QUANTIFYING UNCERTAINTY IN PARETO ESTIMATES OF GLOBAL LAKE AREA

A PREPRINT

J. Stachelek
Earth System Observations
Los Alamos National Laboratory
Los Alamos, NM, USA, 87544
jsta@lanl.gov

October 6, 2021

ABSTRACT

Size is a critical factor determining the rate and occurrence of specific lake processes such as carbon sequestration and greenhouse gas emissions. Because we do not have a complete census of all lakes, upscaling estimates of such processes at broad spatial scales requires the use of lake size-abundance distributions rather than empirical measurements of area. Existing lake census efforts are incomplete such that as lakes become smaller they are more likely to by omitted either because they are too small to be resolved from remote sensing products or because of limited ground surveying effort (i.e. "censoring" of small lakes relative to large lakes).

The present study explores one potential shortcoming of prior approaches estimating global lake area using lake size-abundance distributions. Namely, these prior approaches rely on frequentist curve fitting techniques combined with an ad-hoc cutoff determination strategy (visual inspection to determine a likely censoring point). This yields an over-exact lake area estimate that is overwhelmingly reported with no uncertainty bounds. I address this shortcoming by fitting models in a Bayesian framework where each parameter contributes uncertainty to model estimates. I show that such although such models produce a more realistic estimate of lake area uncertainty they underestimate true total lake area. The degree of this underestimation is likely related to the proportion of the dataset subject to censoring. Ultimately, this may explain the fact that total lake area estimates have increased through time as the resolution of lake databases has improved.

1 Introduction

Area is a critical factor determining the contribution of lakes to global biogeochemical cycling. As such, existing evidence suggests that lakes are key components of overall carbon burial, sequestration, and greenhouse gas emissions (**??**). The calculation of such fluxes requires two terms 1) an areal flux rate and 2) the distribution of total lake area. As with all estimates at broad spatial scales, uncertainty estimates are of great interest both because they allow for intercomparisons among similar lake studies and because they allow for comparisons between lakes and other ecosystems.

The techniques typically used to determine uncertainty in areal flux rates and the techniques typically used to determine uncertainty in total lake area distributions differ greatly. In the former case, uncertainties can typically be calculated in straightforward manner from literature compilations (**??**). In the latter case, uncertainty in total lake area distributions is challenged by the fact that no existing database is a complete census of all lakes (**?**) and the distribution of individual lake areas spans a range of approximately 7 orders of magnitude. The largest lakes are so big they could otherwise be classified as inland seas ($> 10^4\ km^2$) while the smallest are barely larger than a regulation sized soccer (football) field ($> 10^{-3}\ km^2$). A consequence of these challenges is that uncertainty varies within the distribution as a function of lake size. The area of the largest lakes is known exactly while the area of the smallest lakes is incomplete below a certain unknown threshold. This can occur either because lakes are too small to be resolved from remote sensing products or

because of limited ground surveying effort. The omission or "censoring" of small lakes occurs because we know that small lakes exist but below a certain threshold we have limited knowledge of their exact areas (**?**).

Estimating total lake area from a sample of lakes requires a conceptual model of how lakes are formed (i.e. the data generating process). Typically, lake areas are treated as arising from a fractal generating process due to the fact that landform topography, which determines the placement of lakes, can itself be treated as a fractal generating process. Indeed, many other geomorphological phenomena that are dependent on landform topography such as coastline length are often well-described by fractal generating processes (**?**). A challenge in modelling such data generating processes for lake areas is that large lakes likely follow a different data generating process than that of the smallest lakes. Whereas small lakes are constrained by landform topography, large lakes are essentially unconstrained by local landform topography and are instead constrained by the placement and arrangement of continents (**?**). As large lakes become even larger, they have a greater probability of intersecting a continent edge and becoming an estuary or embayment rather than a lake. Consquently, lake databases are said to be truncated on large lakes because we know that large lakes are essentially fixed in space and cannot occur in any given location (**?**).

From the preceding discussion it is clear that estimating total lake area requires a method of dealing both with the fact that lake databases are 1) truncated at large lakes and 2) censored at small lakes. Previous studies estimating global lake area have essentially not dealt with the first issue but instead modified their estimation process to limit its effect on the results (see methods section). They have dealt with the second issue by specifying an ad-hoc cutoff value below which empirically determined lake areas are discarded and subsequently back calculated. Hereafter, I refer to this strategy as the "cutoff" method.

