Fundamentals of Biodiversity Data Science

Timothée Poisot

2023-08-05

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

Data science is now an established methodology to study biodiversity, and this is a problem. Well, this is an opportunity when it comes to advancing our knowledge of biodiversity (Tuia et al. 2022), but this is a problem for us, biodiversity scientists, as we suddenly need to develop competences in an entirely new field. And as luck would have it, there are easier fields to master than data science.

But what do we mean by *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. 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 beyond recognition.

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. Science is (or so we like to say) neutral, systematic, and rigorous. Science is baking. Data science? It's cooking. There might be a recipe, but it's a recommendation at best,and after all you know better than to follow a list of instructions, don't you? Data science is craft. It's art. Data vibes.

2 Creating groups: the *k*-means algorithm

As we mentioned in the introduction, a core idea of data science is that things that look the same (in that, when described with data, they resemble one another) are likely to be the same. Although this sounds like a simplifying assumption, this can provide the basis for a very powerful technique in which we *create* groups in data that have no labels. This task is called unsupervised clustering: we seek to add a *label* to each observation, in order to form groups, and the data we work from do *not* have a label that we can use to train a model.

2.1 A digression: which birds are red?

Before diving in, it is a good idea to ponder a simple case. We can divide everything in just two categories: things with red feathers, and things without red feathers. An example of a thing with red feathers is the Northern Cardinal (*Cardinalis cardinalis*), and an example of things without red feathers are the iMac G3, Haydn's string quartets, and of course the Northern Cardinal (*Cardinalis cardinalis*).

See, biodiversity data science is complicated, because it tends to rely on the assumption that we can categorize the natural world, and the natural world (mostly in response to natural selection) comes up with ways to be, well, diverse. In the Northern Cardinal, this is shown in males having red feathers, and females having mostly brown feathers. Before moving

forward, we need to consider ways to solve this issue, as this issue will come up *all the time.*

The first mistake we have made is that the scope of objects we want to classify, which we will describe as the "domain" of our classification, is much too broad: there are few legitimate applications where we will have a dataset with Northern Cardinals, iMac G3s, and Haydn's string quartets. Picking a reasonable universe of classes would have solved our problem a little. For example, among the things that do not have red feathers are the Mourning Dove, the Kentucky Warbler, and the House Sparrow.

The second mistake that we have made is improperly defining our classes; bird species exhibit sexual dimorphism (not in an interesting way, like wrasses, but you let's still give them some credit for trying). Assuming that there is such a thing as a Northern Cardinal is not necessarily a reasonable assumption! And yet, the assumption that a single label is a valid representation of non-monomorphic populations is a surprisingly common one, with actual consequences for the performance of image classification algorithms (Luccioni and Rolnick 2023). This assumption reveals a lot about our biases: male specimens are over-represented in museum collections, for example (Cooper et al. 2019). In a lot of species, we would need to split the taxonomic unit into multiple groups in order to adequately describe them.

The third mistake we have made is using predictors that are too vague. The "presence of red feathers" is not a predictor that can easily discriminate between the Northen Cardinal (yes for males, sometimes for females), the House Finch (a little for males, no for females), and the Red-Winged Black Bird (a little for males, no for females). In fact, it cannot really capture the difference between red feathers for the male House Finch (head and breast) and the male Red Winged Black Bird (wings, as the name suggests).

The final mistake we have made is in assuming that "red" is relevant as a predictor. In a wonderful paper, Cooney et al. (2022) have converted the color of birds into a bird-relevant colorimetric space, revealing a clear

latitudinal trend in the ways bird colors, as perceived by other birds, are distributed. This analysis, incidentally, splits all species into males and females. The use of a color space that accounts for the way colors are perceived is a fantastic example of why data science puts domain knowledge front and center.

Deciding which variables are going to be accounted for, how the labels will be defined, and what is considered to be within or outside the scope of the classification problem is *difficult*. It requires domain knowledge (you must know a few things about birds in order to establish criteria to classify birds), and knowledge of how the classification methods operate (in order to have just the right amount of overlap between features in order to provide meaningful estimates of distance).

2.2 The problem: classifying pixels from an image

Throughout this chapter, we will work on a single image – we may initially balk at the idea that an image is data, but it is! Specifically, an image is a series of instances (the pixels), each described by their position in a multidimensional colorimetric space. Greyscale images have one dimension, and images in color will have three: their red, green, and blue channels. Not only are images data, this specific dataset is going to be far larger than many of the datasets we will work on in practice: the number of pixels we work with is given by the product of the width and height of the image!

In fact, we are going to use an image with a lot more dimensions: the data in this chapter are coming from a Landsat 8 image, for which we have access to 7 different bands (the full data product has more bands, but we will not use them all.

Band number	Information
1	Aerosol

Band number	Information
2	Visible blue
3	Visible red
4	Visible green
5	Near-infrared (NIR)
6	Short wavelength IR (SWIR 1)
7	SWIR 2

From these channels, we can reconstruct an approximation of what the landscape looked like (by using the red, green, and blue channels).

The data extracted from the Landsat 8 files are *raw data*. Depending on the question we have in mind, they may not be the *right* data, in that they may not hold information that is relevant to our question.

