A truncated Pareto model to estimate the under recovery of large diamonds

Andréhette Verster [1], Daan de Waal [1], Robert Schall [1], Chris Prins [2]

Abstract The metallurgical recovery processes in diamond mining may under certain circumstances cause an under recovery of large diamonds. In order to predict high quantiles or tail probabilities we use a Bayesian approach to fit a truncated Generalized Pareto Type distribution to the tail of the data consisting of the weights of individual diamonds. Based on the estimated tail probability, the expected number of diamonds larger than a specified weight can be estimated. The difference between the expected and observed frequencies of diamond weights above an upper threshold provides an estimate of the number of diamonds lost during the recovery process.

Keywords Bayesian estimation • Bayesian prediction • Generalized Pareto Type distribution • tail probability • threshold

1 Introduction

The nature of metallurgical recovery processes in diamond mining may cause under recovery of large diamonds (that is, for diamonds above a certain large carat value, say between 30 and 60cts per stone). Diamonds not recovered by the mining process end up on the dumps, together with the tailings. Because of the potentially large monetary value of even a small number of large diamonds the question arises whether re-mining of a mine dump can be made more profitable by the recovery of such diamonds. To answer this question, estimation of the expected number of large diamonds is of interest.

In early, unpublished work done in the 1980's, Sichel, Kleingeld and Ravencroft (1980) (unpublished handwritten notes not available to us; personal communication) used the Double Truncated Pareto distribution to model the distribution of diamond weights. In the present paper we use a Bayesian approach to fit a truncated Generalized Pareto Type distribution to the tail of the data consisting of the weights of individual diamonds from a specific mine. The difference between the expected and observed frequencies of diamonds above an upper threshold provides an estimate of the number of diamonds lost during the recovery process. The method is applied to an observed data set from a diamond mine.

2 Generalized Pareto model

The cumulative distribution function (cdf) F(x) of the Generalized Pareto Type (PT) distribution for a random variable (in our application: diamond weights exceeding a lower threshold) is given by

(1)

The Generalized Pareto Type distribution is fitted to data exceeding a lower threshold . It has the advantage that, given the lower threshold, it has only one parameter, , which is the shape parameter or extreme value index of the distribution (Verster and De Waal 2009).

Any tail probability of the distribution can be estimated once is estimated. In the present work we estimate through a Bayesian approach, namely as the mode of the posterior distribution of , given a lower threshold and an ordered random sample above the threshold.

The posterior distribution of is derived using the maximal data information (MDI) prior (Zellner 1977)

(2)

The proof that [Eq. (2)] is the MDI prior uses the fact that , and that , where denotes the density function of .

Given a random sample above the threshold , the likelihood function is

(3)

so that the posterior density can be written as

(4)

The mode of the posterior distribution is taken as the estimate of .

To validate the choice of lower threshold we draw a quantile-quantile plot (QQ-plot) of the observed data against the estimated data at the empirical cdf probabilities

(Beirlant *et al*. 2004). For given , the estimated quantiles from (1) become

. (5)

A plot of against , yields the QQ-plot which is used to judge model fit: The graph should follow the 45^{o} line if the model fits well, and the correlation coefficient

is a goodness of fit statistic (Beirlant *et al*. 2004).

Example

The Generalized Pareto Type distribution was fitted to diamond size data from a diamond mine. To maintain confidentiality of the data a small lognormal error was added to data points, and a number of data values were randomly removed from the data set to veil the sample size. All conclusions made from the data remained unaffected by these modifications of the sample.

The modified sample contained diamond weights of 11cts or more. The lower threshold for fitting the Pareto Type of distribution was chosen as . With this threshold, the Bayesian estimate of is . Figure 1 shows the posterior distribution of with a mode of 0.44.

Figure 1 Generalized Pareto model: Posterior distribution of

(lower threshold)