In the present study, I explore two different methods for calculating uncertainty bounds around global lake area. First, I calculate these bounds using a frequentist approach for estimating total lake area (i.e. the cutoff method) via a simulation study. Second, I compare these frequentist uncertainty bounds with those calculated using a Bayesian framework. I carry out these demonstrations using a simulated dataset so that the ability of each method to recover the "true" parameter values can be evaluated. The use of a simulated dataset has the further benefit of avoiding potential confounding factors such as heterogeneity of survey effort or unknown data precision.

2 Methods

Lake areas are typically treated as arising from a scale-invariant fractal generating process (**?????**). This means that the number of lakes in one size class is proportional to the number of lakes in the preceding size class irrespective of their magnitudes. The numerical form describing such a process is a power-law function. One of the statistical tools often used to model data that follow a power-law function is the Pareto distribution which has a probability density function (pdf) of:

$$pdf(A) = \alpha x_{min}^{\alpha} A^{-(\alpha+1)} \tag{1}$$

where A is lake area, a controls the “shape” of the distribution and x_{min} controls the “scale” of the distribution (**?**). Lake area studies using the Pareto distribution do not typically use the pdf directly. Instead, they use the inverse (complementary) cumulative distribution function (ccdf) (i.e. quantile function):

$$ccdf(A) = \frac{x_{min}}{(1-A)^{\frac{1}{\alpha}}} \tag{2}$$

The reason for using the ccdf is two-fold. First, it stabilizes model estimates in the lower tail of the distribution (**?**). This can be seen from the simulated data in Figure **??** where the Pareto pdf contains a lot of noise in the tail but the ccdf appears smoothed. The smoothing of the tail is also desirable because it functions as a way of dealing with the truncated nature of lake databases (i.e. the area of large lakes is known exactly). The second reason for using the ccdf is that it provides a computational shortcut for estimating the Pareto shape parameter a because it is numerically equivalent to the slope of the ccdf in log-log space (**?**).

For evaluation purposes, I generated a simulated dataset of 10,000 lake areas following the Pareto distribution using inverse transform sampling (**?**). Lakes in my simulated dataset have a minimum and maximum area of approximately 1 and 81000 km^2 respectively. This maximum was chosen to be approximately as large as Lake Superior but less than the Caspian Sea following **?**. The "true" total area of these lakes is approximately 230000 km^2. I simulated a censored lake dataset by excluding lakes smaller than e^1. This excludes (i.e. censors) approximately 60% of the total dataset. I approximated the "true" lake area total by constructing the empirical distribution function (edf) of the data which approximates the underlying Pareto cdf (**?**). Then I used this estimate of the cdf slope to generate cdf estimates for the

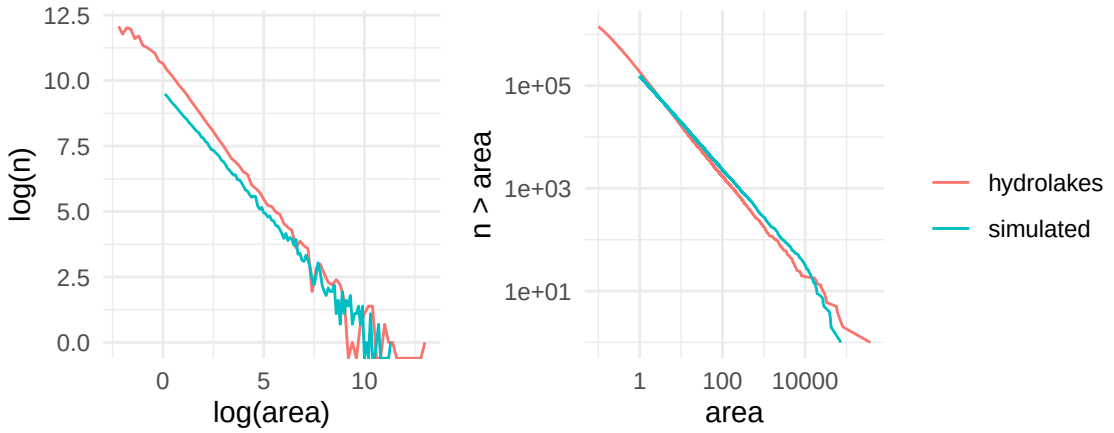

Figure 1: Realization of a Pareto (A) probability density function and (B) complementary cumulative distribution function.

censored lakes. I combined these cdf estimates with the edf values from the known lakes before calculating the sum of the inverted distribution (Figure **??**).

