomial distribution can be approximated by a normal distribution, $\mid \varepsilon(C|X_i^m) \mid > 1.96$ is equivalent to the 95% confidence interval. For a multivariate niche the corresponding distributions of interest are: $P(C\mathbf{X})$, $P(C|\mathbf{X})$, and $P(\mathbf{X}|C)$. Although these exist formally from a frequentist perspective, i.e., $P(C|\mathbf{X}) = N(C\mathbf{X})/N(\mathbf{X})$, in the case of a high-dimensional habitat, both $N(C\mathbf{X})$ and $N(\mathbf{X}) = 0, 1$, which means that a direct statistical estimation is impossible. To overcome this problem, in [40], the likelihood $P(\mathbf{X}|C)$ has been estimated by assuming a factorisation of the form

$$P(\mathbf{X} \mid C) = \prod_{i=1}^{N_\xi} P(\xi \mid C) \tag{2}$$

where ξ is a combination of a small number of variables. Thus, the abiotic and biotic habitat variables are partitioned into a set of N_ξ non-overlapping combinations. In the case that $N_\xi = N$, this corresponds to the well known Naive Bayes approximation [41] (Although a complete factorisation of the likelihood may seem a strong assumption, the Naive Bayes approximation has been shown to be a robust performer even in cases where there are strong correlations between variables. An explanation for its surprisingly good performance can be found in [41]). Usually, in order to calculate $P(C|\mathbf{X})$, as the evidence function $P(\mathbf{X})$ is independent of C, to omit it, the following "score" function is often used

$$\begin{aligned} S(C \mid \mathbf{X}) &= \ln\left(\frac{P(C \mid \mathbf{X})}{P(\bar{C} \mid \mathbf{X})}\right) = \ln\left(\frac{P(\mathbf{X} \mid C)}{P(\mathbf{X} \mid \bar{C})}\right) + \ln\left(\frac{P(C)}{P(\bar{C})}\right) \\ &= \sum_{i=1}^{m} s(X_i^m) + \ln\left(\frac{P(C)}{P(\bar{C})}\right) \end{aligned} \tag{3}$$

where $\bar{C}$ is the set complement of C, with $P(\bar{C}) = 1 - P(C)$, and $s(X_i^m) = \ln(\frac{P(X_i^m|C)}{P(X_i^m|\bar{C})})$ is the contribution ("score") to the overall $S(C|\mathbf{X})$ from the variable X_i^m. If $s(X_i^m) > 0, < 0$ then the factor X_i^m contributes positively/negatively to the occurrence of C. Everything now is based on simple cell counts: $P(X_i^m|C) = N(CX_i^m)/N(C)$, $P(C) = N(C)/N_s$, $P(X_i^m) = N(X_i^m)/N_s$, where $N(C)$ is the number of cells with a presence of the target C, $N(X_i^m)$ is the number of cells with a presence of the variable X_i^m and N_s is the total number of cells in the spatial grid. $P(C|\mathbf{X})$ can be determined directly from $S(C|\mathbf{X})$ by deriving the relation $P(C|S(C|\mathbf{X})) = N(CS)/N(S)$, discretizing the score range into bins and then counting how many cells in a given score range also have a presence of the target. $\Delta(C, \mathbf{X}) = (P(C|\mathbf{X}) - P(C))$ is a measure of how niche-like the conditions specified by $\mathbf{X}$ are. The most niche-like conditions, $\mathbf{X}_n$, are those where $P(C|\mathbf{X})$ reaches its maximum value, while we can term those conditions, $\mathbf{X}_{an}$, where $P(C|\mathbf{X})$ reaches its minimum value, as the most "anti-niche" like. $P(C|\mathbf{X})$ represents a height function for a given point in an N-dimensional Hutchinsonian ecological niche space. The corresponding "Niche Landscape" is our ENM. As every spatial cell α can be assigned a corresponding niche profile, $\mathbf{X}(\alpha)$, $P(C|\mathbf{X}(\alpha))$ for different spatial cells also now yields an SDM over our spatial region of interest.

A significant advantage of assuming a factorisation of the likelihood is that the subsequent model is completely transparent, with each factor contributing separately, so that each factor can be compared and contrasted with the rest. Moreover, the same applies for groups of factors, so we can compare the relative predictive and discriminative value of climate factors versus socio-economic factors, or air contamination versus mobility.

Of fundamental importance in the construction of $P(C|\mathbf{X})$ is the statistical ensemble from which the counts will be made. Although we consider a spatial ensemble, the data assigned to each cell invariably have a temporal dimension. In the case of standard SDM/ENM, with a target class defined through point collection data, each data point, $d(\mathbf{x}, t)$, is associated with a spatial specification, usually latitude and longitude, and a collection date, t. Similarly, abiotic factors are also time-stamped. A model for $P(C|\mathbf{X})$ assumes that the distributions in time of both C and $\mathbf{X}$ are both statistically constant and representative.

In the case that the target variable is metric, such as number of cases, N_C, the classifier $S(C|\mathbf{X})$ can also be used to predict $N_C(\alpha, t)$ for a given spatial cell and over a particular time period. In the case of a spatial model, the cells may be divided up into training and test sets. The points $(N_C(\alpha), S(C \mid \mathbf{X}(\alpha)))$ for each cell, α, of the training set can then be plotted, and a regression performed to determine the relationship $N_C = F(S(C|\mathbf{X}))$. This relation can now be used to predict N_C for any cell in the test set given we have $S(C \mid \mathbf{X})$ there.

2.3. Dynamical ENMs and SDMs

A key aspect of the COVID-19 pandemic is its highly dynamic nature, which can manifest itself in several different ways. First of all, the habitat variables themselves may be time dependent, $X_i \equiv X_i(t)$. Secondly, the target itself may be dynamical, $C \equiv C(t)$. In the latter case, this may represent the fact that a disease or a species/disease is present at a given spatial point at one point in time but not another. Thus, the case of an invasive species would fit into this category. In the former, the disappearance of food resources or climate change would naturally fit. In general, we wish to model $P(C(t)|\mathbf{X}(t'))$. It is important to point out that there is a difference, however, between having a ENM that is dynamical versus just considering a different configuration of the habitat variables substituted into a given model. In the case of climate change, for instance, this is usually performed [42] by determining a static ENM, equivalent to $P(C|\mathbf{X}(t))$, which can be applied to any spatial point, α, then using a climate change model to determine $\mathbf{X}(\alpha, t) \to X(\alpha, t')$ for some future time t'. In other words, we assume the ENM does not change, only the spatial distribution of the habitat variables, which is then used to determine the new spatial distribution of the target with the same original model. In other words, we determine an SDM at t' using an ENM derived at time t.

To move beyond this limited context, we must return to the question of spatio-temporal cells as opposed to just a spatial grid. Considering a timespan T, we divide T into N_T intervals. We now have a total of $N_{Ts} = N_s \times N_T$ spatio-temporal cells. The simplest division is into uniform intervals, such as a year into 12 months. In a static setting, the interpretation of $P(C|\mathbf{X})$ is that it represents an equilibrium relation between C and $\mathbf{X}$, even though the data used to calculate it, more often than not, spans substantial and, effectively, non-commensurate time periods. For example, using WorldClim data from a given year as habitat variables for a niche model of a species represented by collection data taken from a 150-year time period. Without assuming equilibrium we should calculate $P(C(t)|\mathbf{X}(t))$. However, we may also consider calculating $P(C(t)|\mathbf{X}(t'))$, i.e., to predict the effect of the habitat variables $\mathbf{X}(t')$ at time $t' < t$ on our target $C(t)$. Of course, an important question at the heart of this is: how are changes in $\mathbf{X}(t')$ reflected in $C(t)$? This requires longitudinal observations and/or an understanding of the underlying causal relationships between the X_i and C.

There are four distinct paths to incorporating time into the SDM/ENM: (i) assume equilibrium, and thereby ignore time dependence; (ii) construct a model $P(C(t)|\mathbf{X}(t))$ using a time slice/history to predict a spatial distribution $C(\alpha, t)$ on that time slice, as a function of the habitat variables on the same time slice; (iii) predict in time, assuming niche conservatism, i.e., construct $P(C(t)|\mathbf{X}(t))$ and use it to predict the spatial distribution, $C(\alpha, t')$ at some later time using the habitat $\mathbf{X}(\alpha, t')$; (iv) predict in time, without assuming niche conservatism, by constructing $P(C(t)|\mathbf{X}(t'))$, where $t' < t$. Models of type (i) and (ii), we term *spatial prediction* models. They are analogous to standard SDM/ENMs, in that they predict a spatial distribution. However, in distinction, in case (ii) they do so using a time slice of data that permits us to compare and contrast $C(t)$, $\mathbf{X}(t)$ and $P(C(t)|\mathbf{X}(t))$ over time. For instance, we may use data only from May 2021 and compare with data from May 2020, or we may consider data from all of 2021, etc. Models of type (iii) and (iv) we term *time prediction* models. In this case they are an ENM/SDM equivalent to a SIR-type model, predicting the evolution of "where" in time, as opposed to "who". The case with niche conservatism would be equivalent to a SIR type model, where the parameters of the model do not change, whereas the non-conservative case would correspond to the case where they do change and are fitted to dynamic data.

The notion of a *time prediction* model also naturally leads us to consider ENM/SDMs that are associated with changes in the distribution of a species/disease. For instance, we may consider the set of cells, $C(t-1)$, that have a presence in the time interval $t-1$ and the set of cells, $C(t)$, that have a presence in the time interval t. $\Delta_C(t, t-1)$ may then represent those cells that had a presence/absence at $t-1$ and, in contrast, an absence/presence at t. This could, for example, model the range expansion of a species. In this case, an ENM $P(\Delta_C(t, t-1)|\mathbf{X}(t-1))$ represents a model for predicting changes in the distribution due to the habitat variables. This model can then be applied to produce an SDM $P(\Delta_C(\alpha, t+1, t)|\mathbf{X}(\alpha, t))$ using the habitat variables at t and predicting the change in distribution between t and $t+1$.

We give more details about how to construct the different spatio-temporal models in the Supplementary Material. As discussed above, spatially, we used a grid corresponding to the municipalities of Mexico. For the temporal partition, we chose a month, though any other timescale of interest could have been used.

2.4. Testing Model Performance

The way in which our models are tested depends on the type of statistical ensemble that is used to train and then test the model. In the case of an ENM on a given time slice, $P(C(t)|\mathbf{X}(t)$, the model will be created on a training set that corresponds to a randomly chosen fraction, f_{train}, of spatial cells with data associated with a time period t. The model is then tested on the remaining fraction of cells $f_{test} = 1 - f_{train}$ from the same time period. The test and train sets can be chosen in multiple ways. Here, we consider a 70%/30% split. Standard performance statistics can then be determined, such as from a confusion matrix

or from the area under the ROC, among others [43,44]. However, we can also use the ENM $P(C(t)|\mathbf{X}(t))$ trained on 100% of spatial cells and apply it to a future time period t'. In this case, the entire set of spatial cells at t' is the test set. The luxury of a problem such as COVID-19 is that we have relatively accurate data about the spatial distribution of the disease as a function of time. In this case, the ENM $P(C(t)|\mathbf{X}(t))$ is used by substituting at t', $\mathbf{X}(\alpha, t')$ for each spatial cell, α, to obtain for that cell $P(C(\alpha, t')|\mathbf{X}(\alpha, t))$. We can now test the quality of the prediction using any of the above standard metrics, given that $P(C(\alpha, t)|\mathbf{X}(\alpha, t'))$ is a classifier, and we can compare with the actual distribution $C(\alpha, t')$. We will expect the time t ENM to yield good performance at t' only if the niche is conserved over this time period.

