Fundamentals of Biodiversity Data Science

Timothée Poisot

2023-08-05

Table of contents

Preface

Data science 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. And as luck would have it, there are easier fields to master than data science. 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 *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. This entire process is driven by (or subject to, even) 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.

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.

To understand why, it helps to think of science (the application of the scientific method, that is) as cooking. There are general techniques one must master, and specific steps and cultural specifics, and there is a final product. 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 (they are mostly recommendations, in fact) is often justifiable if you feel like it. But this improvisation requires good skills, a clear mental map of the problem, 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 gave in the Department of Biological Sciences at the Université de Montréal, as well as a few workshops I ran with the Québec Centre for Biodiversity Sciences. In teaching data synthesis, data science, and machine learning to biology students, I realized that the field was missing a stepping stone to proficiency. There are excellent manuals covering the mathematics of data science and machine learning **REFS**; there are many good papers giving overviews of some applications of data science to biological problems **REFS**; and there are, of course, thousands of tutorials about how to write code. But one thing that students commonly called for was an attempt to tie concepts together. This is this attempt.

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

1.3 How to read this book

1.4 Some rules about this book

1.4.1 No code

1.4.2 No simulated data

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 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 9 image Vermote et al. (2016), 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
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) – this information is presented in Figure 2.1 . Or is it? If we were to invent a time machine, and go stand directly under Landsat 9 at the exact center of this scene, and look around, what would we see? We would see colors, and they would admit a representation as a three-dimensional vector of red, green, and blue. But we would see so much more than that! And even if we were to stand within a pixel, we would see a *lot* of colors. And texture. And depth. We would see something entirely different from this map; and we would be able to draw a lot more inferences about our surroundings than what is possible by knowing the average color of a 30x30 meters pixel.

But just like we can get more information that Landsat 9, so to can Landsat 9 out-sense us when it comes to getting information. In the same way that we can extract a natural color composite out of the different channels, we can extract a fake color one to highlight differences in the landscape; in Figure 2.2, we show such a fake color composite, that is particularly efficient at drawing our attention to the location of water in this area.

Both Figure 2.2 and Figure 2.1 represent the same physical place at the same moment in time; but through them, we are looking at this place with very different purposes. This is not an idle observation, but a core notion in data science: what we measure defines what we can see. In order to

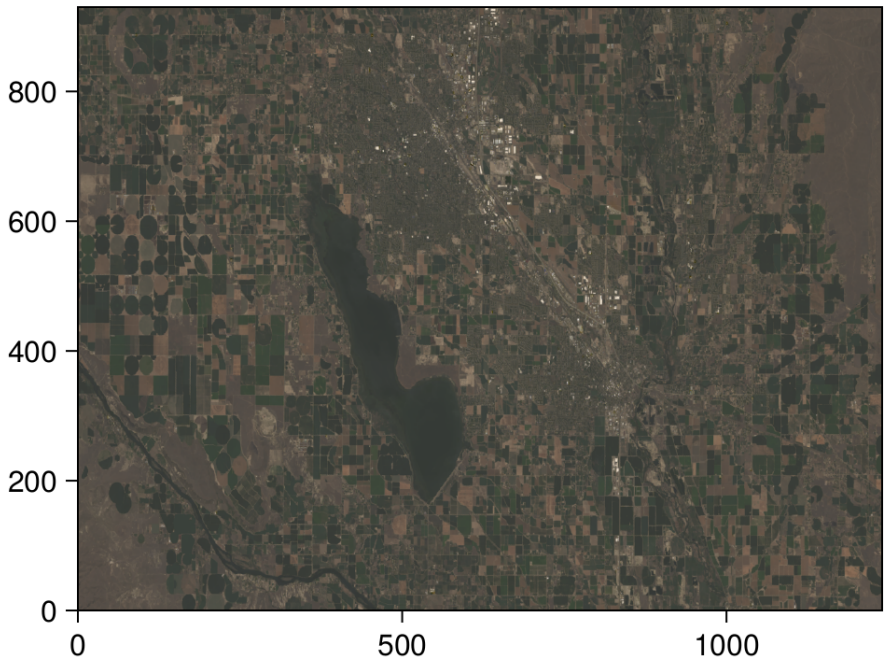

Figure 2.1: The Landsat 9 data are combined into the "Natural Color" image, in which the red, green, and blue bands are mapped to their respective channels. This looks eerily like the way we perceive the landscape. Note how little difference there is between the large body of water and the surrounding crops. This emphasizes that the combination of channels we use will limit our ability to separate the pixels into groups.

