ANZIAM J. 56(2015), 373–380
doi:10.1017/S1446181115000073

A POISSON–PARETO MODEL OF CHLOROPHYLL-A FLUORESCENCE SIGNALS IN MARINE ENVIRONMENTS

S. WOODCOCK ✉[1], B. MANOJLOVIC[1], M. E. BAIRD[2] and P. J. RALPH[3]

(Received 30 September, 2013; revised 3 March, 2015)

Abstract

Because of its central role in the global carbon cycle, quantifying the biomass of photosynthetic microalgae in the oceans is crucial to our ability to estimate the oceans' carbon drawdown. Many traditional methods of primary production assessment have proven to be extremely time consuming and, consequently, have handled only very small sample sizes. The recent advent of in situ bio-optical sensors, such as the water quality monitor (WQM), is now providing lower cost and higher throughput data on these crucial biological communities. These WQMs, however, only quantify the total fluorescence of all individual cells within their optical sample windows, irrespective of size. In this paper, we further develop an established model, based on Pareto random variables, of the size structure of the microalgae community to understand the effect of the WQMs' sampling and data pooling on their estimates of algal biomass. Unfortunately, evaluating sums of Pareto variables is a notoriously difficult problem. Here, we utilize an approximation for the right-tail of the resulting distribution to derive parameter estimates for the underlying size structure of the microalgae community.

2010 *Mathematics subject classification*: 92B05.

Keywords and phrases: Poisson–Pareto random variables, shifted Hill's estimators.

1. Introduction

Photosynthetic carbon fixation in the oceans plays an important role in the global carbon cycle [8]. Phytoplankton use chlorophyll to capture light and typically range in size from 0.2 μm up to 100 μm [19]. Although marine photosynthetic organisms are almost entirely single-celled microalgae and comprise less than 1% of the total global plant biomass, due to their high turnover they account for more than 40% of total

[1]School of Mathematical Sciences, University of Technology Sydney, Sydney, Australia;
e-mail: Stephen.Woodcock@uts.edu.au, Bojana.Manojlovic@uts.edu.au.
[2]CSIRO Oceans and Atmosphere Flagship, GPO Box 1538, Hobart 7001, Australia;
e-mail: Mark.Baird@csiro.au.
[3]Plant Functional Biology and Climate Change Cluster, University of Technology Sydney, Sydney, Australia; e-mail: Peter.Ralph@uts.edu.au.

global carbon fixation [7]. As such, the importance of understanding and quantifying the communities in the oceans cannot be overstated.

The traditional method of quantifying microalgae density is through manual microscopic cell counts. Although extremely detailed (information such as taxon identification and morphology are available), this approach has a number of drawbacks. Most notably, it is an extremely laborious and, hence, costly method for data collection. It takes many hours of a laboratory-based worker's time to provide cell counts which can be obtained much more cheaply via automated methods. Furthermore, in practice, it tends to introduce detection limitations as well; there is a smallest cell size which will be detected: this is a function of the microscope used [11].

In the last few decades, an alternative technology for quantifying microalgae in the oceans has been developed. Bio-optical sensors such as the WQM (WetLabs, USA) [15] use fluorescence signals to detect chlorophyll-a (Chl-a). Chl-a is the principal photosynthetic pigment common to all microalgae. When cells are excited by light with a wavelength of 470 nm, the Chl-a molecule re-emits light at a longer wavelength of 695 nm; this is Chl-a fluorescence. The amount of light returned at this longer wavelength is used as a proxy for total Chl-a concentration [13]. These Chl-a fluorescence readings correlate with the microalgal biomass present.

The advantages of employing WQMs are numerous. Total Chl-a measurements can be obtained much more quickly and for a fraction of the cost of direct cell counts. Furthermore, the impact of small cells which may not have been accurately quantified by traditional microscopic counts may lead to underestimation of the biomass, but bio-optical sensors capture all cell sizes. The major drawback of bio-optical sensors is that they return just a single total Chl-a estimate, summed over an indeterminate number of individual cells. We do not, therefore, explicitly gain information on the exact profile of the microalgal community. This creates a need for translating these single Chl-a measurements into likely profiles of the biological community observed. In this paper, we further develop the prevailing model of the size profile of microalgae to incorporate the sampling effects inherent to fluorescence signals and provide parameter estimates for the underlying size distribution.

