12 IKC
12TH INTERNATIONAL
KIMBERLITE
CONFERENCE

30 YEARS OF DIAMONDS IN CANADA
8–12 July 2024 • Yellowknife

12th International Kimberlite Conference Extended Abstract No. 12IKC-161, 2024

Recovering Population Estimates through Expectation Maximization Techniques

Yuri Kinakin, Gus Fomradas
Rio Tinto Exploration Canada Inc; yuri.kinakin@riotinto.com, gus.fomradas@riotinto.com

Introduction

Compared to commodities such as base metals that are infinitely divisible and amenable to a pulverized assay approach, diamonds present a relatively unique sampling problem. The full size range of diamonds within a deposit is very unlikely to be recovered in a single sample. It is particularly important to understand that, as sample size decreases, the expectation of a sample grade will be lower than the expected value of the population. This is also commonly known as the "sample-size effect."

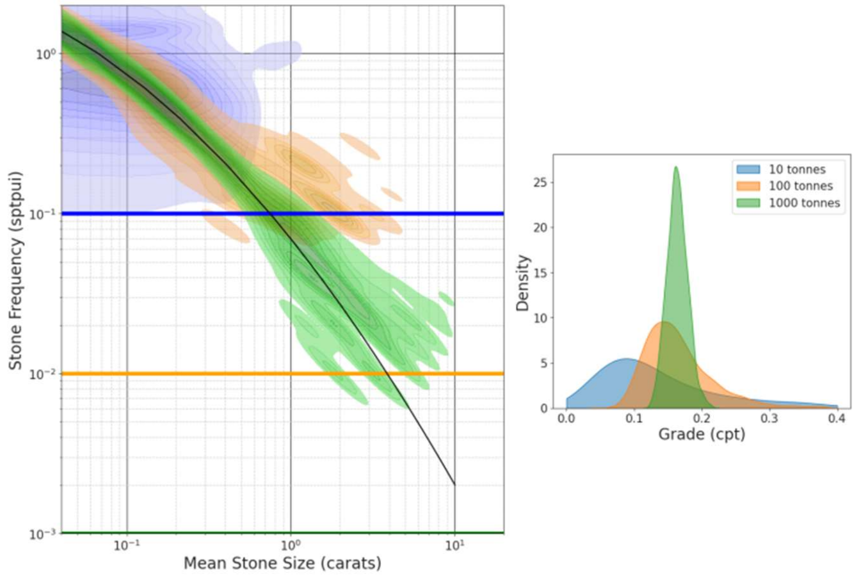

Figure 1: (right) Density plot of diamond size frequency distributions for 1000 realizations each of 10, 100 and 1000 tonne samples drawn from an a-priori distribution (grey line). The minimum value of 1spt is also plotted for each sample size. (left) Resultant simple grades for the same simulations. Note that the expected value for each distribution decreases with decreasing sample size.

As the value of a diamond is at least partially a function of its intact size, when evaluating a deposit it is therefore necessary to use some method to estimate the full size-frequency distribution (SFD). Two approaches used in industry are either to fit a mathematical model to those stones recovered by a sample and extend this to the full expected population, or to take an existing curve from a producing asset and fit that to sample data. (Burgess, 2018). Of the two approaches, the former is the more commonly applied and often the most appropriate in greenfield exploration where there are no nearby deposits against which to benchmark.

Motivation

The utilization of size-frequency distribution analysis and modelling to diamonds was first discussed in the scientific literature by Sichel, who proposed that the distribution of natural diamond distributions follows a long-tailed Poisson type (Sichel, 1973). This work was expanded to include examples of recovered SFDs and associated log-linear/quadratic models from a selection of diamond bearing rocks distributed around the globe (Davy, 1989).

One of the issues that practicioners need to consider when extending the SFD out of the region directly supported by sampling is the exact cut-off region for statistical significance. When a linear least squares (LLS) approach is used to fit a polynomial to the sample data, a judgement call must be made by a practitioner to remove any size classes that have the potential to provide high leverage. While using a mathematical curve fitting approach is often necessary to correct for the sample-size effect, a practioner can easily overestimate the grade and probability of recovery for large stones if unrepresentative size classes (i.e. nuggets) are not first removed before curve fitting. One solution is to use the fact that the poisson distribution tends towards the normal for larger stone counts so, for example, an arbitrary cut-off value (e.g. removed any size classes containing less than 30 stones) could be used. This is, however, not a hard rule and is left up to a qualified person to implement leading to variability and potential bias in estimates.

