Estimating Uncertainty in Global Lake Area

Joseph Stachelek[a]

[a]Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA

This version was compiled on March 5, 2019

Size is a critical factor determining the rate and occurrence of specific lake processes such as carbon sequestration. Unfortunately, the exact lake size-abundance distribution is unknown in part because we do not have a complete census of all lakes. More specifically, as lakes become smaller they are less likely to be included in lake databases either because they are too small to be resolved from remote sensing products or because of limited ground surveying effort. The exclusion of small lakes relative to large lakes is known as censoring.

The present study explores one potential shortcoming of previous approaches at estimating global lake area and dealing with censoring. Namely, that the typical frequentist curve fitting and ad-hoc cutoff determination strategy (visual inspection to determine a likely censoring point) taken by previous studies yields an over-exact lake area estimate with no reported uncertainty. 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.

Introduction. Size is a critical factor determining the rate and occurence of specific lake processes such as carbon burial and sequestration (DelSontro *et al.*, 2018). However, the relative contribution of small and large water bodies to overall nutrient and material cycling is unknown (Alexander *et al.*, 2000). One way to estimate their relative contributions is to combine information on per area processing rates with estimates of total area (Winslow *et al.*, 2015).

One of the challenges in estimating total lake area is that the distribution of individual lake areas spans a range of approximately 7 orders of magnitude whereby 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 field ($> 10^{-3}\ km^2$). Another important characteristic of lake area is that 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. In this sense, lake databases are said to be censored at small lake areas because we know that small lakes exist but below a certain threshold we have limited knowledge of their exact areas (Hamilton *et al.*, 1992).

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 (Newman, 2005).

Another challenge in estimating total lake area is that the area distribution of the largest lakes likely follows a different data generating process than that of the smallest lakes. The reason that largest lakes likely follow a different data generating process is that they are constrained not by local landform topography but by the placement and arrangement of continents. As lakes become larger they have a greater probability of intersecting a continent edge and becoming an estuary or embayment rather than a lake (Goodchild, 1988). In this sense, 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 (Hamilton *et al.*, 1992).

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 truncated at large lakes and 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 the effects of using the typical frequentist approach for estimating total lake area (i.e. the cutoff method) via a simulation study. In addition, I demonstrate an approach to produce uncertainty estimates of total lake area in a Bayesian framework.** I evaluate the effects of the cutoff method on a simulated dataset because it is not sensitive to other potential confounding factors such as heterogeneity of survey effort or data precision.

Methods. Lake areas are typically treated as arising from a scale-invariant fractal generating process (Winslow *et al.*, 2015; Downing *et al.*, 2006; McDonald *et al.*, 2012; Goodchild, 1988; Hamilton *et al.*, 1992). Essentially, 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 (Shalizi, 2017). 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 (Newman, 2005). This can be seen from the simulated data in Figure 1 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. 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 (Downing *et al.*, 2006).

For evaluation purposes, I generated a simulated dataset of 10,000 lake areas following the Pareto distribution using inverse transform sampling (Newman, 2005). 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 Lehner and Döll (2004). 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 (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 (Newman, 2005). Then I used this estimate of the cdf slope to generate cdf estimates for the censored lakes. I combined these cdf estimates with the edf values from the known lakes before calculating the sum of the inverted distribution (Figure 2).

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` (Team and others, 2018). I evaluated uncertainty