2. Methods

There are two factors governing variability between the Chl-a fluorescence measurements: the number of microalgal cells in each sample window and the contribution of each individual cell to the overall signal strength within that. Addressing the first of these, we make one simple assumption. We assume that the presence or absence of an individual cell in the sample is independent of any others. This is reasonable since we know from direct cell counts that the typical volumes of microalgae are orders of magnitude smaller than the sample volume, so there should be no *crowding out* of individuals within the sample window. In other words, we model the number of microalgae cells detected by the fluorescence sensor in each sample by an independent realization of a Poisson random variable.

https://doi.org/10.1017/S1446181115000073 Published online by Cambridge University Press

To model the contribution of each individual cell to the overall fluorescence signal, we assume that the size distribution of individuals is such that the volumes of cells in each sample follow a Pareto distribution. This is a widely accepted distribution for the size structure of microalgae, and arises from the observation that the normalized biomass size spectrum [22] (effectively binning particles by logarithmic size classes) suggests a power law to describe the size structure [6, 24]. The Pareto distribution is a long right-tailed distribution with domain $x \in [m, \infty)$, slope parameter $\alpha > 0$ and probability density function

$$f(x) = \frac{\alpha m^{\alpha}}{x^{\alpha+1}}.$$

Again, we assume that there are no size effects and that, for example, the presence of one large individual cell does not make it any more or less likely that there will be other large cells present. Since the largest microalgal cells have diameters <0.1 mm and the WQM measures around 10^3 mm^3 sample volume (1 ml) of water, this assumption seems reasonable. We, therefore, assume that the volume of each individual cell is taken to be an independent realization of the same Pareto variable. Finally, to convert between microalgal cell volumes and their Chl-a fluorescence, we need to know the fluorescence/volume ratio. Here, we use an allometric relationship determined experimentally by Finkel [9], which gives

$$c \approx 2.06 \times 10^7 v^{-0.320},$$

where v is cell volume (μm^3) and c is Chl-a concentration (pg Chl-a/μm^3) within the cell.

For ease of notation, we drop the scaling factor of 2.06×10^7 throughout the remainder of this paper. Total Chl-a readings can easily be *normalized* by division by this constant and the scaled readings used.

This experimentally obtained scaling factor is consistent with what might be expected. To provide a simple conceptual illustration of the model, assuming that fluorescence from an individual cell scales proportional to its surface area (that is, proportional to the square of its diameter) and that volume scales proportional to the cube of its diameter would give a fluorescence/volume ratio scaling inversely proportional to the diameter and hence inversely proportional to the cube root of the volume.

By establishing that an individual cell's fluorescence scales in this way, we know that if cell volumes are Pareto distributed, then their contributions to the overall fluorescence signal are also Pareto distributed. This is because of a standard property of this distribution, namely, a power-law transformation applied to a Pareto random variable gives a variable which is itself Pareto distributed. That is, if $X \sim \text{Pareto}(m, \alpha)$, then, for $Y = X^n$, $Y \sim \text{Pareto}(m^n, \alpha/n)$ for all $n \geq 0$. Since we have that Chl-a concentration scales proportionally to $v^{-0.320}$, we therefore have the total fluorescence of cell volume v scales proportionally to $v^{-0.320} \times v = v^{0.68}$. Finally, we note that we rely on one key property of both the Pareto distributions, namely that a truncated or (left-tail) censored Pareto random variable is itself a Pareto random variable. That

https://doi.org/10.1017/S1446181115000073 Published online by Cambridge University Press

is, if $X \sim \text{Pareto}(m, \alpha)$, then $X \mid X > T \sim \text{Pareto}(T, \alpha)$ for all $T \geq m$. This assures that we need not be concerned by any detection limitations that arise as a result of the technology employed in fluorescence sensors. For example, if there are individual cells containing Chl-a, but they are too small to be detected by the bio-optical sensor, these will not impact our overall estimate for the key shape parameter α, and the scale parameter m will be the value of the smallest individual cell detected by this sensor.