Aside from the pure assumption of niche conservatism, where we apply the ENM $P(C(t)|\mathbf{X}(t))$, we can compare and contrast ENMs from different time slices to determine the degree to which they are changing. This would correspond to determining if the niche is actually changing over time. This can be performed for each habitat variable by comparing $s(X_i^m(t))$ with $s(X_i^m(t'))$. If the niche is changing then we must consider just how fast it is changing. This can be deduced by comparing the $s(X_i^m(t))$ across time. If $s(X_i^m(t)) \sim s(X_i^m(t'))$ for two time periods t and $t' > t$, then the niche is conserved over the interval $(t' - t)$. If they differ substantially, then we may reduce $(t' - t)$ and determine how much change there has been over this shorter interval.

For time prediction models that do not assume niche conservatism, the training set consists of choosing two time periods $t - 1$ and t. An ENM $P(C(t)|\mathbf{X}(t-1))$ is created on all spatial cells. This model is then applied to all spatial cells for the time periods t and $t + 1$ as test set. Thus, we use the ENM $P(C(t)|\mathbf{X}(t-1))$, substituting $\mathbf{X}(\alpha, t)$ to create the SDM $P(C(\alpha, t+1)|\mathbf{X}(\alpha, t))$ for prediction of $C(\alpha, t+1)$. We can then apply any of the above standard metrics to evaluate the performance of this ENM/SDM.

2.5. Predicting Abundances

ENMs that are based on presence/no presence are, naturally, binary classification models. Although a metric variable, such as abundance, can be treated as such by considering multiple classes, it is also possible to use a binary prediction model, such as the top 10% of municipalities with highest confirmed cases, to construct a relation, $N(S(C|\mathbf{X}))$ between the score, $S(C|\mathbf{X})$, and the number of cases on a training set. This model can then be applied to a test set, predicting the expected number of cases for a given cell, α, using its habitat profile $\mathbf{X}(\alpha)$. This procedure can be used for both *spatial* and *time* prediction models. In the former case, the abundance predictions are associated with a given time slice t using a split of the cells into training and test sets, while for the latter we predict from a training set that consists of all cells on a time slice t to predict abundances on a test set of all cells on a time slice t'.

2.6. Inferring Causality in ENMs

Another criticism of the application of standard SDM/ENMs to the pandemia has been the lack of plausibility of the relation between, say, infection rates and habitat variables, such as climate or contamination. As pointed out in [40], this is a problem with correlative approaches in general. This criticism however, can be applied to a phenomenological study of any complex adaptive system, where cause–effect relations are multi-layered and, therefore, often indirect. Epidemiology and medicine in general are rife with problems of causal inference and there are two basic approaches to it: a "classic" approach [45] and a "modern" approach [46–48]. An important criterion from our perspective is that of "strength of association" [45], where, although a small association does not mean that there is no causal effect, the larger the association, the more likely that it is causal. This is key to our use and understanding of both $\varepsilon(C|X_i^m)$ and $s(X_i^m)$ and their multi-factorial counterparts $\varepsilon(C|\mathbf{X})$ and $s(\mathbf{X})$. Thus, for two niche factors, X_α and X_β, if $s(X_i^m) > s(X_j^n)$, then we will judge X_i^m to be causally closer to C than X_j^n. An illustrative example of this

would be a food chain: carnivore → herbivore → plant food → climate, as considered in [40].

In [40,49], we have proposed a formalism for examining causality that is particularly natural for application to ENMs. If we have two niche factors X_i^m and X_j^n, we can better understand their potential causal relations with a target C by considering the following relations

$$\varepsilon(C|X_i^m X_j^n; P_N) = \frac{N_{X_i^m X_j^n}(P(C|X_i^m X_j^n) - P_N)}{\sqrt{(N_{X_i^m X_j^n} P(C|X_i^m)(1 - P(C|X_i^m)))}} \quad (4)$$

where P_N represents the null hypothesis with respect to which we will determine the predictability of the combination $X_i^m X_j^n$. If $P_N = P(C)$ we are gauging the consistency of $P(C|X_i^m X_j^n)$ with the null hypothesis that the distribution of the target species is independent of the variable combination $X_i^m X_j^n$. However, we may also choose as null hypotheses $P_N = P(C|X_i^m)$ or $P_N = P(C|X_j^n)$, in which case the null hypothesis is that $P(C|X_i^m X_j^n)$ is independent of X_j^n and X_i^m, respectively. With this approach, we can determine the degree to which X_j^n confounds X_i^m or vice versa. For example, if $P_N = P(C|X_i^m)$ and $\varepsilon(C|X_i^m X_j^n; X_i^m) > 1.96$, we can conclude that within a 95% confidence interval that $P(C|X_i^m X_j^n)$ is not consistent with the null hypothesis and therefore the habitat variable X_j^n is predictive of the distribution of C beyond what is explained by the habitat variable X_i^m. As X_i^m may be a biotic variable and X_j^n an abiotic one, we have used this to show that, very often, biotic factors are confounders for abiotic factors and not vice versa [40]. Although, here, we are concentrating on causal relations and confounding with respect to pairs of habitat variables the formalism naturally extends to larger numbers of variables.

In any spatial cell α we may determine if there is a presence or absence of a given coarse grained bin for either variable, $X_i^m(\alpha) = 0, 1$ and $X_j^n(\alpha) = 0, 1$. For example, $X_8^4(\alpha)X_{10}^2(\alpha) = 11$ would represent a presence of bin 4 of variable 8 and a presence of bin 2 of variable 10 in the cell α, while $X_8^4(\alpha)X_{10}^2(\alpha) = 01$ would represent an absence of bin 4 of variable 8 and a presence of bin 2 of variable 10 in the cell. Thus, there are 4 possible combinations for any pair of habitat variables: presence–presence = 11, presence–absence = 10, absence–presence = 01, absence–absence = 00. By comparing and contrasting the different combinations, we may determine, for example, if presence of one habitat variable is more predictive than the other. Note $X_j^n(\alpha) = 0$ does not imply absence of the variable itself in the cell, just the range denoted by the nth coarse grained bin.

2.7. Data and Habitat Variables

In this paper, considering the evolution of the pandemia in Mexico, we used an irregular spatial partition consisting of the 2458 municipalities of Mexico. The advantage of this partition is that it is most aligned with publicly available socio-demographic and socio-economic factors that can serve as corresponding niche variables. Its disadvantage is that there is a potential bias in the spatial distribution of different municipalities according to their area. For the epidemiological data, used to define the target classes, we used data from the General Directorate of Epidemiology, Secretariat of Health in Mexico [50]. For the socio-demographic and socio-economic data, we used 124 variables taken from the 2010 Mexican Census [51] at the municipality level. For mobility data, we used 12 variables provided by the Institute of Geography of UNAM, that represent the average, daily labour flows between a pair of municipalities. For the air contamination factors, we used three atmospheric pollutants (formaldehyde (HCHO) nitrogen dioxide (NO_2), sulphur dioxide (SO_2)) [52–54], while for climactic data, we used 19 bioclimatic variables from the WorldClim database (www.worldclim.org (accessed on March 2022)) with a spatial resolution of 30 arc-seconds (≈1 km) [55], which includes 11 temperature and eight precipitation indices that express annual trends (e.g., annual mean temperature and precipitation), seasonality (e.g., annual

temperature and precipitation ranges), and environmental extremes (e.g., highest and lowest values of temperature for the warmest and coolest months).

As all the socio-demographic, socio-economic, and mobility data were metric and continuous, for each variable, we ranked the 2458 municipalities from highest to lowest value then divided the ranked list into deciles, with 10% of all municipalities in each decile. By doing this, as opposed to dividing the metric interval for the variable into equal parts, we assure equal statistical weight to each coarse grained category. For the air pollution variables, we divided each raster layer into 20 ranges, which were chosen to have roughly equal number of pixels in each one. Climate variables were also divided into 20 ranges. Thus, in this way we mapped both target class and habitat variables in their entirety into a set of binomial presence/absence variables for each cell on our grid. We also included in as a habitat variable the decile of confirmed cases associated with a given municipality, but at period $t-1$, as opposed to the target variable which was associated with period t. In this way, we could show, in principle, how the history of the habitat can also be included as a predictor. Indeed, this is the first step at showing how the present methodology may be developed to include SIRs-type modelling properties.

3. Results

An important aspect of a complex, adaptive phenomenon, such as the COVID-19 pandemic, is that there are many questions of "where", "when", and "why" that require answers, with each question in turn requiring its own SDM/ENM. Here, we will present representative results for the following: (i) Where are the highest number of confirmed cases in a given time period to be found? (ii) Where are confirmed cases in a given time period to be found? (iii) Where will most confirmed cases be found in a future time period? (iv) Where will cases be found in a future time period where previously there were none? In a more ecological language and taking confirmed cases of COVID-19 as the "species" of interest: (i) Where is the highest abundance of the species of interest to be found? (ii) Where is the species to be found in a given time period independently of its abundance? (iii) Where will the species be found in highest abundance in a future time period? (iv) Where will the species be found where it was not present previously? (In the online system, EpI-PUMA, publicly available in a Platform-as-a-Service environment ((http://covid19.c3.unam.mx), accessed on 5 April 2021, the above and many more different SDM/ENMs are available that use the methodology described in this paper). These five representative models illustrate important differences about how ENM/SDMs can be used to answer where, when, and why questions. Firstly, we may use a model developed on a region $R_{train}(t)$, using data from a time period t, to predict on another region $R_{test}(t)$, as is standardly performed, and where the complete spatial region under consideration is $R(t) = R_{train}(t) \cup R_{test}(t)$. Secondly, we may use a model developed on a region $R(t)$ to predict on $R(t')$, the same region at some later time.

3.1. Dynamic Biotic and Abiotic Factors Can Be Included in an SDM/ENM for COVID-19

To answer each of the above questions, we must extend standard ENM/SDM modelling to include both dynamical target classes, $C(t)$, and associated dynamical habitat variables, $\mathbf{X}(t')$, and develop SDM/ENMs, $P(C(t) \mid \mathbf{X}(t'))$, that relate them. With question (i) the target class will be the 10% of Mexican municipalities that have the highest number of cases in month t. In that context we are characterising niche by the relative abundance of the target species, with those locations where it is highest being characterised as most niche-like. This target class is time dependent, as it may be that those places with the highest abundance at one time are not the same as those at another. With question (ii) the target class will be the presence of confirmed cases in month t. Questions (iii) and (iv) refer to these target classes but in a future month t'. In particular, for the presence of confirmed cases, we will be interested in predicting those municipalities that had no previous cases in time period t but do in period t'.

Thus, the targets for our ENM/SDMs are all, in principle, time-dependent. For the habitat variables, we use those presented in the Methods section below, with an illustrative dynamic habitat variable being the decile of municipalities corresponding to a given range of number of confirmed cases in the period $t-1$.

3.2. Predictive SDM/ENMs Can Be Created for COVID-19

SDM/ENMs were created for the target classes (i)–(v). In Figure 1 (left) we see the results for a purely spatial ENM/SDM model with target class, corresponding to question (i), being the 10% of municipalities with the highest number of confirmed cases, considered for three different, representative months of the pandemic: March 2020, June 2020, and January 2021. In this case, performance is based on a 5-fold 70%/30% train/test split of the 2458 municipalities. Figure 1 (left) shows the corresponding ROC curves and the corresponding AUC for five separate SDM/ENMs for each test set, with: socio-demographic/economic variables only, mobility only, climate only, air contamination only, and all factors together, for each of the three considered months. The relative performance of each sub-model is clear, with the mobility and socio-demographic models being the best performers, followed by the air contamination model, then the climate model, with the total model equivalent to the mobility and socio-demographic models. Table 1 shows the most niche-like and most anti-niche-like habitat factors, as ranked by the statistical measure of co-occurrence ε (see Section 2) (see the Methods section) for each month. As can be seen, the top 18 and bottom 18 ranked factors are almost identical for the three different periods.

