A Bayesian Approach to OC4

Erdem M. Karaköylü

September 13, 2018

Contents

1 Background

A number of approaches have been proposed to derive chlorophyll concentration from remote sensing reflectance (Rrs). The two main paradigms are semi-analytical and empirical. Semi-analytical approaches are more easily interpretable as they are built at least in part on a mechanistic understanding of the processes at play. The highly variable nature of satellite oceanography in general, and ocean color in particular, has made this approach challenging in that so far a consistently satisfactory performance remains elusive. The The Regression Model:

$$log_{10}\left(chlor_a\right) = a_0 + \sum_{i=1}^{j} a_i log_{10}\left(\frac{max\left(Rrs\left(\lambda_{blue}\right)\right)}{Rrs\left(\lambda_{green}\right)}\right) \tag{1}$$

2 Issues with ordinary least squares (ols)

There are a number of shortcomings that makes *ols* unsuitable for developing the model above. I'll briefly outline some shortcomings of OLS; first, general issues with *ols*, then issues specific to the development of OC_4

2.1 General shortcomings

- R^2 as a diagnostic encourages a higher number of parameters
- Almost garanteed to overfit on the training data.

2.2 Specific shortcomings

- polynomial regression inherently affected by multicollinearity
- lack of testing simpler formulations
- data paucity

3 Bayesian paradigm

3.1 Pros and cons of being a bayesian

3.1.1 a perception problem

3.1.2 uncertainty

3.1.3 on the use of priors

3.2 Bayes'rule

Bayes' rule is relatively straightforward. Given two events A and B that are not independent, their joint probability $P(A, B)$ can be written in two different ways:

$$P(A, B) = P(A|B) \times P(B) = P(B|A) \times P(A) \tag{2}$$

The above becomes quite handy when one of the conditional probabilities, say $P(B|A)$ is harder to compute than the other. This is easily dealt with by rearraging the terms above, which leads to Baye's rule:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \tag{3}$$

In the context of scientific enquiry, conditional probabilities allow relating hypotheses to collected data, by way of model formulation. Given a set of hypotheses, H, that address a specific question, and given a data set, D, collected to estimate the validity of these hypotheses, (3) can be rewritten as:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \tag{4}$$

A bit of vocabulary is in order here:

- $P(H)$: **the prior** - This is the belief of the experimenter *prior* to seeing the data.
- $P(D|H)$: **the likelihood** - While frequentists will tell you that this is what you want and that you should maximize it to approach the "truth", that is seldom correct and a great way to overfit the model to the training data, thereby leading to poor prediction of unseen data. More about this later.
- $P(D)$: **the evidence** - For the purpose of this paper, it's enough to refer to it as a normalizing constant that basically makes sure that the probabilities computed through (4) sum to 1. One of the main tools of bayesian inference, and one that I will use further below, the Markov Chain Monte Carlo (MCMC) sampler does away with this otherwise often computatiationally intractable construct.
- $P(H|D)$: **the posterior** - This is the set of probabilities that allow to enable the most likely hypotheses, and provides in the process uncertainties around the chose estimates.

3.3 Warm-up: Inferring Earth's Land proportion

I stole and modified this example from McElreath (2015). The goal here is to infer the proportion of land. The hypotheses are then all the possible values that this proportion parameter, ranging from 0 to 1. The steps are as follows:

3.3.1 defining a prior

- $P(H)$
- can be vague or specific, depending on the researcher's prior knowledge
- prior formulation is akin to placing bets on a roulette table
- a vague prior will rapidly get overwhelmed by the collected data
- a vague prior may lead to overfitting
- a strong (regularizing) prior will "calm" the model - a large dataset will overwhelm it
- a strong prior may lead to underfitting

3.3.2 collect data

- random sampling of locations (Fig. 1)
- land? water?
- define the likelihood model (more about this later)
- compute likelihood of data given hypotheses; $P(D|H)$

