Machine Learning for Biodiversity Scientists

An opinionated primer

Timothée Poisot

2024-11-22

Table of contents

List of Figures

Preface

Machine learning is now an established methodology to study biodiversity, and this is a problem.

This may be an opportunity when it comes to advancing our knowledge of biodiversity, and in particular when it comes to translating this knowledge into action (Tuia *et al.* 2022); but make no mistake, this is a problem for us, biodiversity scientists, as we suddenly need to develop competences in an entirely new field in order to remain professionally relevant (Ellwood *et al.* 2019). And as luck would have it, there are easier fields to master than machine learning. The point of this book, therefore, is to provide an introduction to fundamental concepts in data science, from the perspective of a biodiversity scientist, by using examples corresponding to real-world use-cases of these techniques.

But what do we mean by *machine learning* and *data science*? Most science, after all, relies on data in some capacity. What falls under the umbrella of *data science* is, in short, embracing in equal measure quantitative skills (mathematics, machine learning, statistics), programming, and domain expertise, in order to solve well-defined problems. *Machine learning* is a series of techniques (or, more precisely, a high-level approach to these techniques) through which we conduct our data science activities. A core tenet of data science is that, when using it, we seek to "deliver actionable insights", which is MBA-speak for "figuring out what to do next". One of the ways in which this occurs is by letting the data speak, after they have been, of course, properly cleaned and transformed and engineered. This entire process is driven by (or, even, subject to) domain knowledge. There is no such thing as data science, at least not in a vacuum: there is data science as a methodology applied to a specific domain.

Think of data science as being its own epistemology (Desai *et al.* 2022), and machine learning as one methodology we can apply to work within this context.

Before we embark into a journey of discovery on the applications of data science to biodiversity, allow me to let you in on a little secret: data *science* is a little bit of a misnomer. In order to understand why, I need (or at least, I really want) to talk about cooking.

To become a good cook, there are general techniques one *must* master, which we apply to specific steps in recipes; these recipes draw from a common cultural or local repertoire and cultural specifics (but the evolution of recipes is remarkably convergent – most cuisines have a *mirepoix*, bread, and beer). Finally, there is the product, *i.e.* the unique dish that you have cooked. And so it is for data science too: we can abstract a series of processes and guidelines, think about their application within the context of our specific field, study system, or line and research, and all of this will shape the final data product we can serve.

When writing this preface, I turned to my shelf of cookbooks, and picked my two favorites: Robuchon's *The Complete Robuchon* (a no-nonsense list of hundreds of recipes with no place for improvisation), and Bianco's *Pizza, Pasta, and Other Food I Like* (a short volume with very few pizza and pasta, and wonderful discussions about the importance of humility, creativity, and generosity). Data science, if it were cooking, would feel a lot like the second. Deviation from the rules is often justifiable if you feel like it. But this improvisation requires good skills, a clear mental map of the problem, a defined vision of what these deviations will let you achieve, and a library of patterns that you can draw from.

This book will not get you here. But it will speed up the process, by framing the practice of data science as a natural way to conduct research on biodiversity.

1. Introduction

This book started as a collection of notes from several classes I taught in the Department of Biological Sciences at the Université de Montréal, as well as a few workshops I ran for the Québec Centre for Biodiversity Sciences. When teaching data synthesis, data science, and machine learning to biology students, I realized that the field was missing resources that could serve as stepping stones to proficiency.

There are excellent manuals covering the mathematics of data science and machine learning (I will list a few later on). These are important to read, because the field of machine learning is an offshoot of mathematics and computer science, and it is important to become familiar with the core concepts. A little bit of calculus and a whole lot of linear algebra should be more of the same for many ecologists. But these resources are usually less useful as practical guides to the field.

There are many good papers giving overviews of some applications of data science to biological problems (a lot of them are cited in this book). These are important to read, because any attempt to adopt a new methodology (new to us, not new to the field, or new in absolute terms!) must proceed alongside some familiarity of how it has been used by our colleagues. But these articles, although good at showing how these tools are actually used, usually make it difficult to establish more general recommendations.

There are, finally, thousands of tutorials about how to write code to perform any machine learning algorithm you can think of. Some of them are even good. But these tutorials usually suffer (in our case) from being disconnected from the field of biodiversity science, and of course are limited by the language they use, the version of the packages they ran with, and again do not allow for much generalization.

When navigating these resources, one thing that students commonly called for was an attempt to tie concepts together, and to explain when and how human decisions were required in ML approaches (Sulmont *et al.* 2019).

This is particularly true of students with strong domain knowledge that want to understand how machine learning fits with their ability to do research.

This is book is this attempt.

There are, broadly speaking, two situations in which reading this book is useful. The first is when you are done reading some general books about machine learning, and want to see how it can be applied to problems that are more specific to biodiversity research; the second is when you have a working understanding of biodiversity research, and want a stepping stone into the machine learning literature. Note that there is no scenario where you *stop* after reading this book – this is by design. The purpose of this book is to give a practical overview of "how data science for biodiversity happens", and this needs to be done in parallel to even more fundamental readings.

These are examples of books I like. I found them comprehensive and engaging. They may not work for you.

A wonderful introduction to the mathematics behind machine learning can be found in Deisenroth *et al.* (2020), which provides stunning visualization of mathematical concepts. Yau (2015) is a particularly useful book about the ways to visualize data in a meaningful way. Watt *et al.* (2020) is a solid introduction to the underlying theory of applied machine learning. For ecologists, Dietze (2017) is a comprehensive, and still highly readable, treaty on the problems associated to forecasting. The best way to decide on which book to read is often to look at the books that your colleagues have also read; being able to work through material collectively is useful, and knowing that you can practice the craft of data science within a community will make your learning more effective.