I estimated the Pareto shape parameter a in a frequentist framework by calculating the the slope of the edf in log-log space using linear regression in `R` (**?**). I evaluated uncertainty in both a and total lake area in a Bayesian framework using Stan (**?**). Instead of computing on the edf (as in the frequentist case), I computed directly on the pdf with the following Stan model:

```
data {
  int<lower=0> N;
  real x[N];
}
parameters {
  real<lower=0> alpha;
  real<lower=0> xmin;
}
model {
  real lpa[N];

  xmin ~ gamma(.001, .001);
  alpha ~ gamma(.001, .001);

  for (i in 1:N) {
    lpa[i] = pareto_lpdf(x[i] | xmin, alpha);
  }

  target += sum(lpa);
}
```

I used uninformative gamma priors for both the x_{min} and a parameters following **?**. I ran the model with four chains of 8,000 iterations and used the Stan defaults for burn-in and thinning which specify a burn-in of half the iterations and no thinning.

3 Results

Visual inspection of the frequentist method of computing on the edf appeared to produce a reasonable density estimate for small censored lakes (Figure **??**). In addition, estimates of total lake area are somewhat close to the "true" value (Table **??**). However, uncertainty around the frequentist estimates is unrealistically small (Figure **??**).

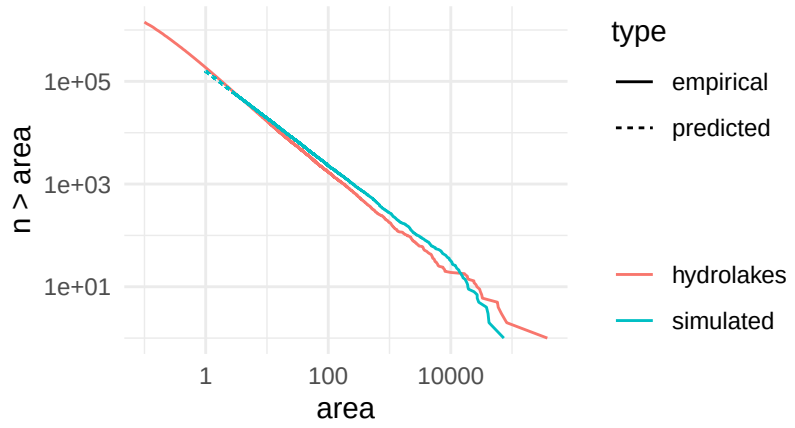

Figure 2: Censored lake area edf (solid line) and cdf estimate (dashed line).

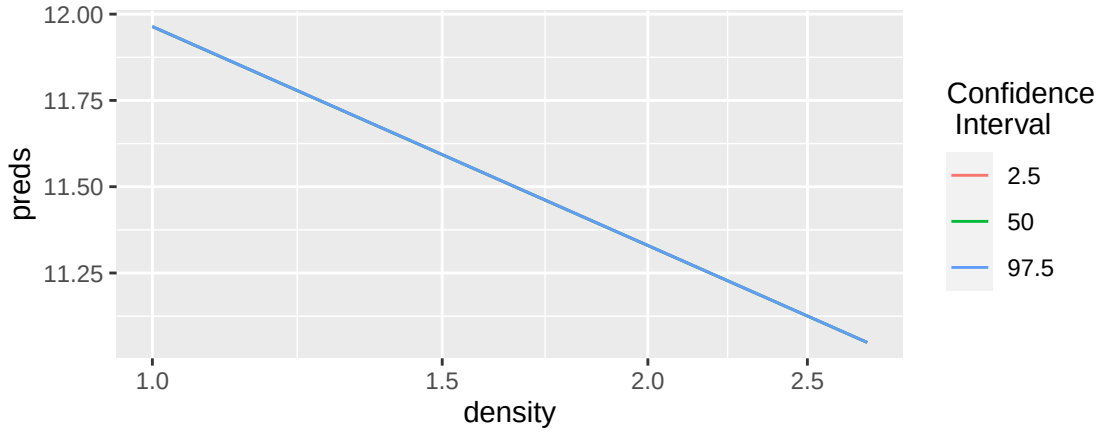

Figure 3: Confidence interval of frequentist density predictions for censored data (the dashed portion in Figure 2. The confidence interval here has essentially no width.