Table 1. Top 18 most niche-like and bottom 18 most anti-niche-like habitat factors for the ENM for top 10% of municipalities with highest number of cases ranked by score in March 2020 and showing the corresponding rank in June 2020 and January 2021.

Percentile	Type	Variable	Rank (March)	Rank (June)	Rank (January)
Top	Mobility	10-Internal-labour-flow 25719:664418	**1**	**1**	**1**
Top	Mobility	10-Labour-flow-inward 3279:508685	**2**	**2**	**2**
Top	Mobility	10-Inward-centrality 0.026:0.347	3	3	5
Top	Mobility	10-Inwards-municipalities-connected 64:849	4	4	6
Top	Socio-Demographics	10-Inhabited private homes that have Internet	5	7	3
Top	Socio-Demographics	10-Inhabited private homes that have a computer, tablet or laptop	6	8	4
Top	Mobility	10-Labour-flow-out 4954:287854	7	9	7
Top	Socio-Demographics	10-Total-centrality 0.053:0.416	8	5	10
Top	Socio-Demographics	10-Total municipalities-connected 130:1018	9	6	11
Top	Socio-Demographics	10-Male population aged 18 and over with post-basic education	10	11	9
Top	Socio-Demographics	10-Population aged 18 and over with post-basic education	11	10	8
Top	Socio-Demographics	10-Inhabited private dwellings that have a cell phone	**12**	**12**	**12**
Top	Socio-Demographics	10-Female population aged 18 and over with post-basic education	**13**	**13**	**13**
Top	Socio-Demographics	01-Illiterate population aged 15 and over	14	15	16
Top	Socio-Demographics	01-Illiterate male population aged 15 and over	15	17	19
Top	Socio-Demographics	01-Population aged 15 and over without schooling	16	23	25
Top	Socio-Demographics	10-Population from 15 to 64 years old	17	14	22
Top	Socio-Demographics	10-Inhabited private dwellings that have a television	18	18	14
Bottom	Mobility	01-Total municipalities-connected 2:14	1773	1702	1699
Bottom	Mobility	02-Labour-flow-inward 22:51	1774	1725	1726
Bottom	Socio-Demographics	10-Population with limitation to walk, go up or down	1775	1726	1727

Table 1. *Cont.*

Percentile	Type	Variable	Rank (March)	Rank (June)	Rank (January)
Bottom	Socio-Demographics	10-Population born in the entity	1776	1727	1728
Bottom	Socio-Demographics	02-Inhabited private homes that have a washing machine	1777	1728	1729
Bottom	Socio-Demographics	01-Inhabited private dwellings that have drainage	1778	1754	1757
Bottom	Socio-Demographics	10-Inhabited private dwellings that do not have drainage	1779	1755	1758
Bottom	Socio-Demographics	01-Inhabited private dwellings that have a refrigerator	1780	1756	1759
Bottom	Socio-Demographics	01-Population from 6 to 11 years old that does not attend school	1781	1610	1587
Bottom	Socio-Demographics	07-Population with limitation	1782	1773	1774
Bottom	Mobility	03-Internal-labour-flow 698:1179	**1783**	**1783**	**1783**
Bottom	Mobility	01-Internal-labour-flow 9:329	**1784**	**1784**	**1784**
Bottom	Mobility	01-Labour-flow-out 1:20	**1785**	**1785**	**1785**
Bottom	Mobility	01-Labour-flow-inward 1:21	**1786**	**1786**	**1786**
Bottom	Mobility	01-Outwards-centrality 0:0.003	1787	1789	1787
Bottom	Mobility	01-Outwards-municipalities-connected 1:7	1788	1790	1788
Bottom	Mobility	01-Inward-centrality 0:0.002	1789	1667	1650
Bottom	Mobility	01-Inwards-municipalities-connected 1:5	1790	1668	1651

In Figure 2 (bottom), we show the values of the score $s(X_i^m(t))$ (see Section 2) for the 10 deciles, $m = [1, 10]$, of the habitat variable X_i =*Internal labour flow of the municipality*, observing that the score contributions are highly conserved across the three time periods. This niche conservation effect is shown in an even more striking fashion in Figure 3, where we see the correlation between the $\varepsilon(X_i^m(t))$ values (left) and the score values, $s(X_i^m(t))$, (right) for the ENM models for each individual habitat variable for the three different periods. For instance, we observe that the $\varepsilon(X_i^m(t))$ values from March 2020 explain 96% of the variance in the $\varepsilon(X_i^m(t))$ values for January 2021. As a further test of niche conservation, we used an ENM trained on all of the cells in one month to predict a future month. In Figure 1 (right), we see the ROC and AUC for the five models using different categories of habitat variable (mobility, socio-demographic, air contamination, climate, and all) trained on data from all spatial cells for the period March 2020 and tested on the periods June 2020 and January 2021. Similarly, we show the result of analogous models trained on the period June 2020 and tested on the period January 2021.

In Figure 4 (left), we see the results for a purely spatial ENM/SDM model with target class, corresponding to question ii), being those municipalities with the presence of confirmed cases of COVID-19 for the months: March 2020, June 2020 and January 2021. Performance is again based on a 70%/30% train/test split of the 2458 municipalities, with Figure 4 (left) showing the corresponding ROC curves and the corresponding AUC for five separate SDM/ENMs for each test set, with: socio-demographic/economic variables only, mobility only, climate only, air contamination only, and all factors together, for each of the three considered months. Again, the relative performance of each sub-model is clear, with the same relative performance as for the previous models. However, we note that for the overall performance of the mobility and socio-demographics of all models is less than their counterparts in the case of top 10% of municipalities with most cases as class, while the performance of the climate and air contamination models is similar.

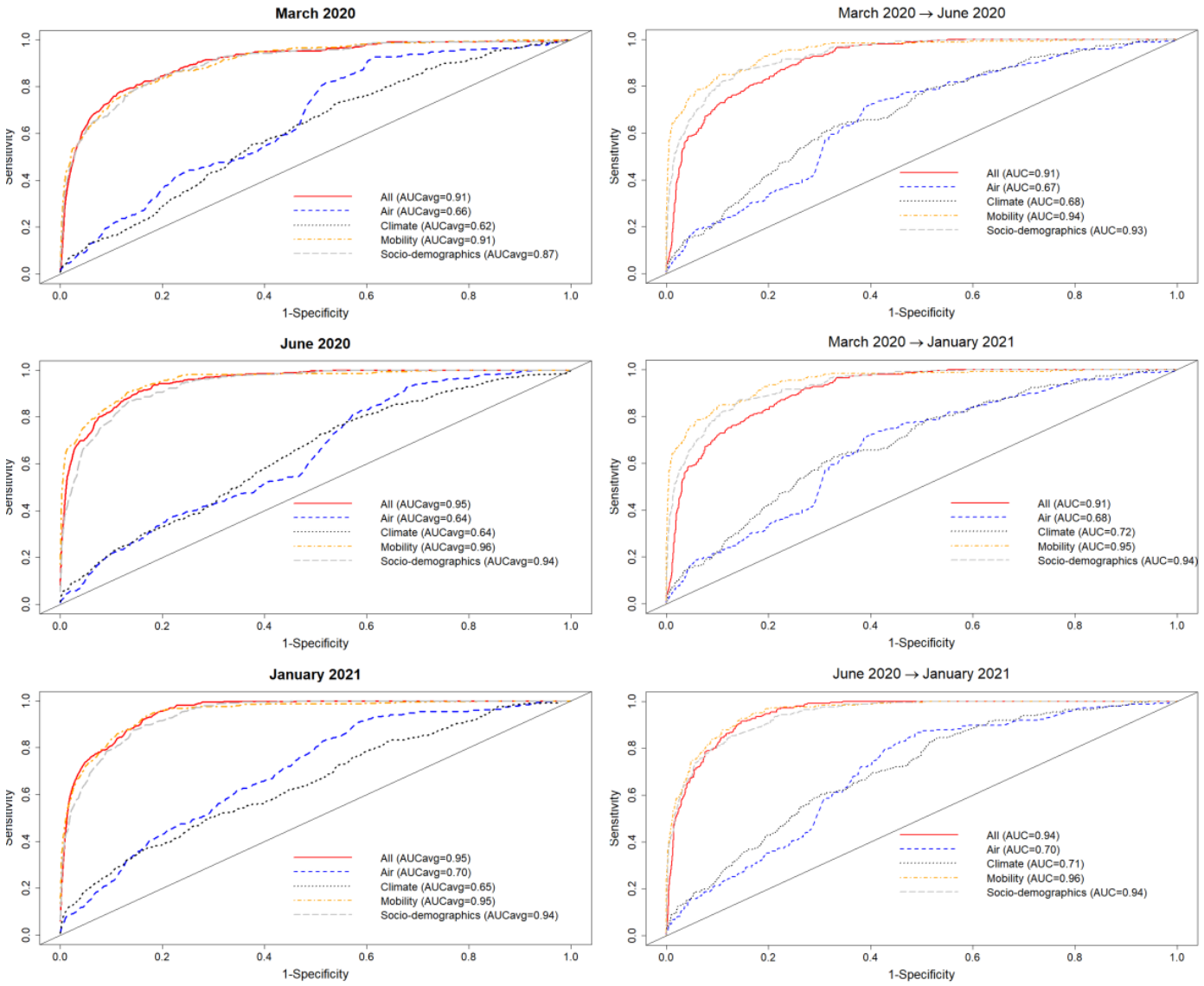

Figure 1. Left: Performance for five different SDMs for the class—top 10% of municipalities with the highest number of cases—according to the habitat variable group considered (mobility, socio-demographics, air, climate, and all) for three different months of the pandemic. Five different train-test 70%–30% splits were considered for each month; Right: Performance for five different SDMs for the class—top 10% of municipalities with the highest number of cases—according to the habitat variable group considered (mobility, socio-demographics, air, climate, and all) using a model trained on one month and tested on a future month.

In Figure 5 (left), in analogy with Figure 3, we see the correlations between $\varepsilon(X_i^m(t))$ values for all habitat variables for the model for the presence of COVID-19 cases for two different months, while in Figure 5 (right), we show the correlations between $s(X_i^m(t))$ values for all habitat variables for the same model. Note that the correlation between the $\varepsilon(X_i^m(t))$ and $s(X_i^m(t))$ distributions in March 2020 and June 2020 and March 2020 and January 2021 are now significantly smaller than their counterparts for the top 10% model.

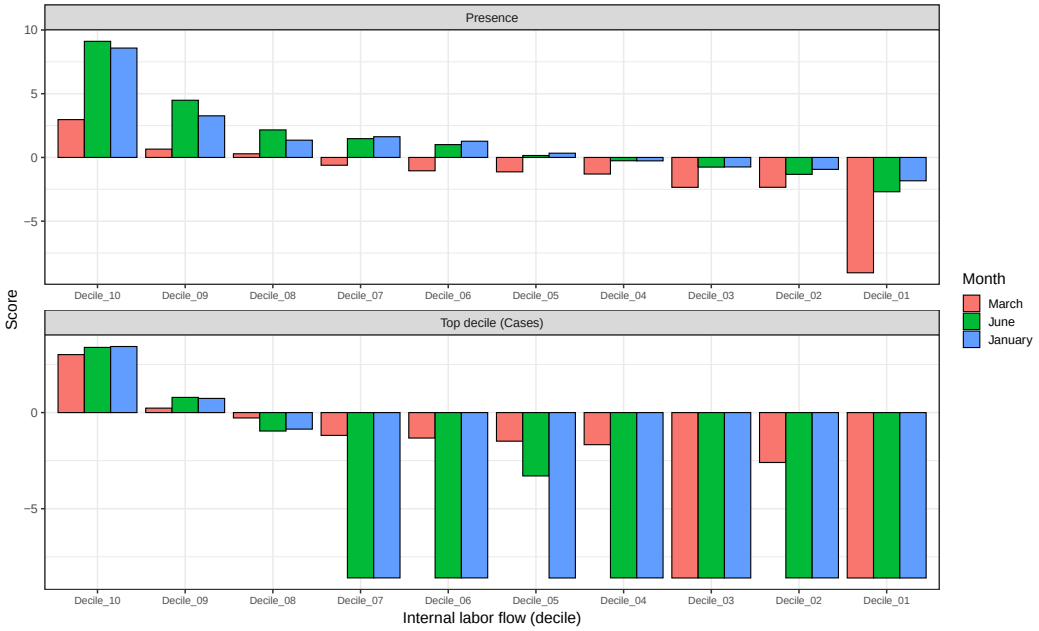

Figure 2. Top: Model scores for the 10 deciles ($Decile_i$) of the habitat variable *Internal labour flow of the municipality* for the model predicting presence of COVID-19 cases; **Bottom**: Model scores for the 10 deciles ($Decile_i$) of the habitat variable *Internal labour flow of the municipality* for the model predicting the top 10% of municipalities with highest number of cases.