When reading this book, I encourage you to read the chapters in order. They have been designed to be read in order, because each chapter introduces the least possible amount of new concepts, but often requires to build on the previous chapters. This is particularly true of the second half of this book.

1.1. Core concepts in data science

1.1.1. EDA

1.1.2. Clustering and regression

1.1.3. Supervised and unsupervised

1.1.4. Training, testing, and validation

1.1.5. Transformations and feature engineering

1.2. An overview of the content

In **?@sec-clustering**, we introduce some fundamental questions in data science, by working on the clustering of pixels in Landsat data. The point of this chapter is to question the way we think about data, and to start a discussion about an "optimal" model, hyper-parameters, and what a "good" model is.

In **?@sec-gradientdescent**, we revisit well-trodden statistical ground, by fitting a linear model to linear data, but uisng gradient descent. This provides us with an opportunity to think about what a "fitted" model is, whether it is possible to learn too much from data, and why being able to think about predictions in the unit of our problem helps.

In **?@sec-crossvalidation**, we start introducing one of the most important bit element of data science practice, in the form of cross-validation. We apply this technique to the prediction of plant phenology over a millenia, and think about the central question of "what kind of decision-making can we justify with a model".

In **?@sec-classification**, we introduce the task of classification, and spend a lot of time thinking about biases in predictions, which are acceptable, and which are not. We start building a model for the distribution of the Reindeer, which we will improve over a few chapters.

In **?@sec-predictors**, we explore ways to perform variable selection, think of this task as being part of the training process, and introduce ideas related to dimensionality reduction. In **?@sec-leakage**, we discuss data leakage, where it comes from, and how to prevent it. This leads us to introducing the concept of data transformations as a model, which will establish some best practices we will keep on using throughout this book.

In **?@sec-tuning**, we conclude story arcs that had been initiated in a few previous chapters, and explore training curves, the tuning of hyper-parameters, and moving-threshold classification. We provide the final refinements to out model of the Reindeer distribution.

In Chapter 2, we will shift our attention from prediction to understanding, and explore techniques to quantify the importance of variables, as well as ways to visualize their contribution to the predictions. In doing so, we will introduce concepts of model interpretation and explainability.

In **?@sec-bagging**, ...

1.3. A note on colors

Type	Meaning	Color
All	generic	
	no data	
Cross-validation	training	
	validation	

Type	Meaning	Color
Species range	testing	
	presence	
	absence	
Range change	loss	
	no change	
	gain	

In addition, there are three important color *palettes*. Information that is *sequential* is nature, which is to say it increases on a continuous scale without a logical midpoint, is rendered with these colors (from low to the left, to high values to the right):

The diverging palette is used for values that have a clear midpoint (usually values centered on 0). The midpoint will always correspond to the central color, and this palette is symmetrical:

Finally, the categorical data are represented using the following palette:

1.4. Some rules about this book

When I started aggregating these notes, I decided on a series of four rules. No code, no simulated data, no long list of model, and above all, no *iris* dataset. In this section, I will go through *why* I decided to adopt these rules, and how it should change the way you interact with the book.

1.4.1. No code

This is, maybe, the most surprising rule, because data science *is* programming (in a sense). But sometimes there is so much focus on programming that we lose track of the other, important aspects of the practice of data science: abstractions, relationship with data, and domain knowledge.

This book *did* involve a lot of code. Specifically, this book was written using *Julia* (Bezanson *et al.* 2017), and every figure is generated by a notebook, and they are part of the material I use when teaching from this content in the classroom. But code is *not* a universal language, and unless you are really familiar with the language, code can obfuscate. I had no intention to write a *Julia* book (or an *R* book, or a *Python* book). The point is to

think about data science applied to ecological research, and I felt like it would be more inclusive to do this in a language agnostic way.

And finally, code rots. Code with more dependencies rots faster. It take a single change in the API of a package to break the examples, and then you are left with a very expensive monitor stand. With a few exceptions, the examples in this book do not use complicated packages either.

1.4.2. No simulated data

I have nothing against simulated data. I have, in fact, generated simulated data in many different contexts, for training or for research. But the limit of simulated is that we almost inevitably fail to include what makes real data challenging: noise, incomplete or uneven sampling, data representation artifacts. And so when it is time to work on real data, everything seems suddenly more difficult.

Simulated data have *immense* training value; but it is also important to engage with the imperfect actual data, as we will overwhelmingly apply the concepts from this book to them. For this reason, there are no simulated data in this book. Everything that is presented correspond to an actual use case that proceeds from a question we could reasonably ask in the context, paired with a dataset that could be used to answer this question.

1.4.3. No model zoo

My favorite machine learning package is *MLJ* (Blaom *et al.* 2020). When given a table of labels and a table of features, it will give back a series of models that match with these data. It speeds up the discovery of models considerably, and is generally a lot more informative than trying to read from a list of possible techniques. If I have questions about an algorithm from this list, I can start reading more documentation about how it works.

Reading a long enumeration of things is boring; unless it's sung by Yakko Warner, I'm not interested, and I refuse to inflict it on people. But more importantly, these enumerations of models often distract from thinking about the problem we want to solve in more abstract terms. I rarely wake up in the morning and think "oh boy I can't wait to train a SVM today"; chances are, my thought process will be closer to "I need to tell the mushroom people where I think the next good foraging locations will be". The rest, is implementation details.

In fact, 90% of this book uses only two models: linear regression, and the Naïve Bayes Classifier. Some other models are involved in a few chapters, but these two models are breathtakingly simple, work surprisingly well, run fast, and can be tweaked to allow us to build deep intuitions about how machines learn. They are perfect for the classroom, and give us the freedom to spent most of our time thinking about how we interact with models, and why, and how we make methodological decisions.