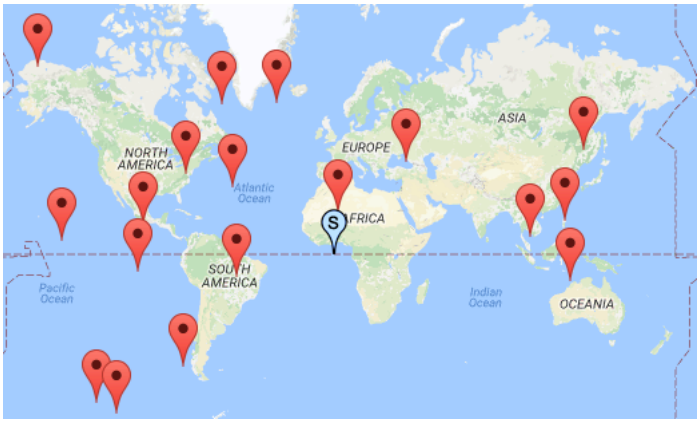

Figure 1: Random sampling of 18 geographic locations

3.3.3 update hypotheses probability given new data

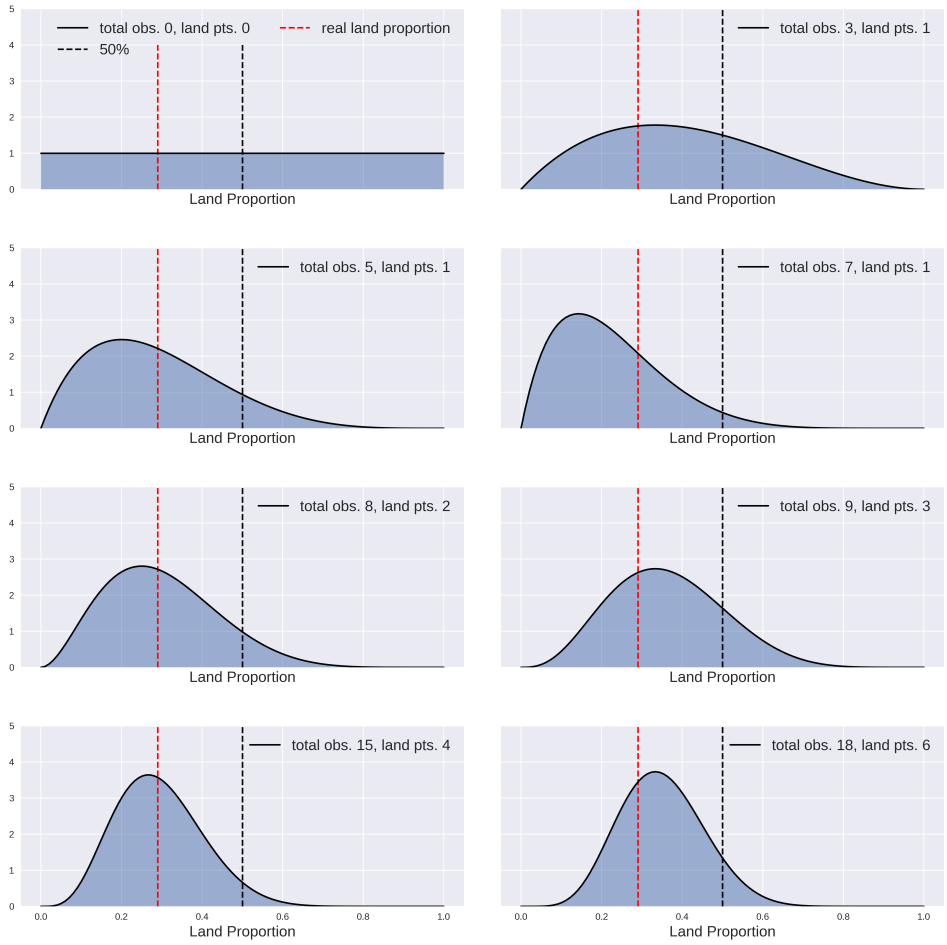

Figure 2: Inference starting from a flat uniform prior (top left panel). The inference progresses top-down and left-right. - - -: true proportion sought; - - -: 50 land/water reference; ——: distribution inferred from combining data-driven likelihood with prior.