We can decompose this image to have a look at the relationship between the red and green channels, for example:

2.3 The theory behind *k*-means clustering

In order to understand the theory underlying k-means, we will work backwards from its output. As a method for unsupervised clustering, k-means will return a vector of *class memberships*, which is to say, a list that maps each observation (pixel, in our case) to a class (tentatively, a cohesive landscape unit). What this means is that k-means is a transformation, taking as its input a vector with three dimensions (red, green, blue), and returning a scalar (an integer, even!), giving the class to which this pixel belongs. These are the input and output of our blackbox, and now we can start figuring out its internals.

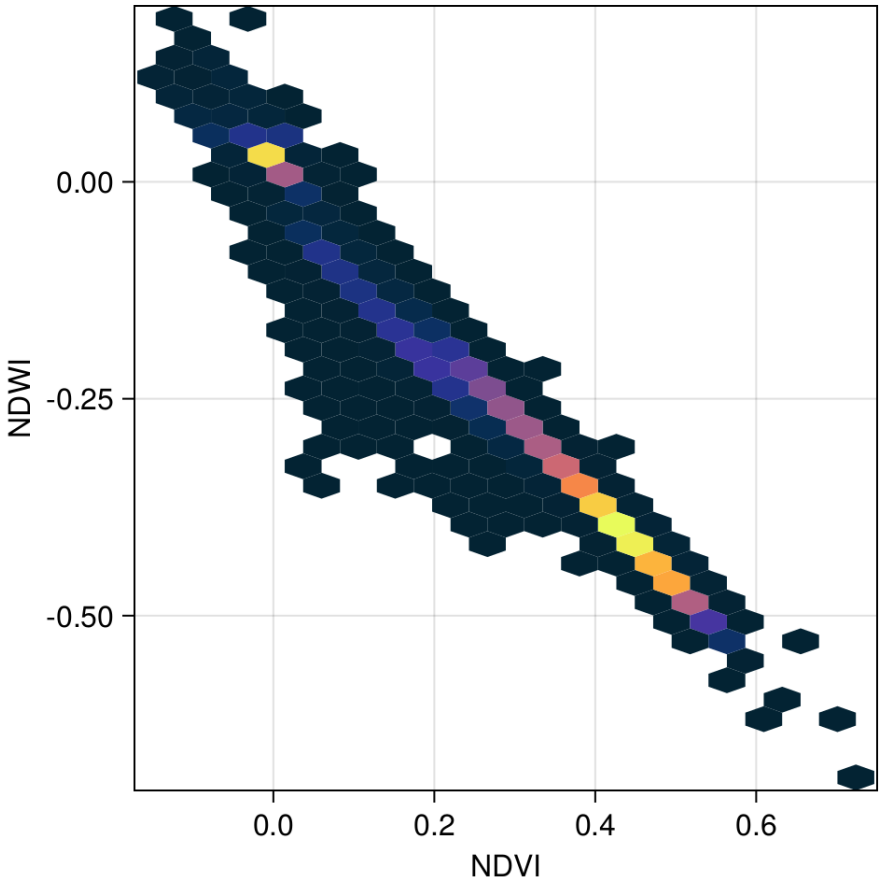

Figure 2.1: The pixels acquired from Landsat 8 exist in a space with many different dimensions (one for each band). Because we are interested in a landscape classification based on water/vegetation data, we use the NDVI and NDWI combinations of bands. Theyr are *derived* data, and represent an instance of feature engineering.