1.4.4. No *iris* dataset

From a teaching point of view, the *iris* dataset is like hearing Smash Mouth in a movie trailer, in that it tells you two things with absolute certainty. First, that you are indeed watching a movie trailer. Second, that you could be watching Shrek instead. There are datasets out there that are *infinitely more* exciting to use than *iris*.

But there is a far more important reason not to use *iris*: eugenics.

Listen, we made it several hundred words in a text about quantitative techniques in life sciences without encountering a sad little man with racist ideas that academia decided to ignore because "he just contributed so much to the field, and these were different times, maybe we shouldn't be so quick to judge?". Ronald Aylmer Fisher, statistics' most racist nerd, was such a man; and there are, of course, those who want to consider the possibility that you can be outrageously racist as long as you are an outstanding scientist (Bodmer *et al.* 2021).

The *iris* dataset was first published by Fisher (1936) in the *Annals of Eugenics* (so, there's a bit of a red flag there already), and draws from several publications by Edgar Anderson, starting with Anderson (1928); Unwin & Kleinman (2021) have an interesting historiographic deep-dive into the correspondence between the two. Judging by the dates, you may think that Fisher was a product of his time. But this could not be further from the truth. Fisher was dissatisfied with his time, to the point where his contributions to statistics were done in service of his views, in order to provide the appearance of scientific rigor to his bigotry.

Fisher advocated for forced sterilization for the "defectives" (which he estimated at, oh, roughly 10% of the population), argued that not all races had equal capacity for intellectual and emotional development, and held a host of related opinions. There is no amount of contribution to science that pardon these views. Coming up with the idea of the null hypothesis does not even out lending "scientific" credibility to ideas whose logical

(and historical) conclusion is genocide. That Ronald Fisher is still described as a polymath and a genius is infuriating, and we should use every alternative to his work that we have.

Thankfully, there are alternatives!

The most broadly known alternative to the *iris* dataset is `penguins`, which was collected by ecologists (Gorman *et al.* 2014), and published as a standard dataset (Horst *et al.* 2020) so that we can train students without engaging with the "legacy" of eugenicists. The `penguins` dataset is also genuinely good! The classes are not so obviously separable, there are some missing data that reflect the reality of field work, and the data about sex and spatial location have been preserved, which increases the diversity of questions we can ask. We won't use `penguins` either. It's a fine dataset, but at this point there is little that we can write around it that would be new, or exciting. But if you want to apply some of the techniques in this book? Go `penguins`.

References

Anderson, E. (1928). The problem of species in the northern blue flags, iris versicolor l. And iris virginica l. *Annals of the Missouri Botanical Garden*, 15, 241.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59, 65–98.

Blaom, A., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D. & Vollmer, S. (2020). MLJ: A julia package for composable machine learning. *Journal of Open Source Software*, 5, 2704.

Bodmer, W., Bailey, R.A., Charlesworth, B., Eyre-Walker, A., Farewell, V., Mead, A., *et al.* (2021). The outstanding scientist, R.A. Fisher: his views on eugenics and race. *Heredity*, 126, 565–576.

Deisenroth, M.P., Faisal, A.A. & Ong, C.S. (2020). Mathematics for machine learning.

Dietze, M. (2017). Ecological forecasting.

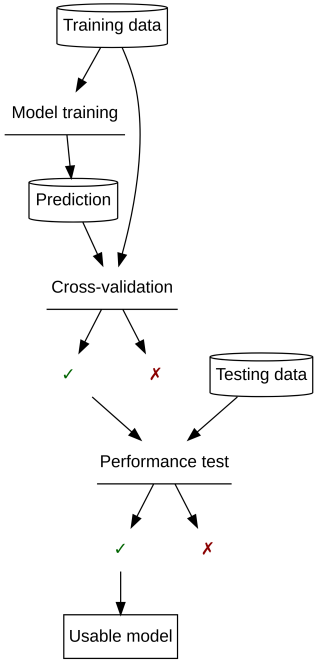
Fisher, R.A. (1936). The Use Of Multiple Measurements In Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.

Gorman, K.B., Williams, T.D. & Fraser, W.R. (2014). Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). *PLoS ONE*, 9, e90081.

Horst, A.M., Hill, A.P. & Gorman, K.B. (2020). *Allisonhorst/palmerpenguins: v0.1.0*. Zenodo.

Sulmont, E., Patitsas, E. & Cooperstock, J.R. (2019). Can you teach me to machine learn? *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*.

Unwin, A. & Kleinman, K. (2021). The Iris Data Set: In Search of the Source of *Virginica*. *Significance*, 18, 26–29.
Watt, J., Borhani, R. & Katsaggelos, A. (2020). Machine learning refined.
Yau, N. (2015). Visualize this.



Flowchart 1.1: An overview of the process of coming up with a usable model. The process of creating a model starts with a trainig dataset made of predictors and responses, which is used to train a model. This model is cross-validated on its training data, to estimate whether it can be fully retrained. The fully trained model is that applied to an independent testing dataset, and the evaluation of the performance determines whether it will be used.