What we have, then, is a parameter estimation problem for a Poisson–Pareto model. We need to establish estimators for the parameters m and α from the observed WQM signals $s_1, s_2, \ldots, s_K$, where each s_j ($j \in \{1, 2, \ldots, K\}$) is an independent realization of S, where S is the sum of the individual fluorescences of a (Poisson) random number of sampled cells and λ is the expected number of cells per sample. That is,

$$S = \sum_{i=1}^{N} F_i \quad \text{for } F_i \sim \text{Pareto}(m^{0.68}, \alpha/0.68) \text{ and } N \sim \text{Poisson}(\lambda).$$

This Poisson–Pareto model is not novel. It has been extensively studied in the context of actuarial science, albeit with a different motivation and purpose. The sizes of insurance claims are often modelled with a Pareto distribution. That is, the vast majority of claims are relatively small and extremely large claims occur very infrequently [10, 16]. With a random number of claims in a given time period, each arising independently, the resulting model for the total value of claims in the time period is Poisson–Pareto. In actuarial science, the use of this model is the exact opposite of our application here. For known parameter values for both the Poisson and Pareto distributions, the primary motivation is for the calculation of *ruin probabilities*, that is, the chance that the total claims over some time period will exceed some tolerable threshold. For the problem faced in interpreting Chl-a fluorescence signals, we instead need to recover the parameter values from observed signals. An analogue for actuarial scientists would be the recovery of the expected number of claims in a year, and both the shape and scale parameters for the distribution of sizes for each claim from a list of annual total claims.

Unfortunately, evaluation of the distributions of sums of independent Pareto variables is notoriously difficult [4, 17, 18, 21, 25]. Except for a small handful of special cases [14], simple analytic forms for these do not exist. Solutions have been found giving ruin probabilities [2, 10]; however, these do not help with the problem of parameter estimation from aggregated sums. Even certain results about the asymptotic behaviour of the sums of a large number of independent Pareto variables are of no value to this exercise. However, it has been shown [17] that for large x, the right-hand tail, $x \gg Nm$, of the distribution of $N > 1$ independent Pareto(m, α) variables, $X_1, X_2, \ldots, X_N$, can be approximated by

$$P(X_1 + X_2 + \cdots + X_N > x + Nm) \approx N\left(\frac{m}{x + m}\right)^{\alpha}.$$

(Since the Pareto distribution is strongly right-skewed, the first of these conditions ($x \gg Nm$) holds, and we know from direct cell counts that the second ($N > 1$) also holds.)

https://doi.org/10.1017/S1446181115000073 Published online by Cambridge University Press

Removing the condition on the value of N, we therefore get an exceedence probability for $S = \sum_{i=1}^{N} F_i$ and $F_i \sim \text{Pareto}(m^{0.68}, \alpha/0.68)$, where $N \sim \text{Poisson}(\lambda)$, of

$$P(S > s + \lambda m) \approx \sum_{N=0}^{\infty} \frac{e^{-\lambda}\lambda^N}{(N-1)!}\left(\frac{m^{0.68}}{s + m^{0.68}}\right)^{\alpha/0.68} = \lambda\left(\frac{m^{0.68}}{s + m^{0.68}}\right)^{\alpha/0.68}.$$

Therefore, we have a cumulative probability density function

$$F(s) = P(S \le s) \approx 1 - \lambda m^{\alpha}(s - \lambda m + m^{0.68})^{-\alpha/0.68}.$$

Given a dataset $S_1, S_2, \ldots, S_K$, we first define the order statistics $S_{(1)} \ge S_{(2)} \ge \cdots \ge S_{(K)}$ with ties arbitrarily broken. We are then able to calculate parameter estimates (denoted by a $\hat{}$ above the respective parameter) via the shifted Hill's estimator [1, 12], that is, essentially the conditional maximum likelihood estimator. Compared to a simple maximum likelihood estimator, the Hill's estimator is a more robust method for measuring the thickness of heavy-tailed distributions, such as the Pareto distribution, since it only uses information from the largest order statistics, taken from the portion of the distribution where the right-tail approximation holds.