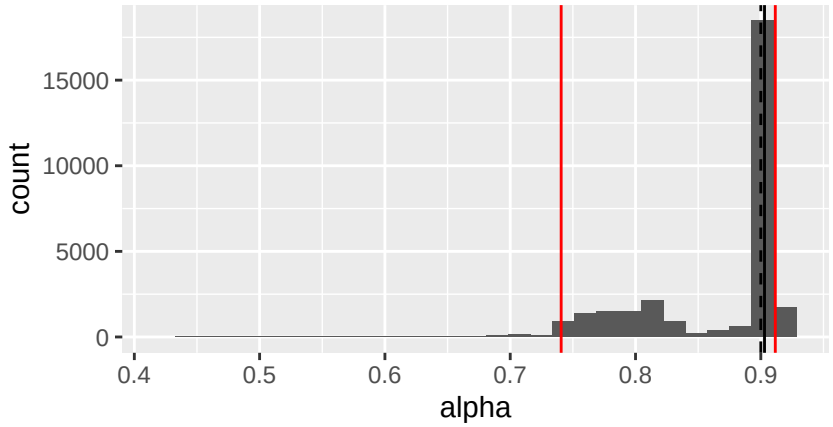

Figure 4: Median (black line) and central 95 percent interval estimates of alpha (red lines). Here the 'true' alpha is 0.9.

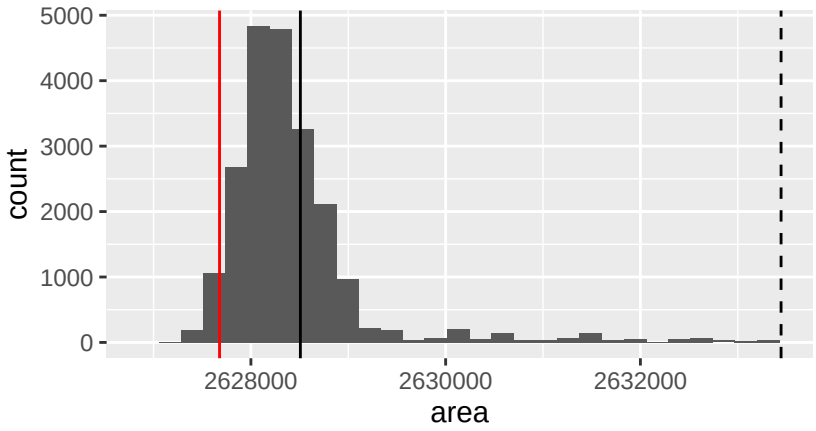

Figure 5: Median (black line) and central 95 percent interval estimates of total lake area (red lines). Here the true total lake area is marked with a dashed vertical line.

Table 1: 'True' total lake area from the uncensored edf and estimated lake area from a combination of the censored edf and the estimated cdf.

True Area (km^2)	Estimated Area (km^2)
2.271e+05	2.263e+05

Instead of the essentially fixed a and total lake area estimates produced by the frequentist approach, I found substantial variability in both a (95% CI: 0.86, 0.92) and total area using a Bayesian approach (Fig **??**, **??**). In particular, the Bayesian 95% credible intervals for both a and total lake area encapsulate the true values (Fig **??**, **??**). Despite better uncertainty estimates using a Bayesian approach, both the frequentist and Bayesian approaches underestimated the true value of a and total lake area.

4 Discussion

I have shown that the typical frequentist cutoff method produces reasonable estimates of the density of small censored lakes but that it does not capture uncertainty in total lake area (Table **??**, Figure **??**). Furthermore, I have shown that models fit using a Bayesian approach indicate substantial uncertainty in both total lake area and the underlying Pareto shape parameter a used to derive these estimates (Figure **??**, **??**).

Although the 95% credible interval of the Bayesian total lake area estimates encapsulate the true total lake area, the median value underestimates the true total lake area (Figure **??**). It is likely that such underestimation, would increase with a greater proportion of censoring. This may explain the steady increase in estimates of global lake area through time from approximately 3 to 5 million km^2 as lake area databases have improved their accuracy (**???**). Future work on estimating global lake area should consider implementing a sensitivity analysis looking at the response of total area estimates to variation in the degree of censoring.

A caveat of the present study is that it is difficult to confirm whether or not any particular dataset follows a fractal generating or Pareto process without strong prior knowledge. Indeed data following many heavy-tailed distributions such as the lognormal or negative exponential can appear to be equivalent to power-law distributions such as the Pareto (**?**). The results of the present study confirm that an apparent change in the shape of the lower tail of a lake area distribution does not necessarily indicate a change in the data generating process. Note that the simulation dataset analyzed herein *is* a truly fractal generated dataset with a homogeneous data generating process yet a cursory look would seem to indicate a change of data-generating process in the lower tail (Figure **??**).

In addition to a sensitivity analysis of censoring, future work might consider more complex models that treat lake areas as a mixture of a Pareto distribution for small lakes and either a negative exponential or lognormal distribution for large lakes. Such an approach has been demonstrated by **?** and **?**. Both studies show that the point at which the distribution mixtures converge can provide valuable inference. In the case of lakes, such a convergence point may indicate a change in the data generating process such as the point at which lake areas are controlled by continent placement rather than fractal landscape morphology (**??**).