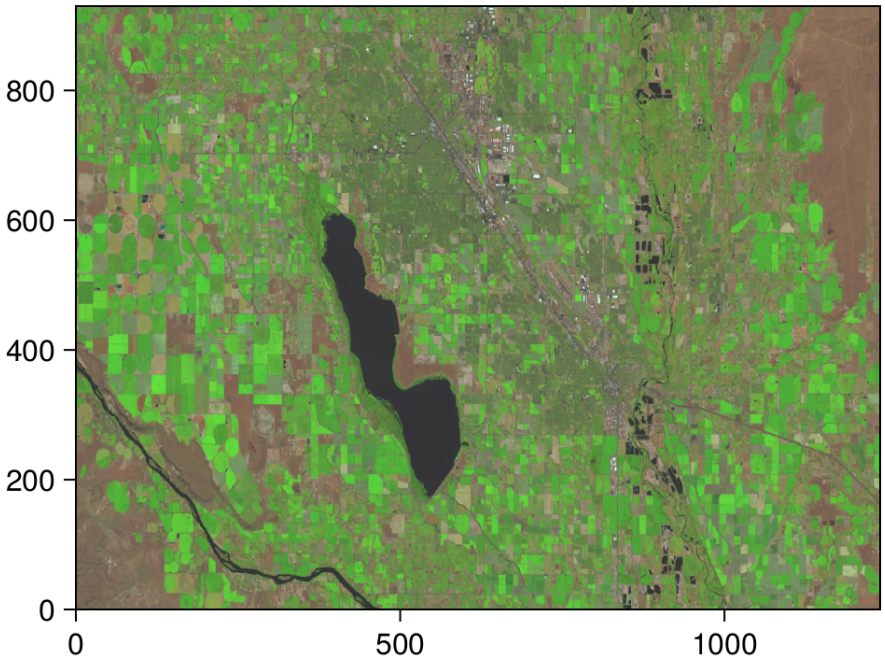

Figure 2.2: The same landscape as above is now presented in a fake color composite, where SWIR is mapped to the red channel, NIR to the green channel, and red to the blue channel. This highlights different values in the landscape, but is no more or less "real" than the true color composite. It is a visualization of the data, that represents different choices, questions, and assumptions. Compared to the Natural Color composite, this version of the data highlights the water, areas with vegetation, and more arid areas.

tell something meaningful about this place, we need to look at it in the "right" way.

So far, we have looked at this area by combining the raw data. Depending on the question we have in mind, they may not be the *right* data. In fact, they may not hold information that is relevant to our question *at all*; or worse, they can hold more noise than signal. Looking at Figure 2.1, we might wonder, "where are the fields?". And based on our knowledge of what plants do, we can start thinking about this question in a different way. Specifically, "is there a series of features of fields that are not shared by non-fields?". But this a complicated question to answer, and so we can simplify this by asking, "how can I combine data from the image to know if there is a plant?".

One way to do this is to calculate the normalized difference vegetation index, or NDVI (Kennedy and Burbach 2020). NDVI is derived from the band data (we will see how in a minute), and is an adequate heuristic to make a difference between vegetation, barren soil, and water. Because we are specifically thinking about fields, we can also consider the NDWI (water) and NDMI (moisture) dimensions: taken together, these information will represent every pixel in a three-dimensional space, telling us whether there are plants (NDVI), whether they are stressed (NDMI), and whether this pixel is a water body (NDWI).

Because there are a few guidelines (educated guesses, in truth, and the jury is still out on the "educated" part) about the values, we can look at the relationship between the NDVI and NDMI data Figure 2.3. For example, NDMI values around -0.1 (note how there is a strong cluster of points here) are low-canopy cover with low water stress; NDVI values from 0.2 to 0.5 are good candidates for moderately dense crops.

By picking these three values, instead of simply looking at the clustering of all the bands in the raw data, we are starting to refine what the algorithm see, through the lens of what we know is important about the system.