Selecting the largest r observations with r as large as possible, but such that the largest $r + 1$ order statistics lie within the right-hand tail, we obtain

$$\hat{\alpha} = \frac{0.68r}{\sum_{i=1}^{r} \ln[(S_{(i)} + \hat{\lambda}\hat{m} - \hat{m}^{0.68})/(S_{(r+1)} + \hat{\lambda}\hat{m} - \hat{m}^{0.68})]}.$$

The other two parameter estimates satisfy

$$\hat{\lambda} = \frac{r}{K\hat{m}^{\hat{\alpha}}}(S_{(r+1)} - \hat{\lambda}\hat{m} - \hat{m}^{0.68})^{\hat{\alpha}/0.68}$$

and

$$\frac{\hat{\alpha}r}{0.68}\sum_{i=1}^{r}(S_{(i)} + \hat{\lambda}\hat{m} - \hat{m}^{0.68}) = (S_{(r+1)} - \hat{\lambda}\hat{m} + \hat{m}^{0.68})\left(\frac{\hat{\alpha}}{0.68} + 1\right).$$

Explicit forms of these are not readily obtainable, although they can be solved numerically. Alternatively, since the smallest individuals from a Pareto size distribution exist in such high abundances, the parameter m may be estimable independently from other samples, given sufficient resolution of microscopy.

3. Numerical results

To verify the suitability of the estimators derived in the previous section, we simulated a synthetic dataset from known parameters, and used the results in the previous section to recover the parameter values. For each simulation, $K = 5000$ WQM readings were simulated. Estimators were calculated using the largest 5%, 10% and 20% of the readings ($r = 250$, $r = 500$ and $r = 1000$). The parameter ranges used in the simulation were $0.6 \le \alpha \le 1.6$, $0.2 \le m \le 125$ (μm) and $40 \le \lambda \le 640$. These ranges were selected for consistency with some estimates previously determined for

https://doi.org/10.1017/S1446181115000073 Published online by Cambridge University Press

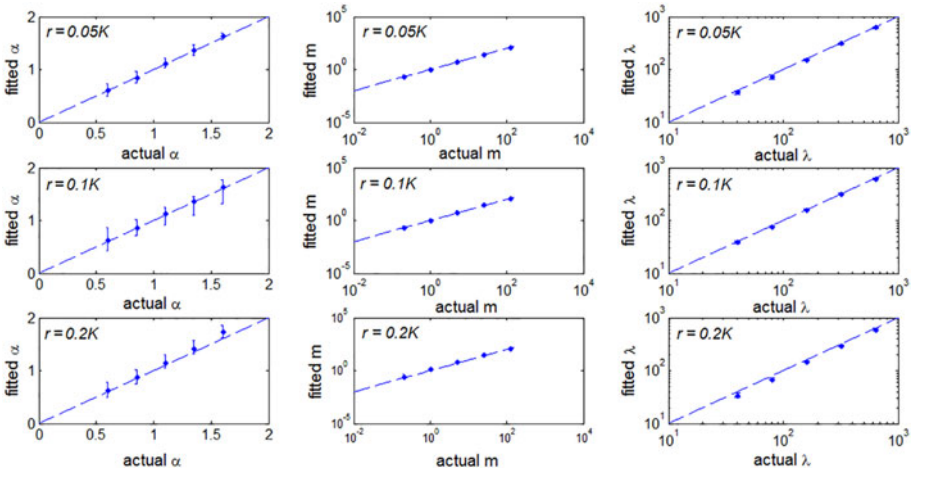

FIGURE 1. Plots of fitted parameter values against actual parameter values for synthetic datasets for three different r values. Except where the plot shows a range of values for that parameter, the three parameters were set as $\alpha = 1.1$, $m = 1\ \mu$m and $\lambda = 160$. Error bars represent 95% confidence intervals.

α [20, 23], m (based on the cyanobacteria *Prochlorococcus* species [3]) and λ [5] (based on the manufacturer's claim of 1 ml sample regions).