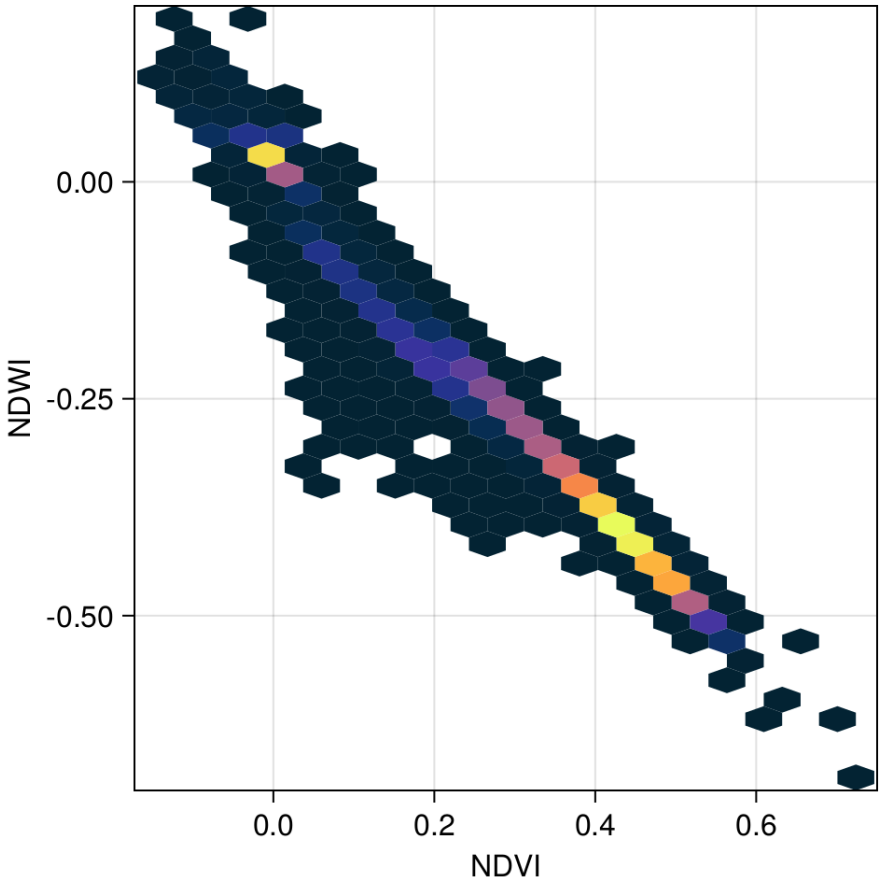

Figure 2.2: The pixels acquired from Landsat 8 exist in a space with many different dimensions (one for each band). Because we are interested in a landscape classification based on water/vegetation data, we use the NDVI and NDWI combinations of bands. Theyr are *derived* data, and represent an instance of feature engineering.

2.3.1 Overview of the algorithms

2.4 Identification of the optimal number of clusters

2.5 Application: optimal clustering of the satellite image data

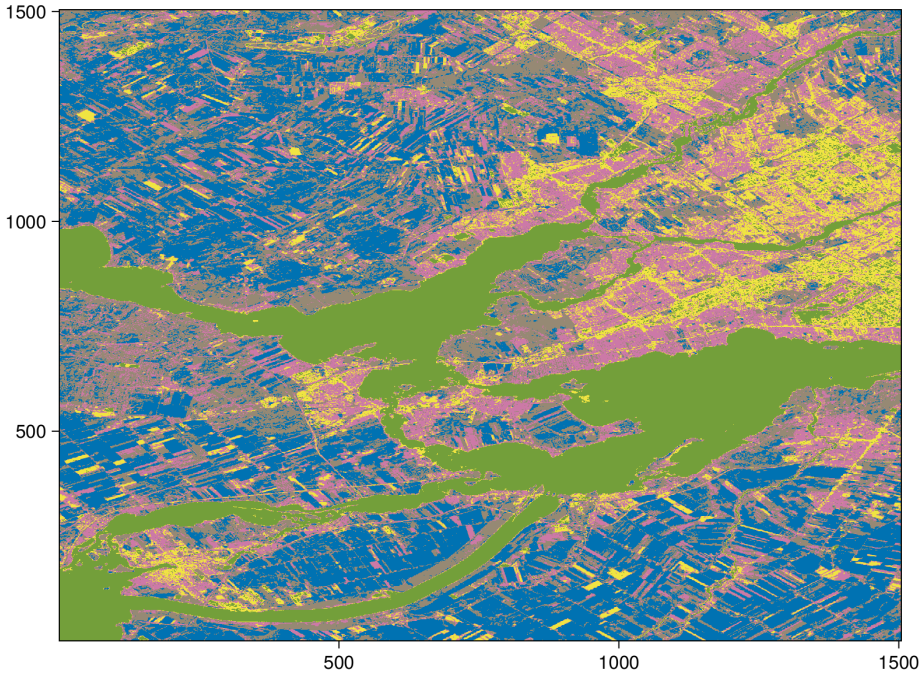

Figure 2.3: After iterating the k-means algorithm, we obtain a classification for every pixel in the landscape. This classification is based on the values of NDVI and NDWI indices, and therefore groups pixels based on a specific hypothesis.

2.6 Alternatives and improvements

EM

k-median

k-medoids

3 Minimizing error: the gradient descent algorithm

https://www.kaggle.com/datasets/abrambeyer/openintro-possum

4 Summary

In summary, this book has no content whatsoever.

References

Cooney, Christopher R., Yichen He, Zoë K. Varley, Lara O. Nouri, Christopher J. A. Moody, Michael D. Jardine, András Liker, Tamás Székely, and Gavin H. Thomas. 2022. "Latitudinal Gradients in Avian Colourfulness." *Nature Ecology & Evolution* 6 (5): 622–29. https://doi.org/10.1038/s41559-022-01714-1.

Cooper, Natalie, Alexander L. Bond, Joshua L. Davis, Roberto Portela Miguez, Louise Tomsett, and Kristofer M. Helgen. 2019. "Sex Biases in Bird and Mammal Natural History Collections." *Proceedings of the Royal Society B: Biological Sciences* 286 (1913): 20192025. https://doi.org/10.1098/rspb.2019.2025.

Luccioni, Alexandra Sasha, and David Rolnick. 2023. "Bugs in the Data: How ImageNet Misrepresents Biodiversity." *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (12): 14382–90. https://doi.org/10.1609/aaai.v37i12.26682.

Tuia, Devis, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, et al. 2022. "Perspectives in Machine Learning for Wildlife Conservation." *Nature Communications* 13 (1): 792. https://doi.org/10.1038/s41467-022-27980-y.