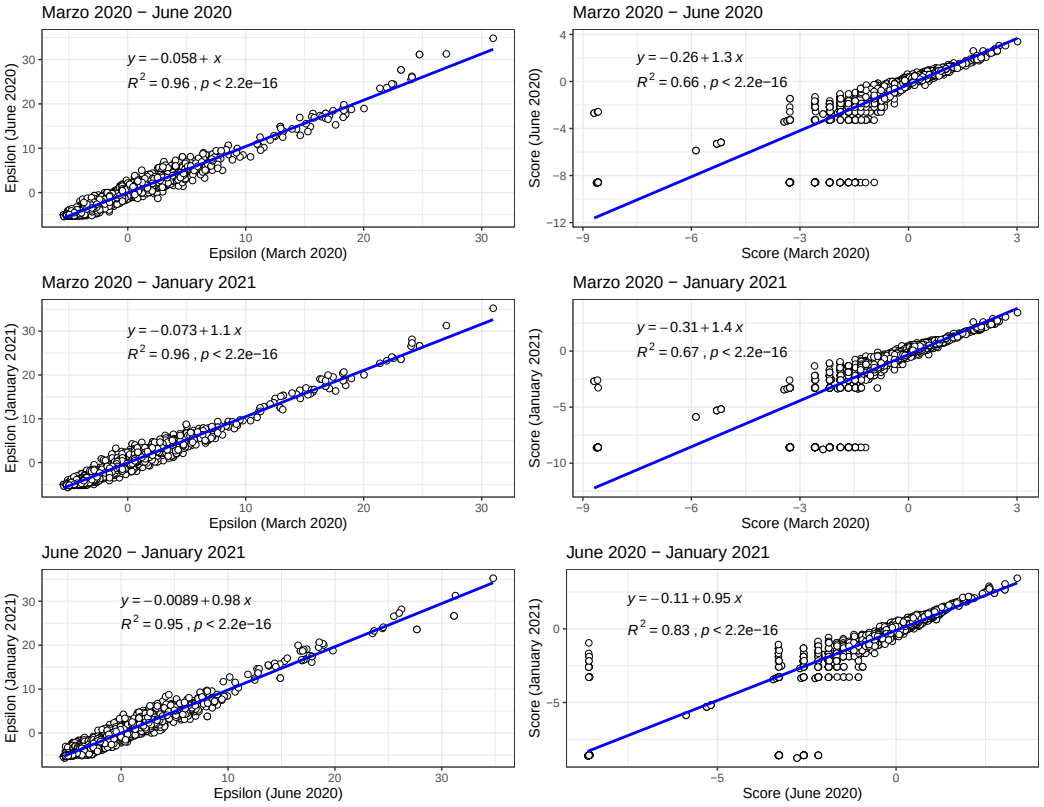

Figure 3. Left: Correlations between $\varepsilon(X_i^m(t))$ values for all habitat variables for the model with class top 10% of municipalities with highest number of cases for two different months; Right: Correlations between $s(X_i^m(t))$ values for all habitat variables for the model with class—top 10% of municipalities with highest number of cases—for two different months.

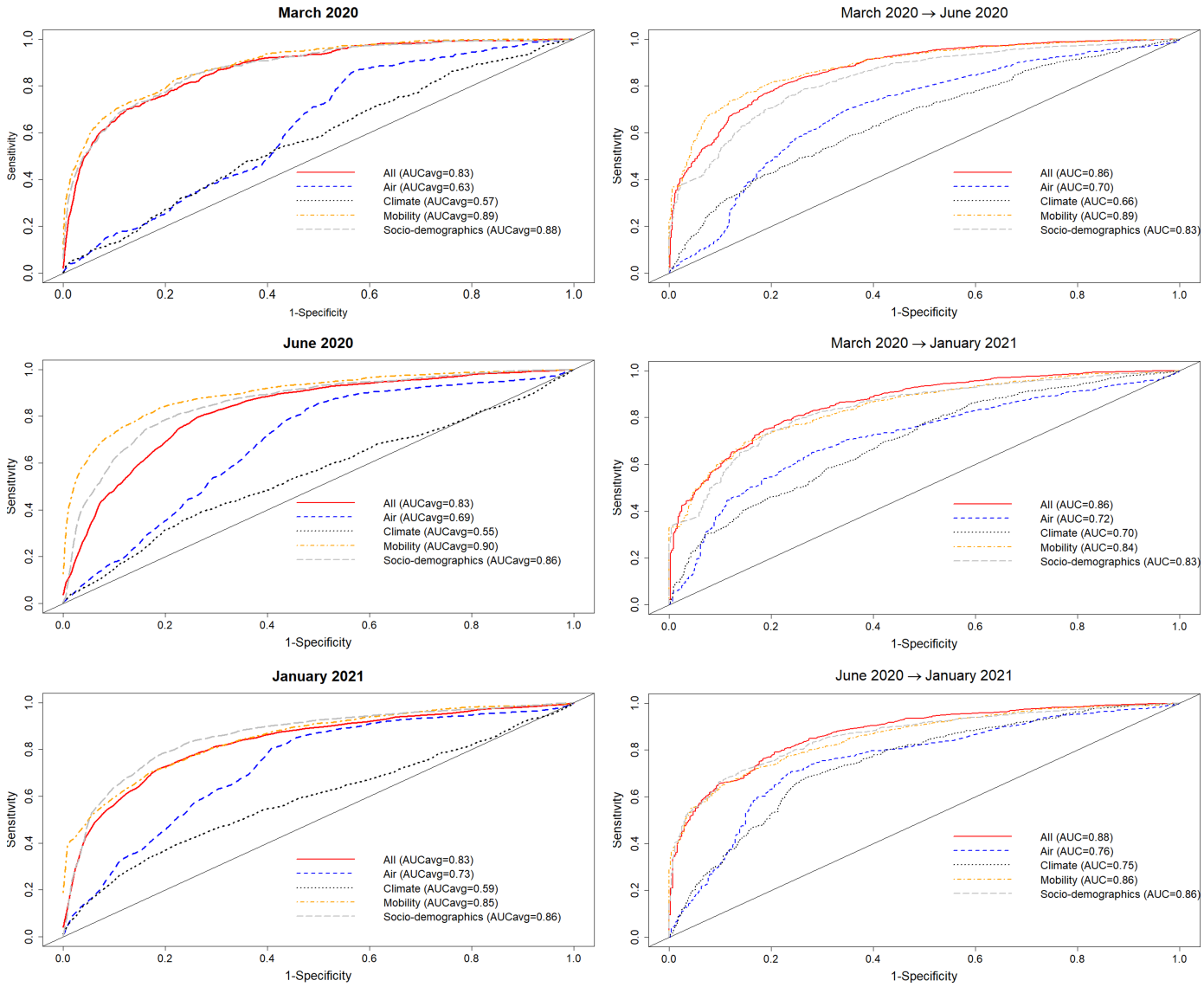

Figure 4. Left: Performance for five different SDMs for the class—presence of COVID-19 cases in the municipality—according to the habitat variable group considered (mobility, socio-demographics, air, climate, and all) for three different months of the pandemic. Five different train-test 70%–30% splits were considered for each month; Right: Performance for five different SDMs for the class presence of COVID-19 cases in the municipality according to the habitat variable group considered (mobility, socio-demographics, air, climate, and all) using a model trained on one month and tested on a future month.

Furthermore, in Figure 2 (Top) we show the values of the score $s(X_i^m(t))$ for the 10 deciles, $m = [1, 10]$, of the habitat variable X_i = Internal labour flow of the municipality to compare and contrast the score contributions of this key niche variable as a function of time. We observe that the relative contributions from each decile differ significantly when compared to the results for The top 10% of municipalities with highest number of cases, as shown in Figure 2 (Bottom).

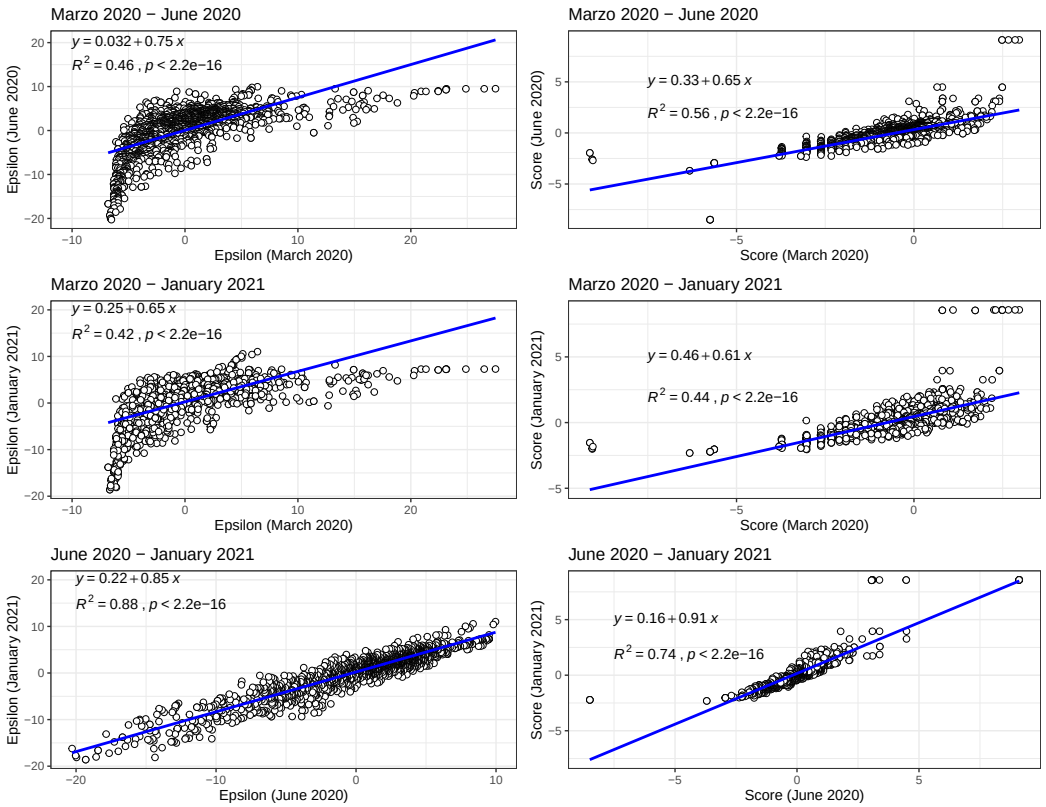

Figure 5. Left: Correlations between $\varepsilon(X_i^m(t))$ values for all habitat variables for the model for the presence of COVID-19 cases for two different months; Right: Correlations between $s(X_i^m(t))$ values for all habitat variables for the model for the presence of COVID-19 cases for two different months.