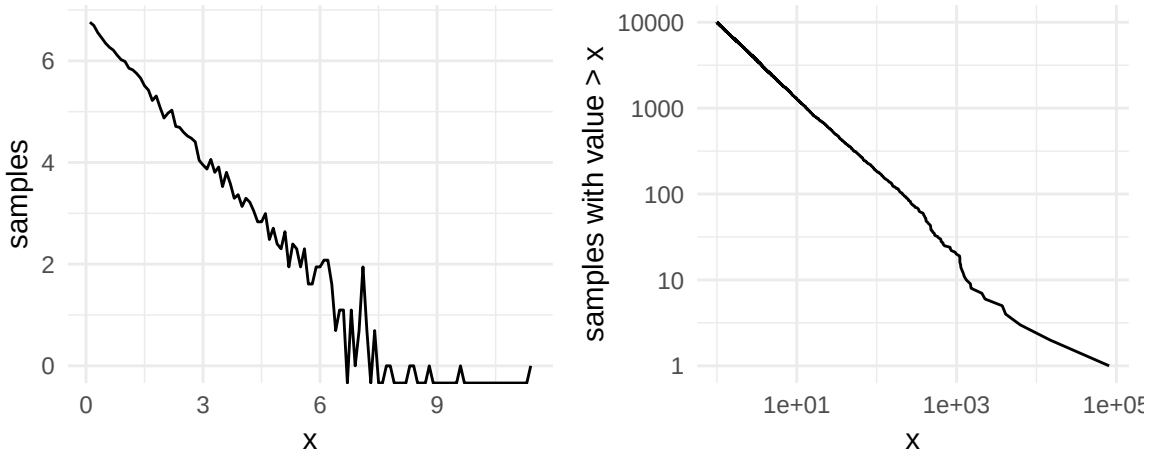

Fig. 1. Realization of a Pareto (A) probability density function and (B) complementary cumulative distribution function.

in both a and total lake area in a Bayesian framework using Stan (Stan Development Team, 2017). 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 vague gamma priors for both the x_{min} and a parameters following Scollnik (2007). 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.

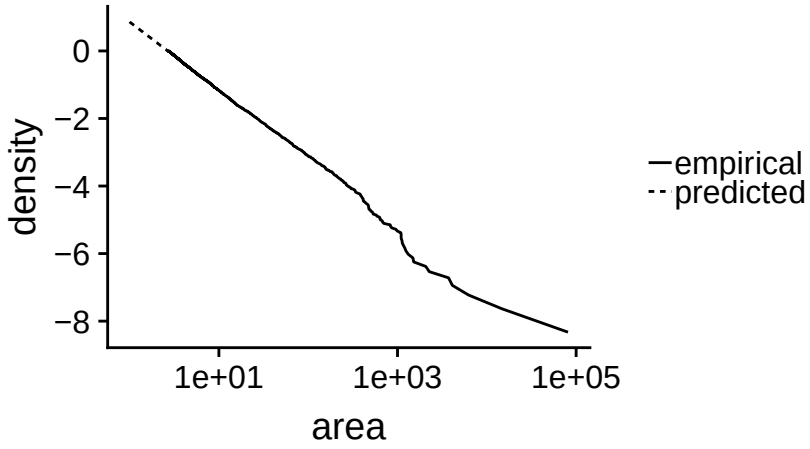

Fig. 2. Censored lake area edf (solid line) and cdf estimate (dashed line).

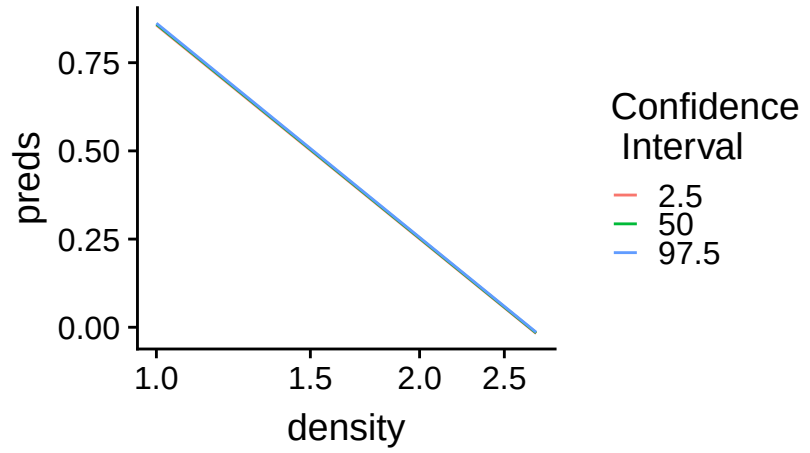

Fig. 3. Confidence interval of frequentist density predictions for censored data (the dashed portion in Figure 2. The confidence interval here has essentially no width.