2. Explaining predictions

In this chapter, we will

navigate the accuracy-explainability for public policy Bell *et al.* (2022)

what is explainable differs between stakeholders Amarasinghe *et al.* (2023)

biodiversity need sustained model uptake Weiskopf *et al.* (2022)

2.0.1. Partial responses

values of variable against mean of all others

2.0.2. Inflated partial responses

sample background variables

still a measure of global model response because the values are kept but the structure is lost

gives a better sense of potentially divergent responses

2.0.3. Shapley values

LOCAL (prediction-scale) importance

Štrumbelj & Kononenko (2013) monte carlo approximation of shapley values

Wadoux *et al.* (2023) mapping of shapley values

Mesgaran *et al.* (2014) mapping of most important covariates

Lundberg & Lee (2017) SHAP

important properties + interpretation

2.0.4. Importance of transfo as part of model

transfo in model = we can still apply these techniques instead of asking "what does PC1 = 0.4 mean"

2.1. Application

2.1.1. Partial responses

2.1.2. Shapley values

```
S = zeros(Float64, (length(variables(model)), length(labels(model))))
for (vidx, vpos) in enumerate(variables(model))
    S[vidx,:] = explain(model, vpos; threshold=false, samples=200)
end
P = features(model, variables(model))
```

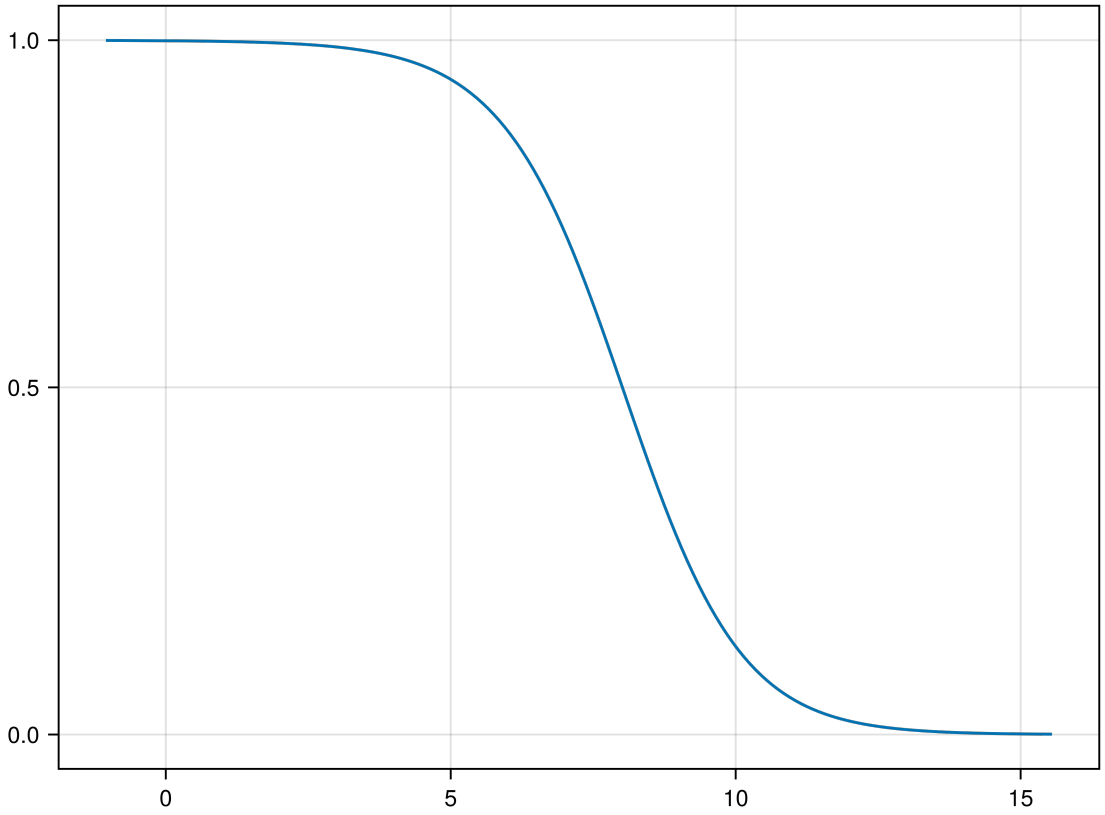


Figure 2.1.: TODO

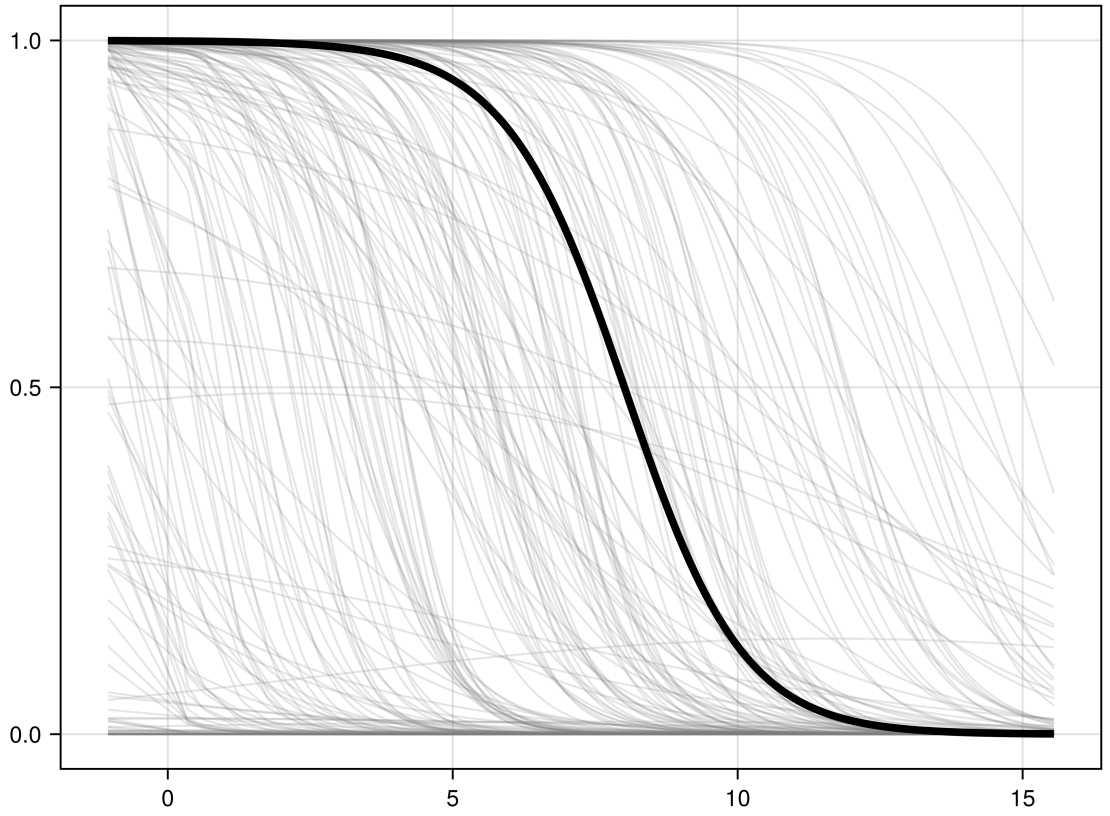


Figure 2.2.: TODO

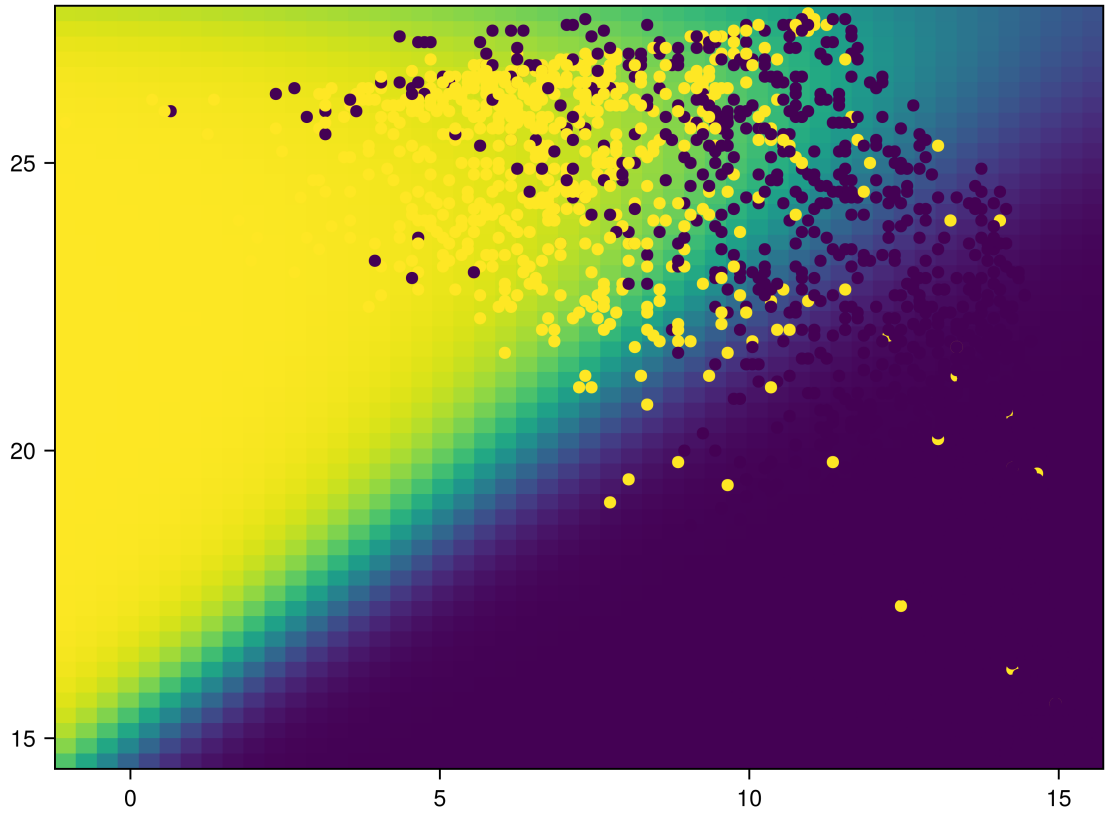


Figure 2.3.: TODO

```
4×2230 Matrix{Float64}:
  6.15002   8.75   8.45001  13.75  …  13.85  14.65   9.25  14.05  13.85
 26.3      22.1   25.2      19.6       21.7   18.6   26.4   18.8   19.3
 42.2      40.7   44.3      45.0       49.6   44.9   47.1   43.6   49.6
 23.15     22.95  24.85     26.45      28.15  26.85  26.35  26.35  26.35
```

TODO redraw the stemplot from the variable selection chapter to compare prediction v. explanation

Table 2.1.: blah blah blah

Variable	**Imp. (Shapley)**	**Imp. (bootstrap)**	**Min.**	**Med.**	**Max.**
BIO 8	39.59%	35.25%	-0.29	-0.09	0.56
BIO 7	33.74%	25.45%	-0.40	-0.05	0.30
BIO 15	14.91%	8.88%	-0.38	0.00	0.18
BIO 5	11.76%	6.61%	-0.24	-0.01	0.20