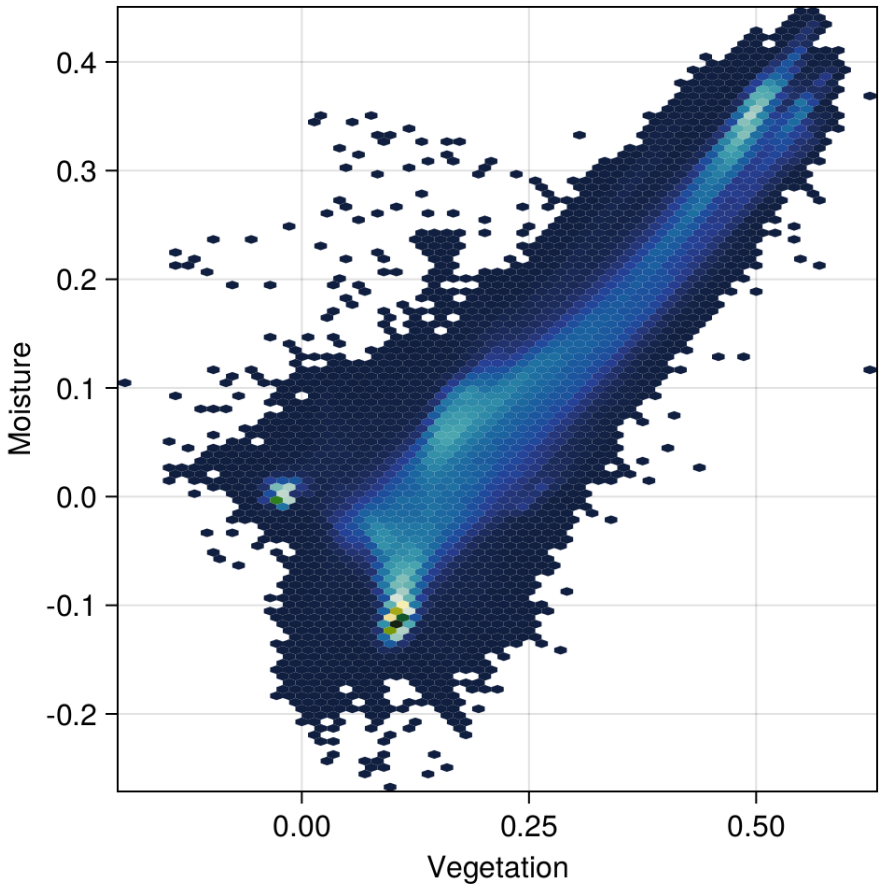

Figure 2.3: The pixels acquired from Landsat 8 exist in a space with many different dimensions (one for each band). Because we are interested in a landscape classification based on water/vegetation data, we use the NDVI, NDMI, and NDWI combinations of bands. These are *derived* data, and represent an instance of feature engineering: we have derived these values from the raw data.

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.

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

2.6 Alternatives and improvements

EM

k-median

k-medoids

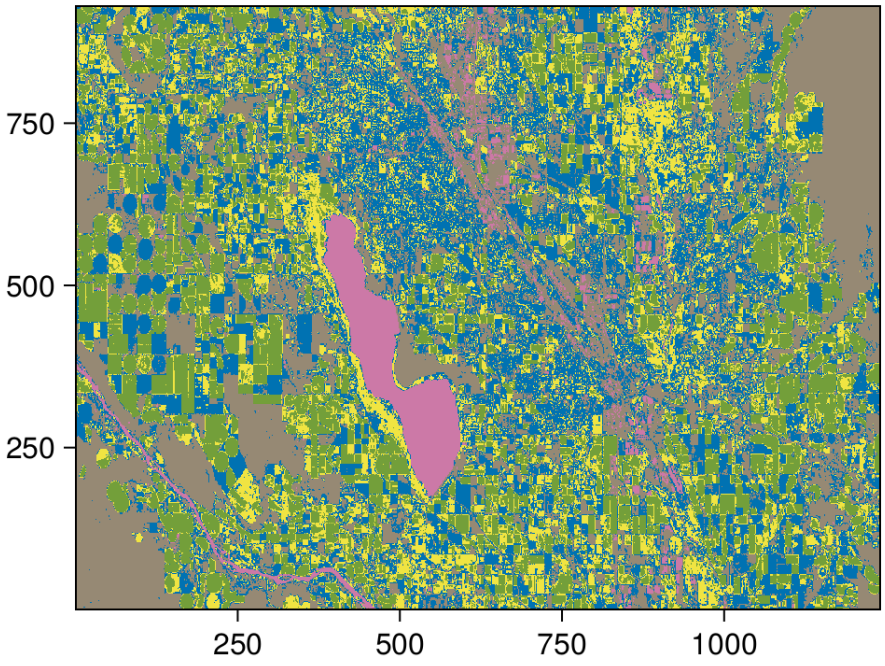

Figure 2.4: After iterating the k-means algorithm, we obtain a classification for every pixel in the landscape. This classification is based on the values of NDVI, NDMI, and NDWI indices, and therefore groups pixels based on a specific hypothesis. This clustering was produced using $k = 5$, *i.e.* we want to see what the landscape would look like when divided into five categories.