Results. Visual inspection of the frequentist method of computing on the edf appeared to produce a reasonable density estimate for small censored lakes (Figure 2). In addition, estimates of total lake area are somewhat close to the "true" value (Table 1). However, uncertainty around the frequentist estimates is unrealistically small (Figure 3).

Instead of the essentially fixed a and total lake area estimates produced by the frequentist approach, I found a substantial variability in both a (95% CI: 0.86, 0.92) and total area using a Bayesian approach (Fig 4, 5). In particular, the Bayesian 95% credible intervals for both a and total lake area encapsulate the true values (Fig 4, 5). Despite better uncertainty estimates using a Bayesian approach, both the frequentist and Bayesian approaches underestimated the true value of a and total lake area.

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 1,

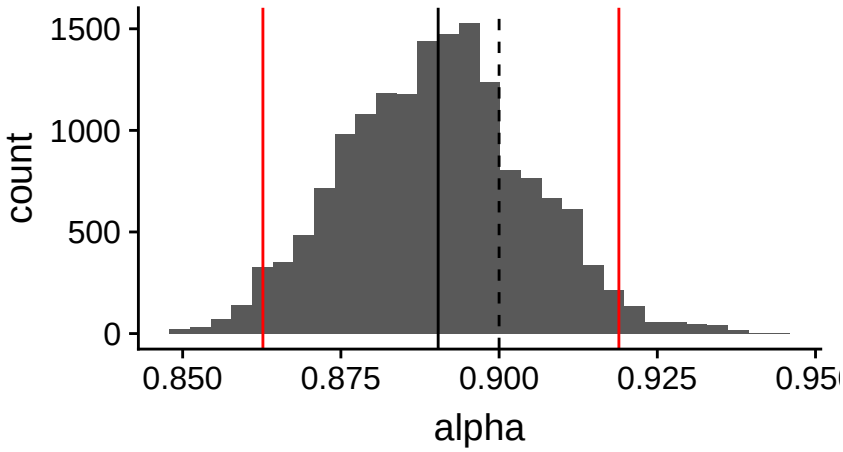

Fig. 4. Median (black line) and central 95 percent interval estimates of alpha (red lines). Here the 'true' alpha is 0.9.

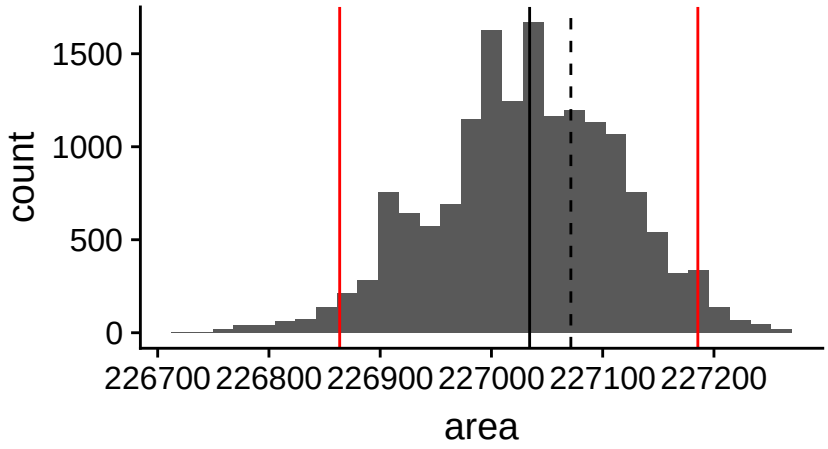

Fig. 5. Median (black line) and central 95 percent interval estimates of total lake area (red lines). Here the true total lake area is marked with a dashed vertical line.

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 (km2)	Estimated Area (km2)
2.271e+05	2.263e+05

Figure 3). 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 4, 5).

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 5). 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 (Lehner and Döll, 2004; Downing *et al.*, 2006; Verpoorter *et al.*, 2014). 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.