Expectation Maximization

An alternative to using LLS to fit a model is by utilizing an expectation-maximization (EM) approach, the benefit of which is a lower sentivitity to nuggets and outliers. This technique is well established in other fields and is referenced in Stichel's original paper on the subject of diamond estimation. Instead of directly fitting a curve to a set of sample points, a population SFD is assumed and the number of number of stones in each size class for a given tonnage is predicted. This prediction is then compared to the sample data, and any misfit is used to update the estimated population SFD. A small misfit on the coarse end of the curve does not overly weight the entire distribution, meaning that nuggets can be left in with minimal impact on the resultant population estimate. A convergence criteria is set and, once the mismatch between prediction and sample data falls below a set level, the population estimate is complete.

If an assumption is made that the population follows a 2nd order polynomial, then the formula for the population takes the form

$$\ln(\lambda_i) = \theta_0 + \theta_1 \ln\,(MSS_i) + \theta_2 \ln\,(MSS_i^2)$$

where λ_i is the expected value in stones-per-tonne for a given size class, MSS_i is the mean stone size for the size class in mm, and θ_0, θ_1 and θ_2 are the constants of the polynomial of the population.

The probability of the number of stones in each size class follows a Poisson distribution, the log-likelihood for which is given by:

$$l(\lambda, x_1, \dots, x_n) = -n\lambda + \ln(\lambda)\sum_{j}^{n} x_j - \sum_{j}^{n} \ln\,(x_j!)$$

where x_j is a single realization of stone counts for a sample. The summation reflects the fact that multiple samples are being recovered from the same population and can be used to increase the log likelihood.

By combining these two equations and differentiating, we are left with the following gradient functions to be applied during the maximization step:

$$\frac{\partial l}{\partial \theta_0} = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{i,j} - n\lambda_i\right)$$

$$\frac{\partial l}{\partial \theta_1} = \sum_{i=1}^{n} \ln(MSS_i)\left(\sum_{j=1}^{n} x_{i,j} - n\lambda_i\right)$$

$$\frac{\partial l}{\partial \theta_2} = \sum_{i=1}^{n} \ln\,(MSS_i^2)\left(\sum_{j=1}^{n} x_{i,j} - n\lambda_i\right)$$

At each iteration, the values of θ_0, θ_1 and θ_2 are adjusted down gradient until convergence. An algorithm was developed and implemented in Python that can be run on a standard laptop computer. A comparison of results between an R^2 minimization and an EM approach are shown graphically in Figure 2.

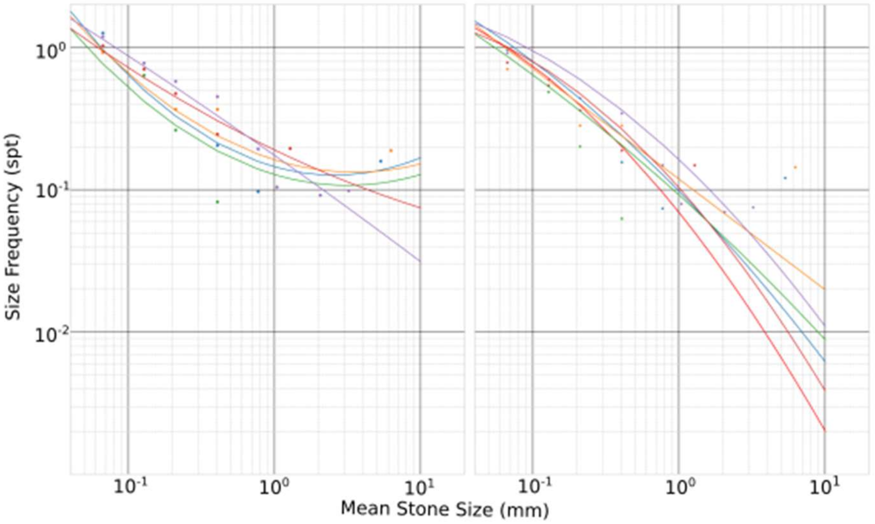

Figure 2: Comparison of R2 population estimates of a single sample (left) versus population estimates of the sample samples using an EM algorithm (right). The true population is shown in red on the right-hand graph.

As shown in Figure 2, the EM approach potentially allows a practitioner to include an entire sample in a model fitting without first applying a cut-off, though this approach is not without downsides. As it is an iterative algorithm with several variable criteria, convergence is not assured for all initial values so both these and an appropriate relaxation need to be appropriately chosen for the algorithm to converge. It is, nevertheless, another useful tool available to a resource geologist for estimating diamond populations from sample data.

References

Burgess, J., Buxton, N., Dyck, D., Oosterveld, M., Routledge, R., and Thurston, M. (2018). Estimation of mineral resources and mineral reserves best practices guidelines. Technical report, Canadian Institute of Mining, Metallurgy and Petroleum.

Davy, A. T. (1989). The Size Distribution of Diamonds in Kimberlites and Lamproites. PhD thesis, Imperial College London.

Sichel, H. (1973). Statistical valuation of diamondiferous deposits. Journal of the South African Institute of Mining and Metallurgy.