Turning now to the above ENMs at time t as predictors of the species distribution at t', in Figure 4 (Right), we see the results of a model trained on data from those municipalities with confirmed cases in March 2020 and June 2020 to predict which municipalities would have cases in June 2020 and January 2021 and January 2021, respectively. In this case, we observe that the ENM for month t is less predictive for month t' than the corresponding model for identifying the top 10% of municipalities with the highest number of cases.

3.3. ENMs Can Be Used to Predict Numbers of Cases

As well as using $S(C|\mathbf{X})$ as a pure classifier (see Section 2), to identify a particular class, such as presence/absence, top 10% highest cases, etc., we can also use it to predict the actual number of cases. To perform this, we regress the total score $S(C|\mathbf{X})$ against the number of cases on our training set and then apply it to the test set. An example is shown in Figure 6, for the classification model with $C =$ top 10% of municipalities with the highest number of cases, where the training set is a randomly chosen 70% of spatial cells and the test set the remaining 30%. All habitat variables were used. Figure 6 (left) shows the relation between score and number of confirmed cases for our three example months—March 2020, June 2020, and January 2021. First, municipalities were ranked according to their score, then grouped into 10 deciles. An exponential function fit was used. Figure 6 (right) shows the results of applying the model to the 30% out of sample test set, showing the relation between predicted and actual abundances.

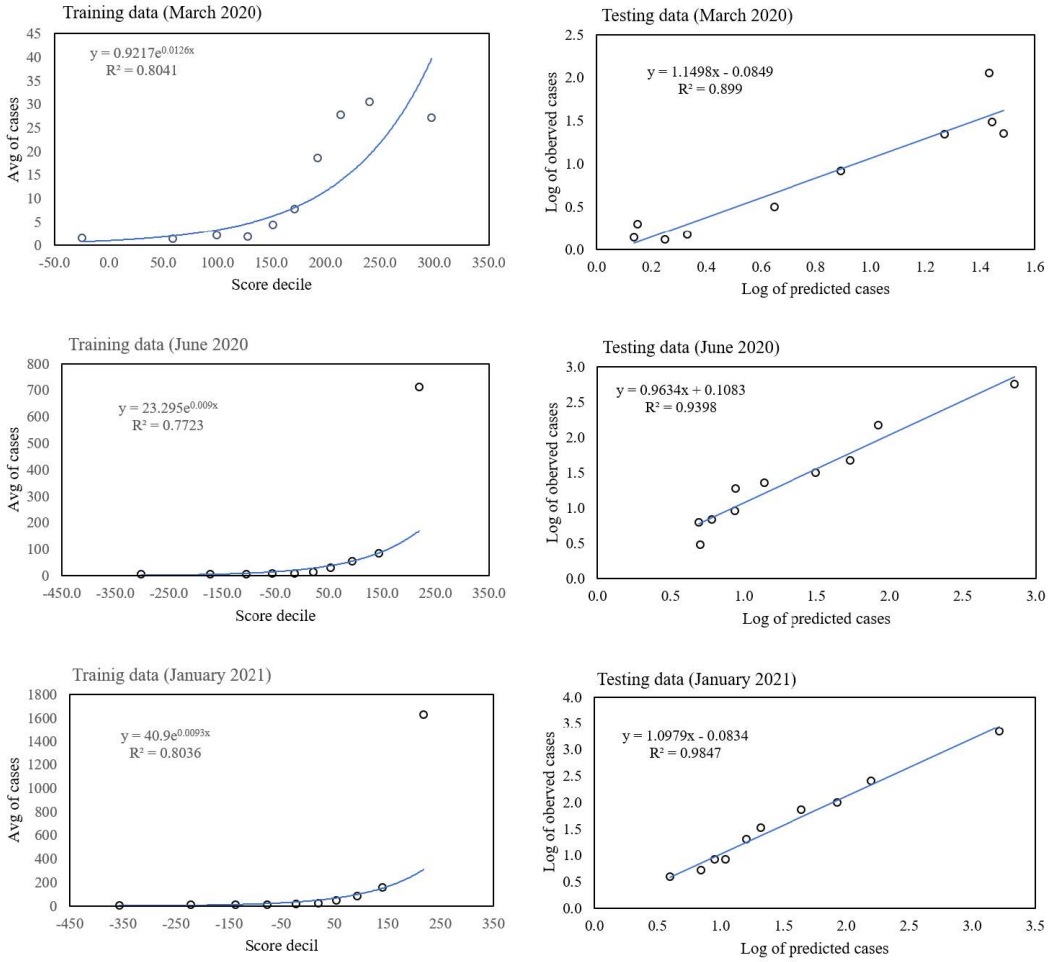

Figure 6. Left—graphs of score from a model for predicting the top 10% of the highest number of cases versus the number of confirmed cases of COVID-19 for the three months March 2020, June 2020, and January 2021 using as training set 70% of spatial cells. An exponential function was used to fit the relation score–number of cases. Right—graphs of number of predicted cases versus the number of actual cases for the 30% hold out set for the same three months.

3.4. Causal Relationships Can Be Deduced

As an illustration of the formalism for disentangling causal relationships (see Section 2), we consider the relative contributions of climatic factors, as represented by the WorldClim variable *Average annual temperature*, and human mobility factors, as represented by *Internal labour flow of the municipality*, to prediction of the top 10% of municipalities with the greatest number of cases. The temperature variable is divided into 20 coarse-grained bins, while the mobility variable is divided into 10 bins. In any spatial cell, α, we may then determine if there is a presence or absence of a given coarse grained bin for either variable. Thus, in Equation (4), we have X_i^m representing the 10 different bins of the mobility variable and X_j^n representing the 20 bins of the temperature variable. As, for a given spatial cell, there are four possibles states, corresponding to presence/absence of the two habitat variables, $X_i^m X_j^n$ can be represented by four 20×10 matrices.

For each cell of each matrix, we calculate $P(C|X_iX_j)$ and also $\varepsilon(C|X_i^m X_j^n)$, using as null hypothesis $P(C)$, as shown in Figure 7, with data from January 2021. Note that the absence of a variable bin, X_i^m, corresponds to those cells where there is no presence of X_i^m. However, there will be a presence of at least one other $X_i^{m'}$, for $m' \neq m$. For example, in the matrix $Clim0\_Mob0$, the cell $Tmp.p04 = 0,\ 08\_Inter - Mob = 0$ corresponds to those cells that are not in the average annual temperature range corresponding to $Tmp.p04$ and also not in the Internal labour flow range corresponding to $08\_Inter - Mob$. However, that

set of cells could have presences of temperature ranges other than $Tmp.p04$ and mobility ranges other than $08\_Inter - Mob$.

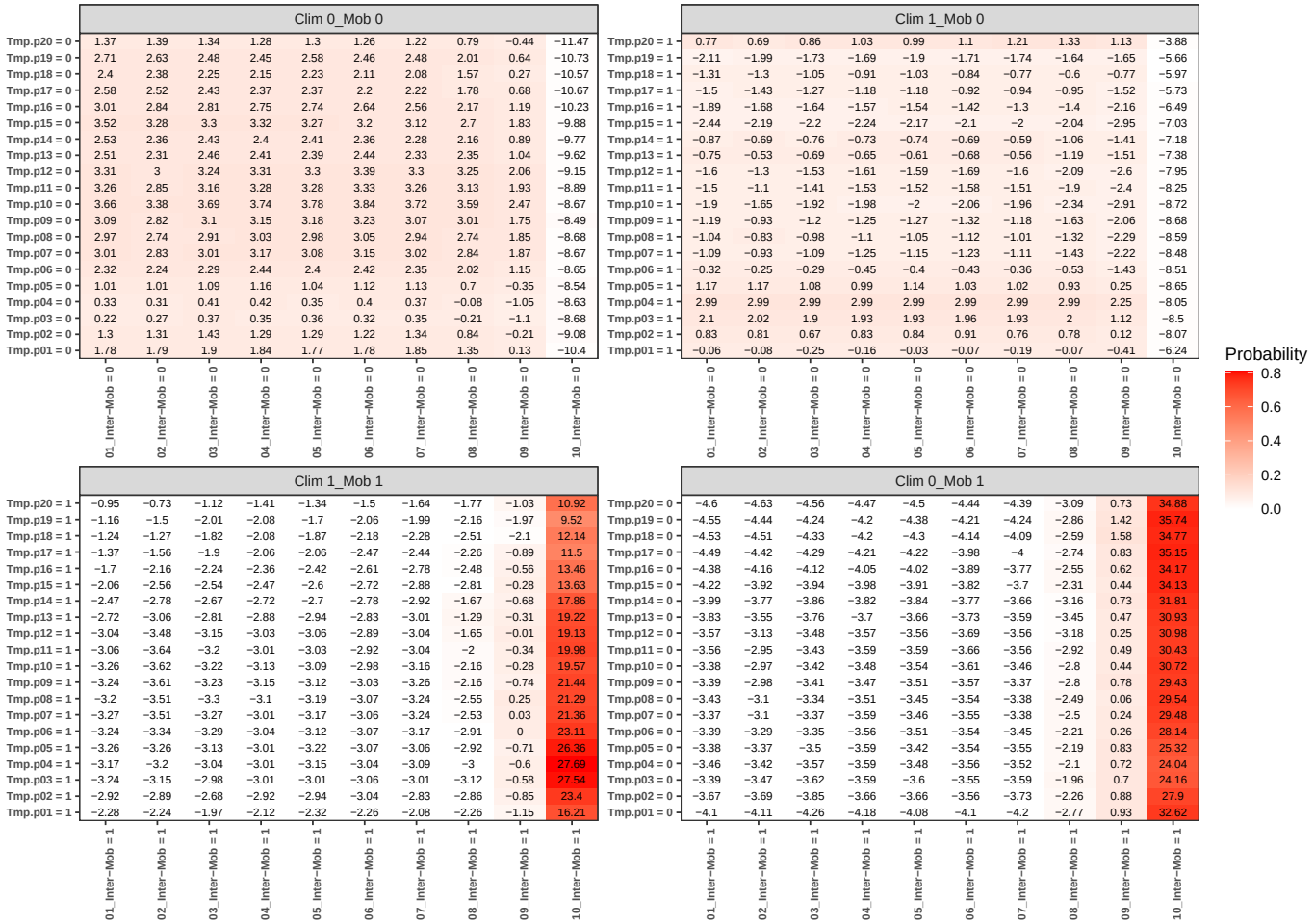

Figure 7. $\varepsilon(C \mid X_i^m X_j^n)$ and $P(C \mid X_i^m X_j^n)$ for each combination of habitat variable ranges $X_i^m =$ *Average annual temperature* and $X_j^n =$ *Internal labour flow of the municipality* for the four combinations $X_i^m = 0, 1$ and $X_i^m = 0, 1$ corresponding to absence/presence for each variable bin, respectively.

The fact that the climate variable is confounded by the mobility variable is manifest in the matrix $Clim1\_Mob0$, where there is a presence of a particular climate variable bin and an absence of the corresponding mobility variable bin. For instance, the cell $Tmp.p04 = 1$, $10\_Inter - Mob = 0$ in that matrix corresponds to those municipalities that are not in the highest decile of mobility, but are in the fourth decile of average annual temperature. $\varepsilon(C|X_i^4 X_j^{10}) = -8.05$ and $P(C|X_i^4 X_j^{10}) \sim 0$ there, indicating that the probability of being in the top 10% of highest number of cases is very low. Indeed, we can see that for any temperature range, for $10\_Inter - Mob = 0$, there is very little chance of the corresponding municipalities being in the top 10%. On the contrary, for $Clim0\_Mob1$, we see that $P(C|X_i^m X_j^{10}) \sim 0.8$ for any m, thus indicating that high mobility is very niche-like, independently of the value of the temperature variable. Although for $Clim1\_Mob0$ the range $Tmp.p04$ can give rise to statistically significant $\varepsilon(C|X_i^4 X_j^n)$ values for $m \neq 10$, by examining the case $Clim1\_Mob1$, we see that the $\varepsilon(C|X_i^4 X_j^n)$ values for $n \neq 9, 10$ are all negative and statistically significant. Thus, climate may appear to be niche-like but, in fact, is confounded by mobility, and now we can intuit why a climate model can create predictability even though climate is not predictive in and of itself.