Overall, parameter recovery was satisfactory, especially for the parameter α. This is, of course, the most *important* parameter in the model, since it describes the size profile of the observed community. The other two parameters are artefacts of the technologies and techniques employed to quantify the detected cells. A different WQM might produce very different values for these two parameters, but it should not produce different estimates for the key parameter α. In general, the estimators tended to fractionally overestimate the value of m and underestimate the value of λ but only within the same order of magnitude. For example, with $r = 250$, the estimated value for $m = 125\ \mu$m was $\hat{m} = 129.5\ \mu$m with (121.6, 135.8) 95% confidence interval. Similarly, the estimated value for $\lambda = 640$ was $\hat{\lambda} = 630.8$ with (601.4, 644.0) 95% confidence interval (see Figure 1).

For all three parameters, we found that parameter recovery was most accurate and with least variance when $r = 250$, that is, only the top 5% of datapoints were used for estimation. Including more datapoints clearly led to the right-tail approximation becoming less valid and hence worsened the estimators.

4. Conclusions

The Poisson–Pareto model presented here is a simple extension of the existing theory of the size distribution of aquatic microalgae. Making the additional assumption that the individual cells sampled by fluorescence sensors do not exhibit any size-dependent or density-dependent effects (that is, they can be reasonably represented by a Poisson variable), we developed a framework for obtaining parameter estimates for

https://doi.org/10.1017/S1446181115000073 Published online by Cambridge University Press

the underlying Pareto distribution from only the aggregated sums of cells detected. As such, we are now able to use data from WQMs, which can be deployed for a fraction of the cost of labour-intensive microscopic techniques, to build up a more accurate picture of the likely size distribution of the microalgae population.

This ability to model newer bio-optical monitoring technologies and to characterize the observed communities from the samples they produce is vital to the future of the multidisciplinary work between applied mathematicians and observational marine scientists. As newer quantitative technologies emerge, and ever larger and more accurate datasets are gathered, the range of statistical and modelling challenges faced will certainly increase.

References

[1] I. B. Aban and M. M. Meerschaert, "Shifted Hill's estimator for heavy tails", *Comm. Statist. Simulation Comput.* **30** (2001) 949–962; doi:10.1081/SAC-100107790.

[2] H. Albrecher and D. Kortschak, "On ruin probability and aggregate claim representations for Pareto size distributions", *Insurance Math. Econom.* **45** (2009) 362–373; doi:10.1016/j.insmatheco.2009.08.005.

[3] M. E. Baird and I. M. Suthers, "A size-resolved pelagic ecosystem model", *Ecol. Model.* **203** (2007) 185–203; doi:10.1016/j.ecolmodel.2006.11.025.

[4] M. Blum, "On the sums of independently distributed Pareto variates", *SIAM J. Appl. Math.* **19** (1970) 191–198; doi:10.1137/0119017.

[5] R. L. Carneiro, A. P. R. da Silva, V. F. de Magalhaes and S. M. F. de Oliveira e Azevedo, "Use of the cell quota and chlorophyll content for normalization of cylindrospermopsin produced by two *Cylindrospermopsis raciborskii* strains grown under different light intensities", *Ecotoxicol. Environ. Contam.* **8** (2013) 93–100; doi:10.5132/eec.2013.01.013.

[6] K. K. Cavender-Bares, A. Rinaldo and S. W. Chisholm, "Microbial size spectra from natural and nutrient enriched ecosystems", *Limnol. Oceanogr.* **46** (2001) 778–789; doi:10.4319/lo.2001.46.4.0778.

[7] P. G. Falkowski, "The role of phytoplankton photosynthesis in global biogeochemical cycles", *Photosyn. Res.* **39** (1994) 235–258; doi:10.1007/BF00014586.

[8] P. G. Falkowski and J. A. Raven, *Aquatic photosynthesis* (Princeton University Press, Princeton, NJ, 2007).

[9] Z. V. Finkel, "Light absorption and size scaling of light limited metabolism in marine diatoms", *Limnol. Oceanogr.* **46** (2001) 86–94; doi:10.4319/lo.2001.46.1.0086.