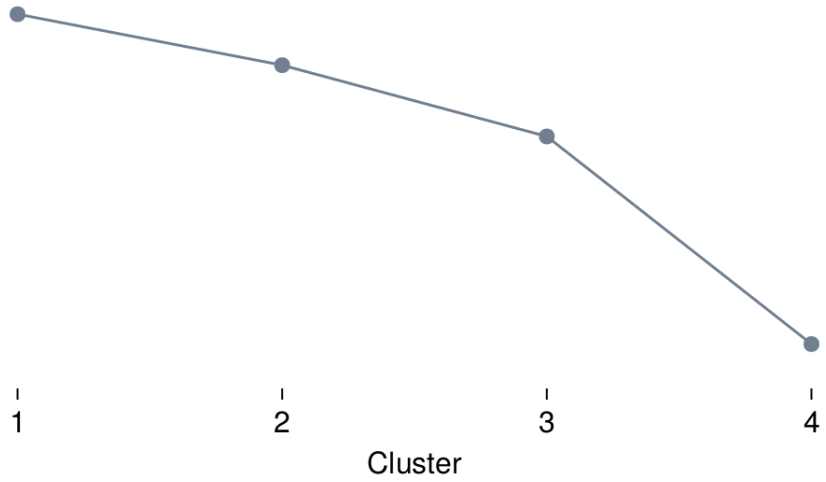

Figure 2.5: Number of pixels assigned to each class in the final landscape classification. In most cases, k-means will create clusters with the same number of points in them. This may be an issue, or this may be a way to ensure that whatever classes are produced will be balanced in terms of their representation.

3 Testing, training, validating

In Chapter 2, we were very lucky. Because we applied an unsupervised method, we didn't really have a target to compare to the output. Whatever classification we got, we had to live with it. It was incredibly freeing. Sadly, in most applications, we will have to compare our predictions to data, and data are incredibly vexatious. In this chapter, we will develop intuitions on the notions of training, testing, and validation; we will further think about data leakage, why it is somehow worse than it sounds, and how to protect against it.

3.1 Training

In data science (in machine learning in particular), we do *fit* models. We *train* them. This is an important difference: training is an iterative process, that we can repeat, optimize, and tweak. The outcome of training and the outcome of fitting are essentially the same (a model that is parameterized to work as well as possible on a given dataset), but it is good practice to adopt the language of a field, and the language of data science emphasizes the different practices in model training.

Training, to provide a general definition, is the action of modifying the parameters of a model, based on knowledge of the data, and the error that results from using the current parameter values. In Chapter 4, for example, we will see how to train a linear model using the technique of gradient descent. Our focus in this chapter is not on the methods we use for training, but on the data that are required to train a model.

Training a model is a process akin to rote learning: we will present the same input, and the same expected responses, many times over, and we will find ways for the error on each response to decrease.

In order to initiate this process, we need an untrained model. Untrained, in this context, refers to a model that has not been trained *on the specific problem* we are addressing; the model may have been trained on a different problem (for example, we want to predict the distribution of a species based on a GLM trained on a phylogenetically related species). It is important to note that by "training the model", what we really mean is "change the structure of the parameters until the output looks right". For example, assuming a simple linear model like $c(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, training this model would lead to changes in the values of β, but not to the consideration of a new model $c(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. Comparing models is the point of validation, which we will address later on.

need for instances

need for responses

3.2 Testing

3.3 Validating

3.4 Strategies to split data

3.5 Data leakage

4 Minimizing error: the gradient descent algorithm

5 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.

Kennedy, Stephanie, and Mark Burbach. 2020. "Great Plains Ranchers Managing for Vegetation Heterogeneity: A Multiple Case Study." *Great Plains Research* 30 (2): 137–48. https://doi.org/10.1353/gpr.2020.0016.

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.

Vermote, Eric, Chris Justice, Martin Claverie, and Belen Franch. 2016. "Preliminary Analysis of the Performance of the Landsat 8/OLI Land Surface Reflectance Product." *Remote Sensing of Environment* 185 (November): 46–56. https://doi.org/10.1016/j.rse.2016.04.008.