A model based on climate only is equivalent to the combination $X_i^m = 1$, $X_j^n = *$, where $*$ denotes that we have marginalized over this value. For instance, $P(C|X_i^m) =$

$\sum_{X_j^n=0,\,1} P(C|X_i^m X_j^n) = P(C|X_i^m X_j^n = 0) + P(C \mid X_i^m X_j^n = 1)$. Although, $P(C|X_i^m X_j^n = 0) \sim 0$ for $n \neq 9,\ 10$, $P(C|X_i^m X_j^{10} = 1) \sim 0.6 - 0.8$ for any m. Thus, the primary reason why the climate model has some degree of predictability is that those places of highest mobility have a climate "profile", i.e., they are not equally distributed across all average annual temperature ranges. It is not, however, the climate that is causal. Another way to see this is shown in Figure 8, where we show the marginal probabilities $P(C|X_i^m)$ and $P(C|X_j^n)$, which would correspond to the results associated with an ENM based only on the climate variable *Average annual temperature* or on the human mobility variable *Internal labour flow of the municipality*, respectively.

We can note that, although the degree of predictability in the climate-only model is weak, there is a variation, with $\varepsilon(C|X_i^m = 1)$ ranging from 0.9 to -2.9. However, the origin of the most niche-like value of 0.9 is associated with the residual predictability from decile 10 of the mobility variable.

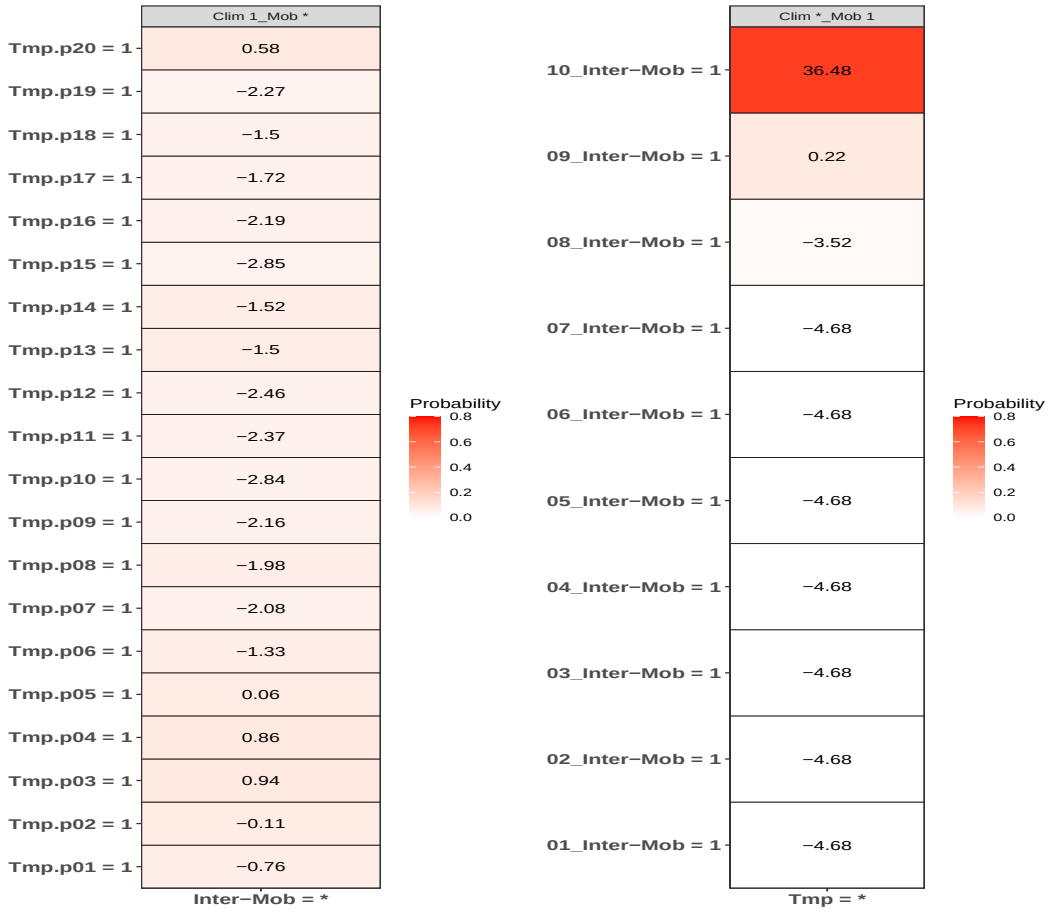

Figure 8. $\varepsilon(C \mid X_i^m = 1)$, $\varepsilon(C \mid X_j^n = 1)$, $P(C \mid X_i^m = 1)$ and $P(C \mid X_j^n = 1)$ for each value of the habitat variable ranges X_i^m = *Average annual temperature* and X_j^n = *Internal labour flow of the municipality* corresponding to a "climate-only" model and a "mobility-only" model.

4. Discussion

Our goal in this article was to demonstrate that the questions: "Does a respiratory virus have an ecological niche, and if so, can it be mapped?" can be answered in the affirmative. We have explicitly created several ENMs and SDMs for COVID-19 that are both predictive and contain habitat factors that are more causally plausible than climate, for instance. In order to achieve this, we introduced several innovations compared to standard niche and species distribution modelling. Firstly, we showed how to extend niche and species distribution modelling to "non-equilibrium" situations, where both target and niche variables are potentially time varying, as well as the relation between them.

Secondly, we created models with habitat variables that were represented by quite different data types and associated spatial resolutions. Finally, we showed how causal relations and confounding can be better understood by introducing a hierarchy of conditional probabilities and the associated intuition that a more causally direct factor should have a bigger effect than an indirect one.

We constructed ENMs as Niche Landscapes, $P(C(t)|\mathbf{X}(t))$ - Bayesian posteriors which serve as a height function on a Hutchinsonian ecological space with 2749 dimensions, spanning air pollution, climate, mobility, socio-economic, and socio-demographic data. Usually in ENM/SDMs, the target class, C, is a binary variable, such as presence/no presence of a species. In Coro [5], the corresponding variable was "high" infection rate, defined in a binary fashion with respect to a reference infection rate. We too, have used binary class variables, by choosing a particular subgroup of spatial cells, corresponding to presence/absence of confirmed cases, or if a spatial cell was in the top 10% of highest total infections (In the EpI-PUMA system, publicly available in a Platform-as-a-Service environment (http://covid19.c3.unam.mx, accessed on 5 April 2021), 72 different SDM/ENMs are available that use the methodology described in this paper.). However, the binary nature of the target variable is a choice rather than a restriction. For instance, taking infection rate as a continuous variable, this can be divided into as many quantiles as we please, say n_c. The target variable now consists of n_c "presence/absence" variables, each with its own ENM/SDM, $P(C_i|\mathbf{X})$, $i \in [1, n_c]$. Even in the case of a binary decomposition of a continuous variable; however, the metric nature of the variable leaves an imprint on $P(C|S(C|\mathbf{X}))$, such that higher score, as a continuous metric variable, corresponds to higher infection rate or number of cases, as we have demonstrated, leading to a model that can predict abundances.

In criticising ENM/SDMs as a useful tool for the COVID-19 pandemic, or any other, it is important to distinguish between applicability of ENM/SDMs in general versus a particular instance of an ENM/SDM. It is appropriate to criticise a model that includes climate, and which has been used to infer a corresponding causal effect on the pandemic, without any analysis of possible confounders. This does not mean, however, that with appropriate habitat variables and a methodology for disentangling confounding, that useful ENM/SDMs cannot be constructed, as we have shown here. Our results clearly show that models for predicting where the highest number of cases will be, that are built on mobility, socio-demographic and socio-economic data, are much more predictive than models built on climate and/or air contamination, as can be seen in the results of Figure 1, and that this has been true throughout the pandemic. As seen in Table 1, the most important niche/anti-niche factors for this target are all associated with the highest/lowest levels of mobility, as proxied by inter- and intra-municipal labour flows, and a particular socio-economic profile. The fact that a climate model can exhibit some degree of predictability does not mean that climate is the direct driver. On the contrary, we can ask if any apparent predictability due to climate factors is confounded by human-based factors. As we have shown in Figures 7 and 8, this is indeed the case. Just the distribution of the habitat variable scores themselves tells us that if there is confounding then it is the human factors that confound climate and not vice versa, as a confounder should be causally closer to its effect than the confounded variable and therefore have a higher score, as is implicit in the Bradford–Hill criteria.

With respect to the socio-demographic and socio-economic factors, we see that they have a similar predictive power to the mobility factors and reflect the socio-demographic and socio-economic conditions where COVID-19 cases are highest. For instance, the habitat variables *% of households that have a computer* and *% of households that have internet access* are both significant niche factors, as are other factors that correspond to a more educated population with higher economic status. Of course, the interpretation of these factors is not as intuitive as mobility and we certainly do not wish to attribute direct causality to them. However, there are a variety of relevant factors for COVID-19 that can be related to internet access and computer usage for example, such as age, educational status, population density [56], as urban areas have better infrastructure, and mobility itself [57,58]. As with

any epidemiological or ecological model, the interpretation of the predictors as representing direct versus indirect interactions is highly non-trivial. The formalism we have introduced for disentangling confounding can help in this regard.

The places where COVID-19 is in highest abundance represent one particular characterisation of the spatio-temporal distribution of the COVID-19 "species" and its associated ecological niche. Where it is present and where it will be in the future compared to where it is now are two others. Hutchinson [7,59] defined niche as that region of ecological space where the net growth rate of the species is $r \geq 0$ at low density. For a pathogen that is not capable of free movement and is dependent on a host, there are two natural growth rates: one associated with the pathogenic load within a host and another that is measured by the number of infected hosts. Obviously, in this paper we are concerned with the latter, where the growth rate is characterised by the basic reproduction rate, R_0, or the effective reproduction rate, R_t [60]. If we defined niche for COVID-19 through an analogue of $r \geq 0$, such as $R_t > 1$, we would clearly be in a situation where we were passing from "niche" to "anti-niche" ($R_t > 1 \rightarrow R_t < 1$) and vice versa, continuously in space and time due to a multitude of factors, including public health interventions, such as lockdowns and vaccinations, as well as resistance to vaccinations, new variants, and a host of others. That these factors alter the ecological conditions in a certain place at a certain time is undeniable. However, to keep track of such changes and how they impinge on how niche-like conditions are in a certain spatio-temporal cell, (α, t), requires the corresponding data. Here, we have preferred to use characteristics of the pathogen distribution that are easier to measure—number or presence of cases—but with which we can characterise the concept of niche, and a set of habitat variables that go beyond those previously considered. We take highest case abundance to distinguish those conditions in ecological space, and in geographical space and time, that favour higher abundance of the pathogen, as proxied by abundance of cases. This is a *relative* measure, as it may be that a municipality, α, has $R_t(\alpha) < 1$ but such that it is higher than any other.

Holt [61] has suggested that besides an "Establishment Niche", that corresponds to the original Hutchinsonian niche, a "Population Persistence Niche", associated with the range of niche space in which populations above some threshold density, $N > 0$, can persist, is a complementary notion. We take where COVID-19 is in highest abundance to be closer to this Population Persistence Niche, whereas where it is present is closer to the original Hutchinsonian characterisation and, especially, as it is portrayed in the majority of ENM/SDMs [37], where presence/absence is used to characterise both. On the other hand, where it is now versus in the past corresponds to neither, but is taken to reflect the potential range expansion/contraction of the species. We believe that these examples, and more, show that there are multiple characteristics of a species distribution that can and should be modelled and can be used to characterise complementary notions of niche.