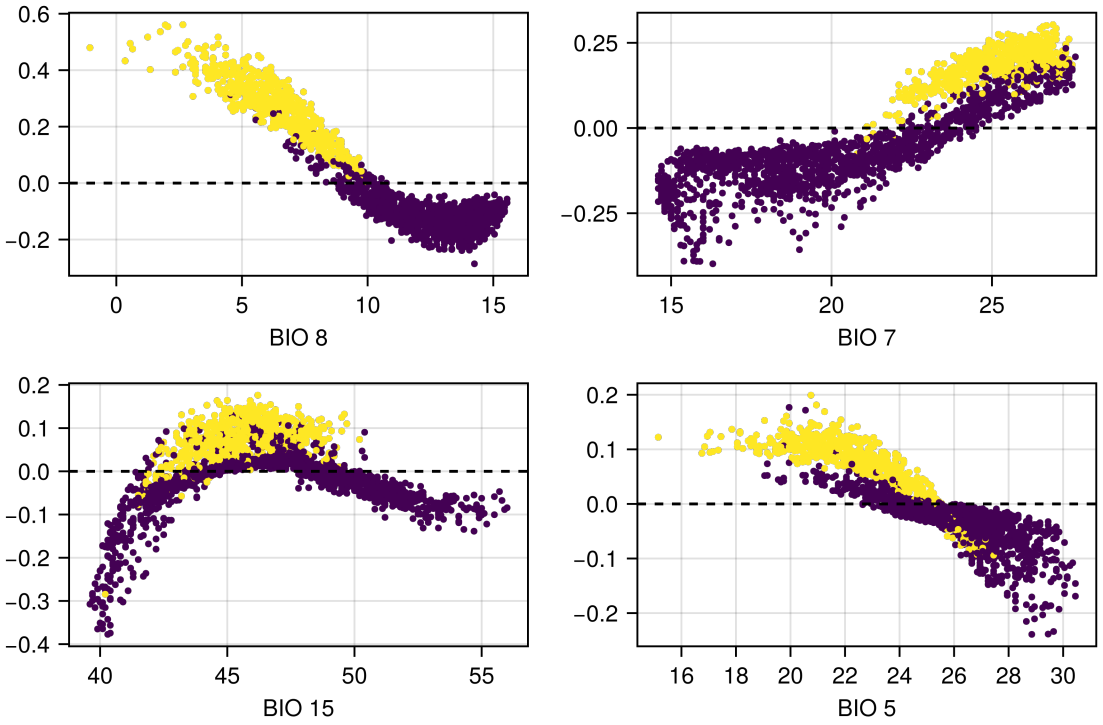
```
f = Figure(; size=(600, 400))
args = (color=predict(model), markersize=5, colorrange=(0., 1.))

ax1 = Axis(f[1,1]; xlabel="BIO $(model.v[varord[1]])")
scatter!(ax1, P[varord[1],:], S[varord[1],:]; args...)
ax2 = Axis(f[1,2]; xlabel="BIO $(model.v[varord[2]])")
scatter!(ax2, P[varord[2],:], S[varord[2],:]; args...)
ax3 = Axis(f[2,1]; xlabel="BIO $(model.v[varord[3]])")
scatter!(ax3, P[varord[3],:], S[varord[3],:]; args...)
ax4 = Axis(f[2,2]; xlabel="BIO $(model.v[varord[4]])")
scatter!(ax4, P[varord[4],:], S[varord[4],:]; args...)

xmin, xmax = extrema(S)
for ax in [ax1, ax2, ax3, ax4]
```

```
    hlines!(ax, [0.0], color=:black, linestyle=:dash)
end

current_figure()
```



2.1.3. Spatial partial effects

```
_layer_path = joinpath(dirname(Base.active_project()), "data", "occurrences", "lay
bio = [SimpleSDMLayers._read_geotiff(_layer_path; bandnumber=i) for i in 1:19]
```

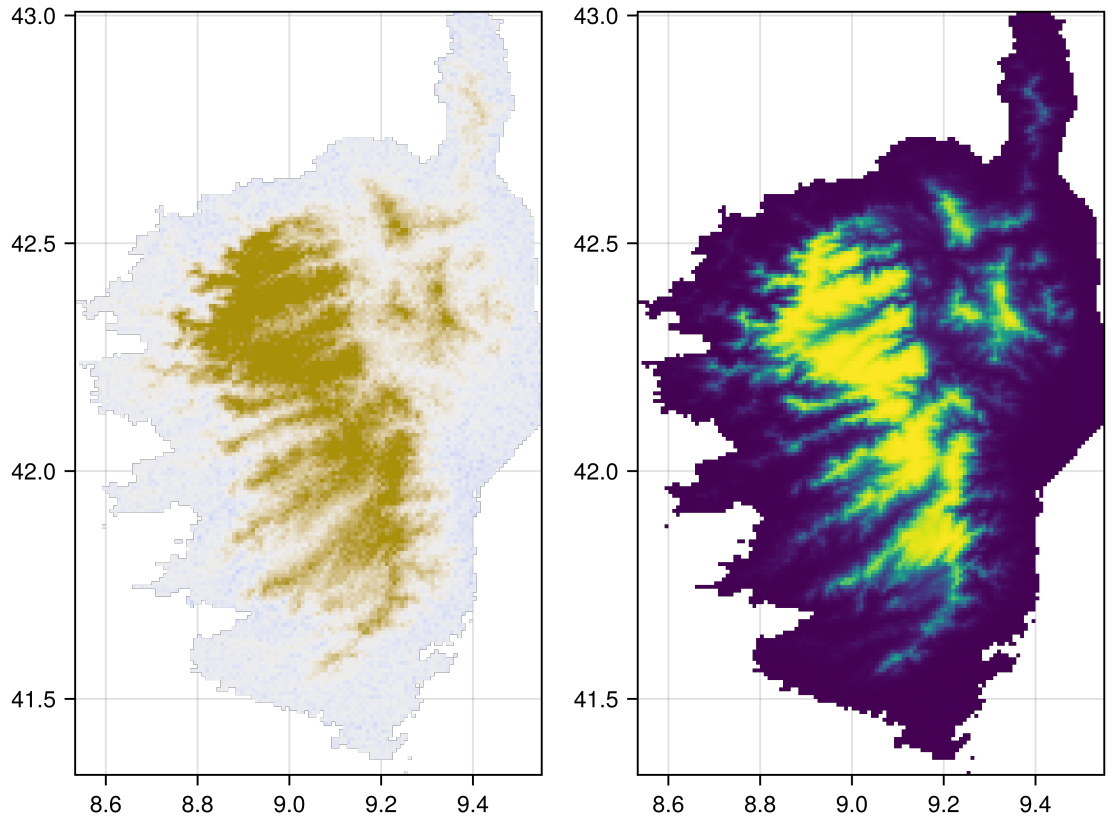
```
19-element Vector{SDMLayer{Float32}}:
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
 SDMLayer{Float32}(Float32[Inf Inf … Inf Inf; Inf Inf … Inf Inf; … ; Inf Inf … Inf Inf;
```

```
V = explain(model, bio; threshold=false, samples=30)
```

```
4-element Vector{SDMLayer{Float64}}:
```

```
SDMLayer{Float64}([0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0; … ; 0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0], Bool[0 0 .
SDMLayer{Float64}([0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0; … ; 0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0], Bool[0 0 .
SDMLayer{Float64}([0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0; … ; 0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0], Bool[0 0 .
SDMLayer{Float64}([0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0; … ; 0.0 0.0 … 0.0 0.0; 0.0 0.0 … 0.0 0.0], Bool[0 0 .
```

```
f = Figure()
a1 = Axis(f[1,1])
a2 = Axis(f[1,2])
heatmap!(a1, V[varord[1]], colormap=bkcol.div, colorrange=(-0.5,0.5))
heatmap!(a2, partialresponse(model, bio, variables(model)[varord[1]]; threshold=false))
current_figure()
```