Future investigators should be mindful that is difficult to confirm whether or not any particular dataset following 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 (Clauset *et al.*, 2009). 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 2).

In addition to a sensitivity analysis of censoring, future work should 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 Bonabeau *et al.* (1999) and Scollnik (2007). Both studies have shown 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 (i.e. the point at which lake area are controlled by continent

placement rather than fractal landscape morphology (Goodchild, 1988; Hamilton *et al.*, 1992).

References

Alexander RB, Smith RA, Schwarz GE (2000). "Effect of Stream Channel Size on the Delivery of Nitrogen to the Gulf of Mexico." *Nature*, **403**(6771), 758.

Bonabeau E, Dagorn L, Freon P (1999). "Scaling in Animal Group-Size Distributions." *Proceedings of the National Academy of Sciences*, **96**(8), 4472–4477. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.96.8.4472.

Clauset A, Shalizi CR, Newman ME (2009). "Power-Law Distributions in Empirical Data." *SIAM review*, **51**(4), 661–703.

DelSontro T, Beaulieu JJ, Downing JA (2018). "Greenhouse Gas Emissions from Lakes and Impoundments: Upscaling in the Face of Global Change: GHG Emissions from Lakes and Impoundments." *Limnology and Oceanography Letters*. ISSN 23782242. doi:10.1002/lol2.10073.

Downing JA, Prairie YT, Cole JJ, Duarte CM, Tranvik LJ, Striegl RG, McDowell WH, Kortelainen P, Caraco NF, Melack JM (2006). "The Global Abundance and Size Distribution of Lakes, Ponds, and Impoundments." *Limnology and Oceanography*, **51**(5), 2388–2397.

Goodchild (1988). "Lakes on Fractal Surfaces: A Null Hypothesis for Lake-Rich Landscapes." *Mathematical Geology*.

Hamilton SK, Melack JM, Goodchild MF, Lewis W (1992). "Estimation of the Fractal Dimension of Terrain from Lake Size Distributions." *Lowland floodplain rivers: Geomorphological perspectives. Wiley*, pp. 145–163.

Lehner B, Döll P (2004). "Development and Validation of a Global Database of Lakes, Reservoirs and Wetlands." *Journal of Hydrology*, **296**(1-4), 1–22. ISSN 00221694. doi:10.1016/j.jhydrol.2004.03.028.

McDonald CP, Rover JA, Stets EG, Striegl RG (2012). "The Regional Abundance and Size Distribution of Lakes and Reservoirs in the United States and Implications for Estimates of Global Lake Extent." *Limnology and Oceanography*, **57**(2), 597–606. ISSN 00243590. doi:10.4319/lo.2012.57.2.0597.

Newman M (2005). "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics*, **46**(5), 323–351. ISSN 0010-7514, 1366-5812. doi:10.1080/00107510500052444.

Scollnik DPM (2007). "On Composite Lognormal-Pareto Models." *Scandinavian Actuarial Journal*, **2007**(1), 20–33. ISSN 0346-1238, 1651-2030. doi:10.1080/03461230601110447.

Shalizi CR (2017). "Advanced Data Analysis from an Elementary Point of View." p. 860.

Stan Development Team P (2017). "Stan Modeling Language Users Guide and Reference Manual, Version 2.17. 0." *Technical report*.

Team RC, others (2018). "R: A Language and Environment for Statistical Computing."

Verpoorter C, Kutser T, Seekell DA, Tranvik LJ (2014). "A Global Inventory of Lakes Based on High-Resolution Satellite Imagery." *Geophysical Research Letters*, **41**(18), 6396–6402. ISSN 00948276. doi:10.1002/2014GL060641.

Winslow L, Read J, Hanson P, Stanley E (2015). "Does Lake Size Matter? Combining Morphology and Process Modeling to Examine the Contribution of Lake Classes to Population-Scale Processes." *Inland Waters*, **5**(1), 7–14. ISSN 20442041, 2044205X. doi:10.5268/IW-5.1.740.