The methods we have exhibited and the corresponding examples also allow us to understand to what degree a niche is conserved. In the case of species abundance, we saw that the niche associated with those places where the relative number of COVID-19 cases was highest was highly conserved, with very little difference across the entire pandemic. Moreover, we showed how this conservatism could be quantified using our statistical diagnostic $\varepsilon(C \mid X_i^m)$ and the score functions, $s(X_i^m(t))$, associated with the habitat variables X_i^m, with the time dependence of the associated score function reflecting the relationship between target and habitat variables. For example, if $s(X_i^m(t))$ is strongly positive at one time, t, and not another, t', then we may say that the variable X_i^m is niche-like with respect to C at time t, but not at t'. Niche conservatism with respect to X_i^m is then quantified by $s(X_i^m(t)) \sim s(X_i^m(t'))$. This niche conservatism is also manifest in that an ENM created at time t provides a SDM that is just as predictive at a later time t' as at t. Thus, from a Hutchinsonian perspective, the relation $P(C(t) \mid \mathbf{X})$ may be conserved in ecological space, even though the spatial distribution of $C(\alpha, t)$ and/or $\mathbf{X}(\alpha, t)$ could change in time.

In the case of presence/absence as target class, the corresponding realised niche is not conserved, in that $s(X_i^m(t)) \neq s(X_i^m(t'))$. This is seen in Figure 5, when compared

to Figure 3, where the regression slopes for the March 2020–June 2020 and March 2020–January 2021 comparisons have lower R^2 values and also slopes < 1, indicating that the scores in January or June are only about 60% of their values in March. The R^2 values and slopes for the ε comparisons in the same figures show the same effect. These differences just reflect the range expansion of COVID-19, where it has been argued that: "there is no unsuitable habitat" for COVID-19 [2]. This is linked, however, to the notion of presence, not to abundance. The fact that the score and ε contributions are diminishing in the presence model from March 2020 to June 2020 and January 2021 is due to the fact that the habitat variables are less able to discriminate between those cells where cases are present versus absent. The differences between June 2020 and January 2021 are much less as, by this time, a large majority of municipalities now had confirmed cases. We can also see this niche non-conservation in Figure 3 (top), using as an example the variable *Internal labour flow of the municipality*. We see that in March 2020 only the first three deciles of municipalities ranked by that score are associated with a positive score—niche-like—whereas in June 2020 and January 2021 60% have positive scores. Similarly, we see that in the most anti-niche-like deciles, *Decile*_02 and *Decile*_01, the scores are becoming less negative, indicating that those municipalities with the lowest internal labour flows are becoming less anti-niche-like. If the pandemic had a presence in *every* municipality, then every $s(X_i^m(t)) \to \infty$, corresponding to the fact that there can be no discrimination between where the species is present and where it is absent. Everywhere is niche-like. This would not be the case, however, for abundance, as is seen in Figure 3 (bottom), where the change is due to the fact that our relative abundance measure is associated with the top 10% of municipalities with the highest number of cases. The range expansion of COVID-19 presence also has an impact on the performance of the corresponding SDMs, as seen in Figure 4, where we see that for the all, socio-demographic, and mobility models, the corresponding AUC for the presence/absence spatial models is much less than their abundance counterparts.

We have also shown how a classification-based ENM can be used to predict abundances, with an example being the number of confirmed cases of COVID-19 for a given month. We see that the score based on all habitat variables explains approximately 90% of the variance. We believe that the ENM/SDM formalism we have developed here has the capacity to be truly epi-ecological/eco-epidemiological given the right habitat variables. The static habitat variables we have chosen cannot account for the dynamic expansion and contraction that is characteristic of epidemics and which naturally emerges from a mechanistic SIR-type modelling formalism. What is required are habitat variables that are the analogue of what enters into a differential equations type modelling environment: changes in abundances from one time period to another, for example, or even changes in those changes, as it it these variables that account for the underlying dynamics. Indeed, an ENM built on a given time slice that does not explicitly account for such variables will either underestimate or overestimate abundances, depending on whether the epidemic is in an expansion or contraction phase. We will return to this in a future publication.

We have also shown how it is possible to distinguish confounding and how this can be used as a tool to disentangle causality. As a test case we considered combining a behavioural/socio-economic/socio-demographic variable—*Internal labour flow of the municipality*—and a climatic variable, showing how climate as an apparently predictive habitat variable is confounded by the more predictive socio-economic and mobility factors.

An apparent limitation of our work is that we have worked at a quite coarse-grained level, that of municipalities. This is a limitation of the data used, however, not the methodology. In principle, a much finer spatial resolution, at the level of "census blocks", for instance, could be used if data were available at that resolution for the target class and habitat. In the same vein, dynamical data that represented both changes in the local environment, such as lockdowns and hospital occupation rates, or changes in behaviour, such as mask wearing compliance, cell phone data as a proxy for short-term mobility, or vaccination rates, or a host of other factors could also be included. It would be interesting for example to determine to what extent adaptations in the pathogen were potentially reflected as changes

in its niche. There is also the question of the validity of the target class data as maintaining accurate counts of confirmed cases is difficult. However, this would have little to no impact on the overall conclusions of our ENM/SDMs.

In summary, the five principal advantages of our methodology relative to standard ENM/SDMs are: (i) First and foremost, our methodology can account for the highly multifactorial nature of COVID-19 as a complex adaptive system, where there are many directly or indirectly causal factors that affect both its spread and its magnitude; (ii) It is Bayesian in nature and therefore has the advantage of being adaptive, where the incorporation of new information can be achieved naturally using Bayes theorem, which, in addition, also allows for the incorporation of human expertise by means of Bayesian priors, as well as the addition of data-based evidence; (iii) By considering the incorporation of new information in time, any time variation in the relation between a target class C and its corresponding niche variables, $\mathbf{X}$ can be tracked to determine how conserved the niche is; (iv) It permits a more profound analysis of causality and confounding through the consideration of a hierarchy of conditional probabilities; (v) It naturally permits the construction of niche models associated with different notions of niche – presence/absence, abundance, etc.—as these are simply representable as distinct classes of interest, C. The models we have presented can predict "absolutes", such as the number of cases in a given municipality, transversally, i.e., on the same time slice. Additionally, they can predict "relatives", such as the relative number of cases in a given municipality longitudinally. We have shown that the corresponding models are accurate, with a high degree of correlation between predictions and actual numbers. Furthermore, that accuracy is intimately related to the incorporation of relevant niche variables such as mobility, socio-economic, and socio-demographic factors. Our goal here was not to offer a gold-standard model for prediction of the pandemic. As mentioned, there are more predictive and directly causal factors than we have included here. However, the niche of COVID-19 is immensely complex and adaptive and the incorporation of the vast array of relevant factors that determine it is a huge challenge in data collection and integration. What we have shown is that if that data can be represented in space and time, $\mathbf{X}(\alpha, t)$, then it may be incorporated into an ENM/SDM using the methodology we have shown here. Moreover, the innovations we have shown are independent of the specific use case of COVID-19 considered here. They represent general extensions of current niche and species distribution modelling, and can be applied to any ecological system, where they are necessary. In particular, they can be applied to situations where the target and/or niche variables are changing in time. Aside from disease dynamics, invasive species, habitat, and biodiversity loss are other prime areas where time dependence is crucial and where our approach can be used. Moreover, we believe that taken in its fullest sense, where a niche/anti-niche represents the complete set of both biotic and abiotic drivers that favour where a "species" is or is not, or at least should be, a universally applicable concept, with the SDM determining the "where" and the ENM the corresponding "why". Indeed, its applicability is only restricted in an ecological setting by just what we mean by "ecological". If we take ecology to cover any interaction between biota and the environment then we should accept factors as mask-wearing compliance as a potential niche factor for example. Furthermore, when thinking of a "Species" distribution model, we may encouraged to be liberal in our notion of "species", where it may represent, for instance, cases of non-transmissable diseases, such as diabetes or heart disease.

In conclusion, there is a difference between stating that ENM/SDMs generally are inappropriate vehicles for modelling a dynamic phenomenon such as the COVID-19 pandemic versus stating that a particular ENM/SDM is inappropriate. We have shown that ENM/SDMs can be generated which overcome such criticisms.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/tropicalmed8030178/s1.

Author Contributions: Conceptualization, C.R.S. and C.G.-S.; methodology, C.R.S., C.G.-S., and P.R.-M.; formal analysis, C.R.S. and C.G.-S.; data curation, C.G.-S. and P.R.-M.; writing—original draft preparation, C.R.S.; writing—review and editing, C.R.S. and C.G.-S.; funding acquisition, C.R.S. and C.G.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by DGAPA-PAPIIT-UNAM grants IV100520 and IA203721. The APC was funded by DGAPA-PAPIIT-UNAM grant IV100520.

Data Availability Statement: The datasets analysed during the current study are available in public repositories. The COVID-19 data are from the General Directorate of Epidemiology repository, [https://www.gob.mx/salud/documentos/datos-abiertos-152127?idiom=es, accessed on 19 March 2021]. The climate data are from Worldclim repository [www.worldclim.org, accessed on 22 March 2021]. The sociodemographic data are from the repository of the National Institute of Statistics and Geography of Mexico [https://www.inegi.org.mx/programas/ccpv/2010, accessed on 25 March 2021]. The atmospheric data are from the Institutional Repository of Geospatial Scientific Data of the ICAyCC-UNAM [https://uniatmos.atmosfera.unam.mx/, accessed on 15 April 2020]. These data can be freely downloaded and do not require a licence. The mobility data are available from the corresponding author on reasonable request.

Acknowledgments: The EpI-PUMA project has benefited from the collaboration of multiple researchers, students, and software developers.

Conflicts of Interest: The authors declare no conflict of interest.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

References

1. Araujo, M.B.; Naimi, B. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. *MedRxiv* **2020**. [CrossRef]
2. Carlson, C.J.; Chipperfield, J.D.; Benito, B.M.; Telford, R.J.; O'Hara, R.B. Species distribution models are inappropriate for COVID-19. *Nat. Ecol. Evol.* **2020**, *4*, 770–771. [CrossRef] [PubMed]
3. Araújo, M.B.; Mestre, F.; Naimi, B. Ecological and epidemiological models are both useful for SARS-CoV-2. *Nat. Ecol. Evol.* **2020**, *4*, 1153–1154. [CrossRef] [PubMed]
4. Chipperfield, J.D.; Benito, B.M.; O'Hara, R.B.; Telford, R.J.; Carlson, C.J. On the inadequacy of species distribution models for modelling the spread of SARS-CoV-2: Response to Araújo and Naimi. *EcoEvoRxiv* **2020**. [CrossRef]
5. Coro, G. A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate. *Ecol. Model.* **2020**, *431*, 109187. [CrossRef]
6. Contina, A.; Yanco, S.W.; Pierce, A.K.; DePrenger-Levin, M.; Wunder, M.B.; Neophytou, A.M.; Lostroh, C.P.; Telford, R.J.; Benito, B.M.; Chipperfield, J.; et al. Comment on "A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate", author Coro. *Ecol. Model.* **2020**, *436*, 109288. [CrossRef]
7. Hutchinson, G.E. *Introduction to Population Ecology*; Yale University Press: New Haven, CT, USA, 1978.
8. Soberón, J. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* **2007**, *10*, 1115–1123. [CrossRef]
9. Alexander, K.A.; Lewis, B.L.; Marathe, M.; Eubank, S.; Blackburn, J.K. Modeling of wildlife-associated zoonoses: Applications and caveats. *Vector-Borne Zoonotic Dis.* **2012**, *12*, 1005–1018. [CrossRef]
10. Escobar, L.E.; Craft, M.E. Advances and limitations of disease biogeography using ecological niche modeling. *Front. Microbiol.* **2016**, *7*, 1174. [CrossRef]
11. Rogers, D.J. Models for vectors and vector-borne diseases. *Adv. Parasitol.* **2006**, *62*, 1–35.
12. Peterson, A.T.; Sánchez-Cordero, V.; Beard, C.B.; Ramsey, J.M. Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerg. Infect. Dis.* **2002**, *8*, 662. [CrossRef]
13. Blackburn, J.K.; McNyset, K.M.; Curtis, A.; Hugh-Jones, M.E. Modeling the geographic distribution of Bacillus anthracis, the causative agent of anthrax disease, for the contiguous United States using predictive ecologic niche modeling. *Am. J. Trop. Med. Hyg.* **2007**, *77*, 1103–1110. [CrossRef] [PubMed]
14. Stephens, C.R.; Heau, J.G.; González, C.; Ibarra-Cerdeña, C.N.; Sánchez-Cordero, V.; González-Salazar, C. Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS ONE* **2009**, *4*, e5725. [CrossRef] [PubMed]
15. Berzunza-Cruz, M.; Rodríguez-Moreno, Á.; Gutiérrez-Granados, G.; González-Salazar, C.; Stephens, C.R.; Hidalgo-Mihart, M.; Marina, C.F.; Rebollar-Téllez, E.A.; Bailón-Martínez, D.; Balcells, C.D.; et al. Leishmania (L.) mexicana infected bats in Mexico: Novel potential reservoirs. *PLoS Neglected Trop. Dis.* **2015**, *9*, e0003438. [CrossRef]