[10] M. J. Goovaerts, R. Kaas, R. J. Laeven, Q. Tang and R. Vernic, "The tail probability of discounted sums of Pareto-like losses in insurance", *Scand. Actuar. J.* **6** (2005) 446–461; doi:10.1080/03461230500361943.

[11] R. R. Guillard and M. S. Sieracki, "Counting cells in cultures with the light microscope", in: *Algal culturing techniques* (ed. R. A. Andersen), (Elsevier, New York, NY, 2005) 239–252; doi:10.1016/B978-012088426-1/50017-2.

[12] B. M. Hill, "A simple general approach to inference about the tail of a distribution", *Ann. Statist.* **3** (1975) 1163–1174; doi:10.1214/aos/1176343247.

[13] O. Holm-Hansen, C. J. Lorenzen, R. W. Holmes and J. D. Strickland, "Fluorometric determination of chlorophyll", *J. Cons. Int. Explor. Mer* **30**(1) (1965) 3–15; doi:10.1093/icesjms/30.1.3.

[14] B. Mandelbrot, "The stable Paretian income distribution when the apparent exponent is near two", *Internat. Econom. Rev.* **4** (1963) 111–115; doi:10.2307/2525463.

[15] C. M. Orrico, C. Moore, D. Romanko, A. Derr, A. H. Barnard, C. Janzen, N. Larson, D. Murphy, R. Johnson and J. Bauman, "WQM: a new integrated water quality monitoring package for long-

term in-situ observation of physical and biogeochemical parameters", *OCEANS 2007* **2007** 1–9; IEEE, doi:10.1109/OCEANS.2007.4449418.

[16] S. W. Philbrick, "A practical guide to the single parameter Pareto distribution", *PCAS LXXII* **44** (1985) 44–77; http://casualtyactuarialsociety.com/pubs/proceed/proceed85/85044.pdf.

[17] C. M. Ramsay, "The distribution of sums of certain i.i.d. Pareto variates", *Comm. Statist. Theory Methods* **35** (2006) 395–405; doi:10.1080/03610920500476325.

[18] C. M. Ramsay, "The distribution of sums of i.i.d. Pareto random variables with arbitrary shape parameter", *Comm. Statist. Theory Methods* **37** (2008) 2177–2184; doi:10.1080/03610920701882503.

[19] C. S. Reynolds, *The ecology of freshwater phytoplankton* (Cambridge University Press, UK, 1984).

[20] J. Rodriguez, F. Jimenez, B. Bautista and V. Rodriguez, "Planktonic biomass spectra dynamics during a winter production pulse in Mediterranean coastal waters", *J. Plankton Res.* **9** (1987) 1183–1194; doi:10.1093/plankt/9.6.1183.

[21] B. Roehner and P. Winiwarter, "Aggregation of independent Paretian random variables", *Adv. Appl. Probab.* **17** (1985) 465–469; doi:10.2307/1427153.

[22] R. W. Sheldon, A. Prakash and W. H. J. Sutcliffe, "The size distribution of particles in the ocean", *Limnol. Oceanogr.* **17** (1972) 327–340; doi:10.4319/lo.1972.17.3.0327.

[23] I. M. Suthers, C. T. Taggart, D. Rissik and M. E. Baird, "Day and night ichthyoplankton assemblages and zooplankton biomass size spectrum in a deep ocean island wake", *Mar. Ecol. Prog. Ser.* **322** (2006) 225–238; doi:10.3354/meps322225.

[24] B. Vidondo, Y. T. Prairie, J. M. Blanco and C. M. Duarte, "Some aspects of the analysis of size spectra in aquatic ecology", *Limnol. Oceanogr.* **42** (1997) 184–192; doi:10.4319/lo.1997.42.1.0184.

[25] I. V. Zaliapin, Y. Y. Kagan and F. P. Schoenberg, "Approximating the distribution of Pareto sums", *Pure Appl. Geophys.* **162** (2005) 1187–1228; doi:10.1007/s00024-004-2666-3.