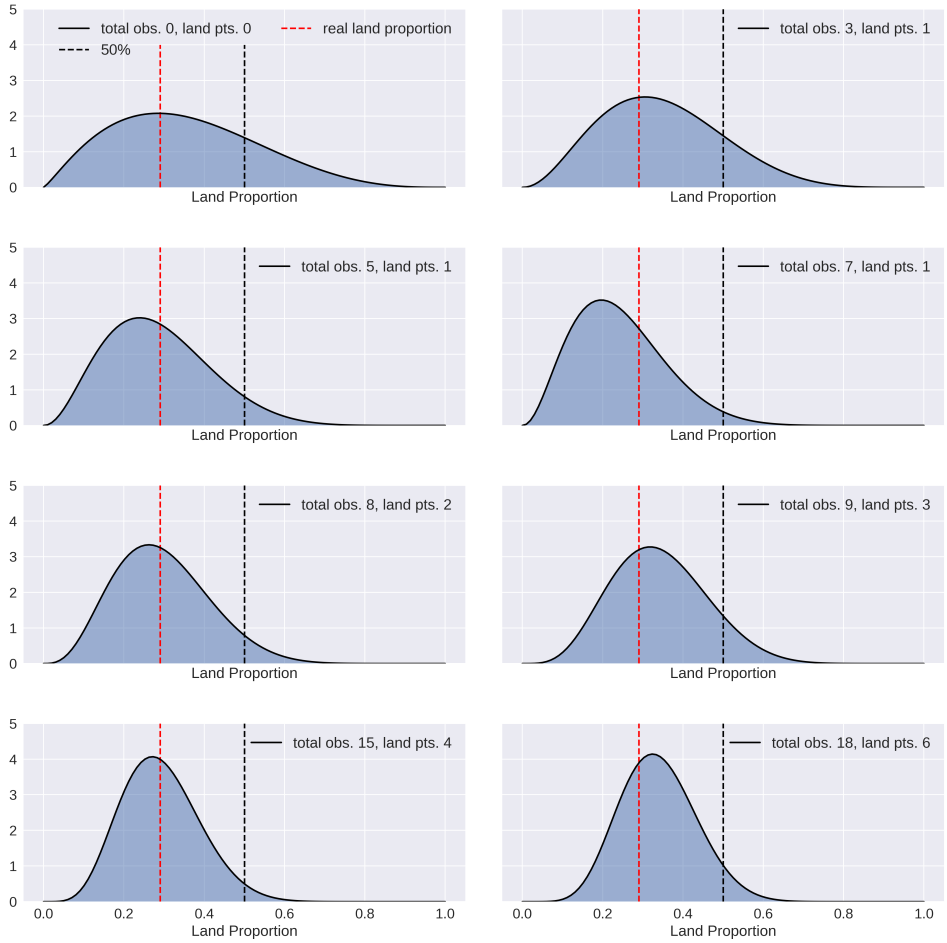

Figure 3: Starting from a weak beta prior

4 Bayesian development of OC4

4.1 Data overview

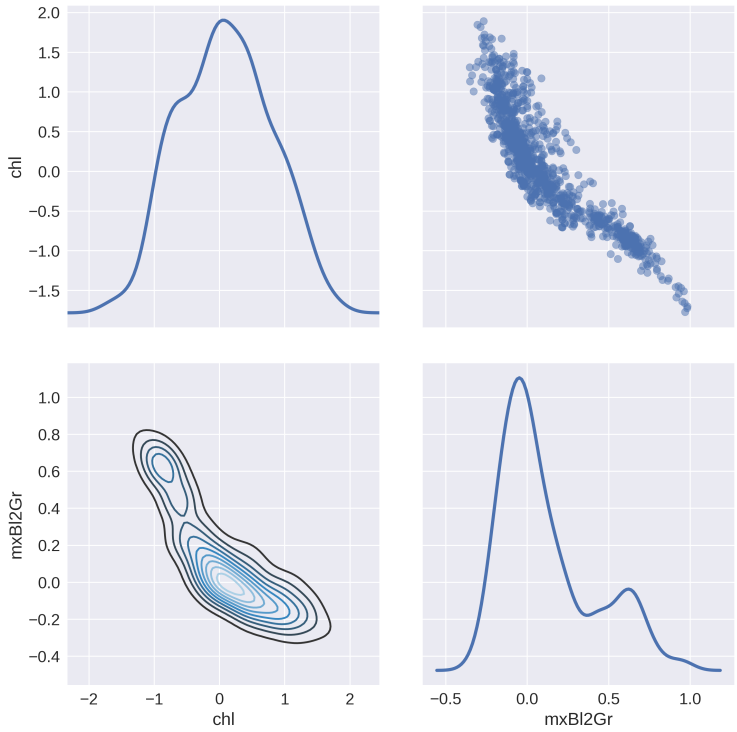

Figure 4: NOMAD data used for model development.
chl: $log_{10}(chl_a)$; mxBl2Gr: $log_{10}\frac{max(Rrs_{443,490,510})}{Rrs_{555}}$

4.2 Models and priors

4.3 Fitting and Diagnostics

5 Bayesian Model Evaluation

5.1 Posterior predictive checks (PPC)

5.1.1 in-sample

5.1.2 out-of-sample

5.2 Information criteria and model evaluations

6 Developing Alternative Empirical Models

6.1 Linear Models

6.2 Hierarchical Models

6.3 Partially Pooled Models

6.4 Bayesian Neural Networks