2.1.4. Most important variable locally

2.2. Conclusion

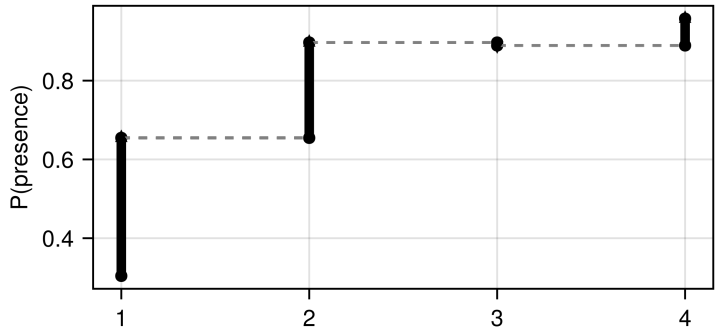


Figure 2.4.: Effect of each variable (sorted by importance as in Table 2.1) on the change of the score for a single prediction. Recall that this is expressed as the change from the *average* prediction made by the model.

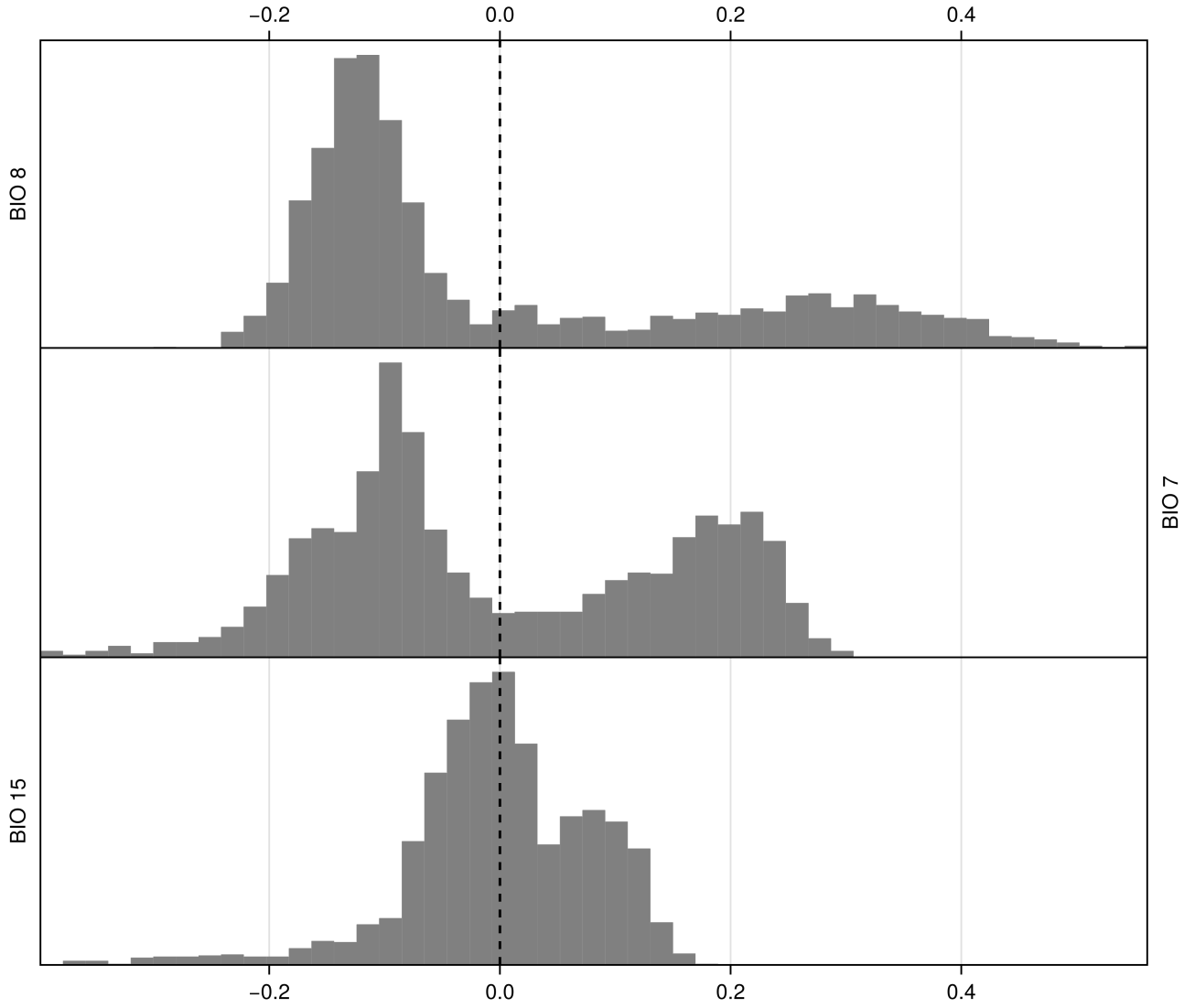


Figure 2.5.: Effect of each variable (sorted by importance as in Table 2.1) on the change of the score for a single prediction. Recall that this is expressed as the change from the *average* prediction made by the model.