16. Stephens, C.R.; González-Salazar, C.; Sánchez-Cordero, V.; Becker, I.; Rebollar-Tellez, E.; Rodríguez-Moreno, Á.; Berzunza-Cruz, M.; Domingo Balcells, C.; Gutiérrez-Granados, G.; Hidalgo-Mihart, M.; et al. Can you judge a disease host by the company it keeps? Predicting disease hosts and their relative importance: A case study for Leishmaniasis. *PLoS Neglected Trop. Dis.* **2016**, *10*, e0005004. [CrossRef] [PubMed]
17. Rengifo-Correa, L.; Stephens, C.R.; Morrone, J.J.; Téllez-Rendón, J.L.; Gonzalez-Salazar, C. Understanding transmissibility patterns of Chagas disease through complex vector–host networks. *Parasitology* **2017**, *144*, 760–772. [CrossRef] [PubMed]
18. González-Salazar, C.; Stephens, C.R.; Sánchez-Cordero, V. Predicting the potential role of non-human hosts in Zika virus maintenance. *EcoHealth* **2017**, *14*, 171–177. [CrossRef]
19. González-Salazar, C.; Stephens, C.R.; Meneses-Mosquera, A.K. Assessment of the potential establishment of Lyme endemic cycles in Mexico. *J. Vector Ecol.* **2021**, *46*, 207–220. [CrossRef]
20. Rengifo-Correa, L.; González-Salazar, C.; Stephens, C.R. Disentangling the contributions of biotic and abiotic predictors in the niche and the species distribution model of Trypanosoma cruzi, etiological agent of Chagas disease. *Acta Trop.* **2023**, *238*, 106757. [CrossRef]
21. Araújo, M.B.; Pearson, R.G.; Rahbek, C. Equilibrium of species' distributions with climate. *Ecography* **2005**, *28*, 693–695. [CrossRef]
22. Morgenstern, H. Ecologic studies in epidemiology: Concepts, principles, and methods. *Annu. Rev. Public Health* **1995**, *16*, 61–81. [CrossRef]
23. Freedman, D.A. Ecological inference and the ecological fallacy. *Int. Encycl. Soc. Behav. Sci.* **1999**, *6*, 1–7.
24. Tu, J.V.; Ko, D.T. Ecological studies and cardiovascular outcomes research. *Circulation* **2008**, *118*, 2588–2593. [CrossRef]
25. Dogan, M.; Rokkan, S. Quantitative ecological analysis: Contexts, trends, tasks. *Soc. Sci. Inf.* **1967**, *6*, 35–47. [CrossRef]
26. Dogan, M.; Rokkan, S. *Quantitative Ecological Analysis in the Social Sciences*; Mass MIT Press: Cambridge, UK, 1969.
27. Moein, S.; Nickaeen, N.; Roointan, A.; Borhani, N.; Heidary, Z.; Javanmard, S.H.; Ghaisari, J.; Gheisari, Y. Inefficiency of SIR models in forecasting COVID-19 epidemic: A case study of Isfahan. *Sci. Rep.* **2021**, *11*, 4725. [CrossRef]
28. Zeb, A.; Alzahrani, E.; Erturk, V.S.; Zaman, G. Mathematical model for coronavirus disease 2019 (COVID-19) containing isolation class. *BioMed Res. Int.* **2020**, *2020*, 3452402. [CrossRef]
29. Zhang, Z.; Zeb, A.; Hussain, S.; Alzahrani, E. Dynamics of COVID-19 mathematical model with stochastic perturbation. *Adv. Differ. Equations* **2020**, *2020*, 451. [CrossRef] [PubMed]
30. Zhang, Z.; Zeb, A.; Egbelowo, O.F.; Erturk, V.S. Dynamics of a fractional order mathematical model for COVID-19 epidemic. *Adv. Differ. Equations* **2020**, *2020*, 420. [CrossRef]
31. Li, X.P.; Al Bayatti, H.; Din, A.; Zeb, A. A vigorous study of fractional order COVID-19 model via ABC derivatives. *Results Phys.* **2021**, *29*, 104737. [CrossRef]
32. Cramer, E.Y.; Ray, E.L.; Lopez, V.K.; Bracher, J.; Brennen, A.; Castro Rivadeneira, A.J.; Gerding, A.; Gneiting, T.; House, K.H.; Huang, Y.; et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2113561119. [CrossRef] [PubMed]
33. Zhang, Z.; Gul, R.; Zeb, A. Global sensitivity analysis of COVID-19 mathematical model. *Alex. Eng. J.* **2021**, *60*, 565–572. [CrossRef]
34. Alali, Y.; Harrou, F.; Sun, Y. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Sci. Rep.* **2022**, *12*, 2467. [CrossRef]
35. Cooper, I.; Mondal, A.; Antonopoulos, C.G. A SIR model assumption for the spread of COVID-19 in different communities. *Chaos Solitons Fractals* **2020**, *139*, 110057. [CrossRef]
36. Zeb, A.; Kumar, P.; Erturk, V.S.; Sitthiwirattham, T. A new study on two different vaccinated fractional-order COVID-19 models via numerical algorithms. *J. King Saud Univ.-Sci.* **2022**, *34*, 101914. [CrossRef] [PubMed]
37. Peterson, A.; Soberón, J.; Pearson, R.; Anderson, R.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M. *Ecological Niches and Geographic Distributions*; Princeton University Press: Princeton, NJ, USA, 2011.
38. Stephens, C.R.; Sierra-Alcocer, R.; González-Salazar, C.; Barrios, J.M.; Salazar Carrillo, J.C.; Robredo Ezquivelzeta, E.; del Callejo Canal, E. SPECIES: A platform for the exploration of ecological data. *Ecol. Evol.* **2019**, *9*, 1638–1653. [CrossRef]
39. Gotelli, N.J. Null model analysis of species co-occurrence patterns. *Ecology* **2000**, *81*, 2606–2621. [CrossRef]
40. Stephens, C.R.; González-Salazar, C.; Villalobos, M.; Marquet, P. Can Ecological Interactions be Inferred from Spatial Data? *Biodivers. Inform.* **2020**, *15*, 11–54. [CrossRef]
41. Stephens, C.R.; Huerta, H.F.; Linares, A.R. When is the Naive Bayes approximation not so naive? *Mach. Learn.* **2018**, *107*, 397–441. [CrossRef]
42. Anderson, R.P. A framework for using niche models to estimate impacts of climate change on species distributions. *Ann. N. Y. Acad. Sci.* **2013**, *1297*, 8–28. [CrossRef] [PubMed]
43. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
44. Fielding, A.H.; Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **1997**, *24*, 38–49. [CrossRef]
45. Hill, A.B. The environment and disease: Association or causation? *J. R. Soc. Med.* **1965**, *108*, 32–37. [CrossRef] [PubMed]
46. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **1974**, *66*, 688. [CrossRef]
47. Rubin, D.B. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* **1978**, *6*, 34–58. [CrossRef]

48. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
49. Stephens, C.R.; Sánchez-Cordero, V.; González Salazar, C. Bayesian inference of ecological interactions from spatial data. *Entropy* **2017**, *19*, 547. [CrossRef]
50. SSA-DGE. Datos Abiertos Dirección Dirección General de Epidemiología, Secretaria de Salud, Gobierno de Mexico. Available online: https://www.gob.mx/salud/documentos/datos-abiertos-152127?idiom=es (accessed on 19 March 2021).
51. INEGI. Censo de Población y Vivienda 2010; Instituto Nacional de Estadística y Geografía. Available online: https://www.inegi.org.mx/programas/ccpv/2010 (accessed on 25 March 2021)
52. Rivera, C.; Stremme, W.; Gruter de la Mora, M.; Fernández, E.A.; Elizarras, R.L.G.; Castelán, H. Databases, metadata and interactive cartographic visualizations of Formaldehyde total column (HCHO in units of molecules/cm^2) measured by the sensor: OMI (Ozone Monitoring Instrument) on board the Aura satellite. In *Institutional Repository of Geospatial Scientific Data of the ICAyCC, UNAM. Computing Unit for Atmospheric and Environmental Sciences;* Institute of Atmospheric Sciences and Climate Change, UNAM: Mexico City, Mexico, 2019.
53. Rivera, C.; Stremme, W.; Gruter de la Mora, M.; Fernández, E.A.; Elizarras, R.L.G.; Castelán, H. Databases, metadata and interactive cartographic visualizations of Nitrogen dioxide tropospheric columns (NO_2 in units of molecules/cm^2) measured by the sensor: OMI (Ozone Monitoring Instrument) on board the Aura satellite. In *Institutional Repository of Geospatial Scientific Data of the ICAyCC, UNAM. Computing Unit for Atmospheric and Environmental Sciences*; Institute of Atmospheric Sciences and Climate Change, UNAM: Mexico City, Mexico, 2019.
54. Rivera, C.; Stremme, W.; Gruter de la Mora, M.; Fernández, E.A.; Elizarras, R.L.G.; Castelán, H. Databases, metadata and interactive cartographic visualizations of Columns in the Planetary Boundary Layer of SO_2 (Sulfur dioxide in units of molecules/cm^2) measured by the sensor: OMI (Ozone Monitoring Instrument) on board the Aura satellite. In *Institutional Repository of Geospatial Scientific Data of the ICAyCC, UNAM. Computing Unit for Atmospheric and Environmental Sciences*; Institute of Atmospheric Sciences and Climate Change, UNAM: Mexico City, Mexico, 2019.
55. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
56. Martínez-Domínguez, M.; Mora-Rivera, J. Internet adoption and usage patterns in rural Mexico. *Technol. Soc.* **2020**, *60*, 101226. [CrossRef]
57. Xu, Y.; Belyi, A.; Bojic, I.; Ratti, C. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Comput. Environ. Urban Syst.* **2018**, *72*, 51–67. [CrossRef]
58. Aromí, J.D.; Bonel, M.P.; Cristia, J.; Llada, M.; Palomino, L. *Socioeconomic Status and Mobility during the COVID-19 Pandemic: An Analysis of Eight Large Latin American Cities*; Technical Report, IDB Working Paper Series; Inter-American Development Bank (IDB): Washington, DC, USA, 2021.
59. Hutchinson, G. Concluding remarks cold spring harbor symposia on quantitative biology. *GS SEARCH* **1957**, *22*, 415–427.
60. Adam, D. A guide to R–the pandemic's misunderstood metric. *Nature* **2020**, *583*, 346–349. [CrossRef] [PubMed]
61. Holt, R.D. Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19659–19665. [CrossRef] [PubMed]