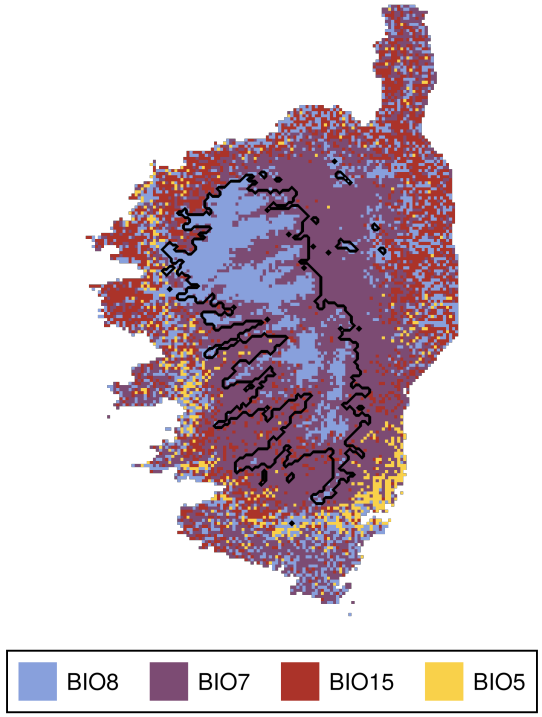


Figure 2.6.: TODO

References

Amarasinghe, K., Rodolfa, K.T., Lamba, H. & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5.

Bell, A., Solano-Kamaiko, I., Nov, O. & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Lundberg, S.M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: *Advances in neural information processing systems* (eds. Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., et al.). Curran Associates, Inc.

Mesgaran, M.B., Cousens, R.D. & Webber, B.L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, 20, 1147–1159.

Štrumbelj, E. & Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.

Wadoux, A.M.J.-C., Saby, N.P.A. & Martin, M.P. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*, 9, 21–38.

Weiskopf, S.R., Harmáčková, Z.V., Johnson, C.G., Londoño-Murcia, M.C., Miller, B.W., Myers, B.J.E., *et al.* (2022). Increasing the uptake of ecological model results in policy decisions to improve biodiversity outcomes. *Environmental Modelling & Software*, 149, 105318.

A. Instructor notes

References

Amarasinghe, K., Rodolfa, K.T., Lamba, H. & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5.

Anderson, E. (1928). The problem of species in the northern blue flags, iris versicolor l. And iris virginica l. *Annals of the Missouri Botanical Garden*, 15, 241.

Bell, A., Solano-Kamaiko, I., Nov, O. & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59, 65–98.

Blaom, A., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D. & Vollmer, S. (2020). MLJ: A julia package for composable machine learning. *Journal of Open Source Software*, 5, 2704.

Bodmer, W., Bailey, R.A., Charlesworth, B., Eyre-Walker, A., Farewell, V., Mead, A., *et al.* (2021). The outstanding scientist, R.A. Fisher: his views on eugenics and race. *Heredity*, 126, 565–576.

Deisenroth, M.P., Faisal, A.A. & Ong, C.S. (2020). Mathematics for machine learning.

Desai, J., Watson, D., Wang, V., Taddeo, M. & Floridi, L. (2022). The epistemological foundations of data science: a critical review. *Synthese*, 200.

Dietze, M. (2017). Ecological forecasting.

Ellwood, E.R., Sessa, J.A., Abraham, J.K., Budden, A.E., Douglas, N., Guralnick, R., *et al.* (2019). Biodiversity Science and the Twenty-First Century Workforce. *BioScience*, 70, 119–121.

Fisher, R.A. (1936). The Use Of Multiple Measurements In Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.

Gorman, K.B., Williams, T.D. & Fraser, W.R. (2014). Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). *PLoS ONE*, 9, e90081.

Horst, A.M., Hill, A.P. & Gorman, K.B. (2020). *Allisonhorst/palmerpenguins: v0.1.0*. Zenodo.

Lundberg, S.M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: *Advances in neural information processing systems* (eds. Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., et al.). Curran Associates, Inc.

Mesgaran, M.B., Cousens, R.D. & Webber, B.L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, 20, 1147–1159.

Štrumbelj, E. & Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.

Sulmont, E., Patitsas, E. & Cooperstock, J.R. (2019). Can you teach me to machine learn? *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., *et al.* (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13, 792.

Unwin, A. & Kleinman, K. (2021). The Iris Data Set: In Search of the Source of *Virginica*. *Significance*, 18, 26–29.

Wadoux, A.M.J.-C., Saby, N.P.A. & Martin, M.P. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*, 9, 21–38.

Watt, J., Borhani, R. & Katsaggelos, A. (2020). Machine learning refined.

Weiskopf, S.R., Harmáčková, Z.V., Johnson, C.G., Londoño-Murcia, M.C., Miller, B.W., Myers, B.J.E., *et al.* (2022). Increasing the uptake of ecological model results in policy decisions to improve biodiversity outcomes. *Environmental Modelling & Software*, 149, 105318.

Yau, N. (2015). Visualize this.