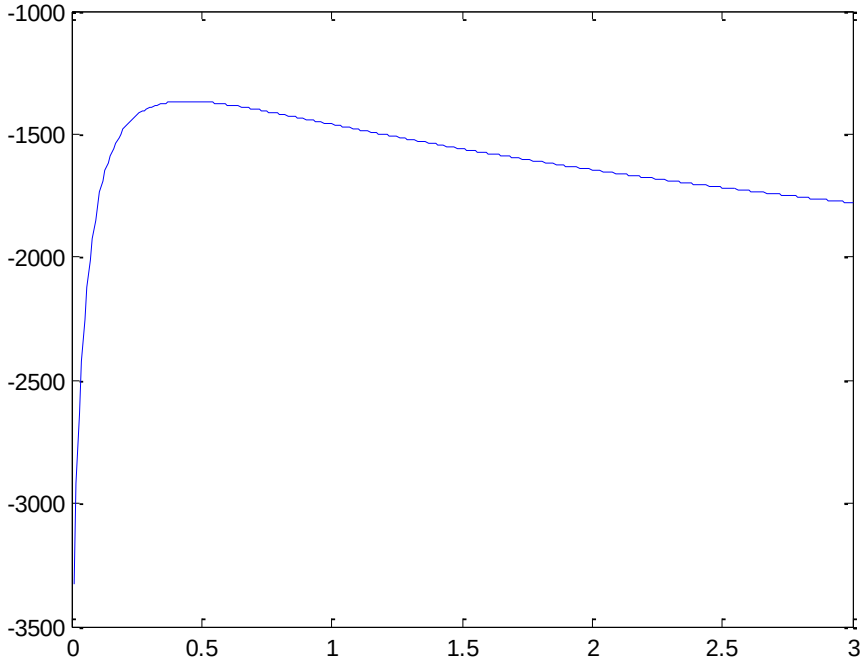
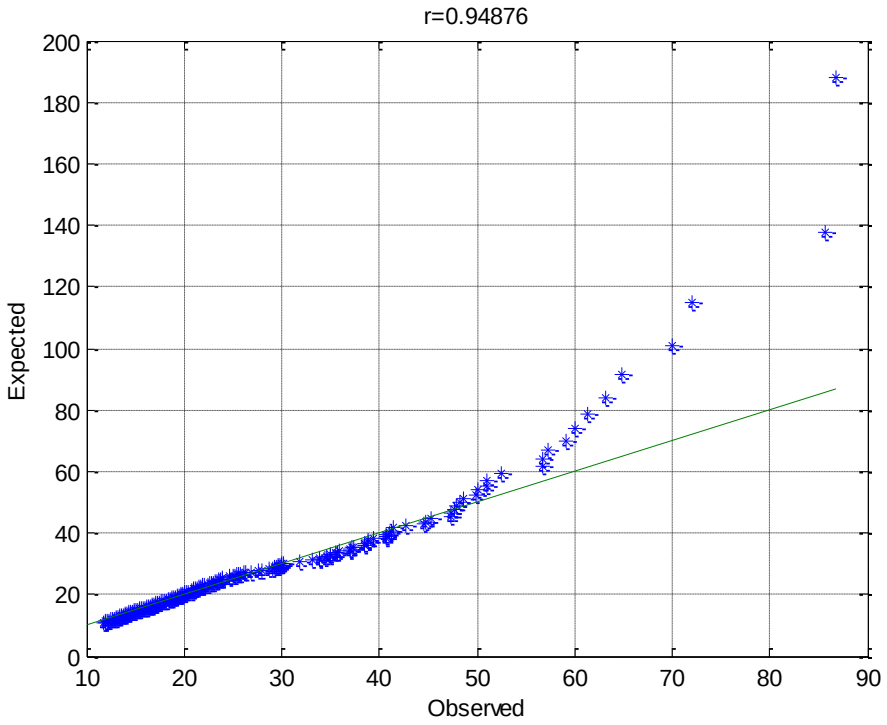


Figure 2 shows the QQ-plot which has a correlation coefficient of 0.9488. Inspection of the QQ-plot suggests that the Generalized Pareto Type distribution fits the data well up to diamond weights of about 45 to 50 cts. Thus the lower threshold seems appropriate, but above a weight of 50 cts the expected weights of diamonds are larger than the observed weights. The latter observation is a first indication of potential under

recovery of large diamonds. This aspect of the data set is explored further in the next section.

Figure 2 Generalized Pareto model: QQ-plot of observed vs. expected quantiles



The expected number of diamonds greater than a certain weight can be estimated by multiplying the tail probability with the sample size, , which is the number of observations larger than the lower threshold . Based on the estimate for of , the expected number of diamonds greater than 30cts and 60cts is estimated as 56.2 and 12.6 respectively. In contrast, the observed numbers of diamonds greater than 30cts and 60cts in the sample are 56 and 7 respectively.

3 Truncated Generalized Pareto model

As pointed out in section 2, the QQ-plot (Fig. 2) suggests that the Generalized Pareto Type distribution fits the data well for weights below 45 to 50 cts, but the fit is not good for weights above this range. Furthermore, based on the fit of the Generalized Pareto Type distribution to the whole data set above the lower threshold of the expected number of diamonds greater than 60cts is estimated as 12.6, while the observed number of diamonds greater than 60cts is only 7. This confirms our expectation that there is a certain loss of large diamonds in the recovery process. Therefore, only the first part of the data (diamonds with weights below a certain upper threshold) may be reliable, while diamonds weights above a certain upper threshold may constitute unreliable data in the sense that an unknown number of diamond weights are missing from the sample. This characteristic of the data suggests that the model should be fitted only to the range of diamonds weights representing reliable data.

In earlier work, Sichel, Kleingeld and Ravencroft (1980; personal communication) considered the Double Truncated Pareto model with the following probability density function

(6)

Here and are parameters of the distribution estimated from the double truncated data, and and are the lower and upper limit of truncation respectively. In the same spirit, we now adapt the Generalized Pareto Type model, presented in Section 2: We incorporate both a lower and an upper threshold for the data, and then fit the truncated Generalized Pareto distribution to the data between the thresholds.

Considering data truncated in the interval , where is the lower and the upper threshold, the truncated distribution function is

, (7)

Thus the posterior distribution for , from [Eq. (3)] above, is

(8)

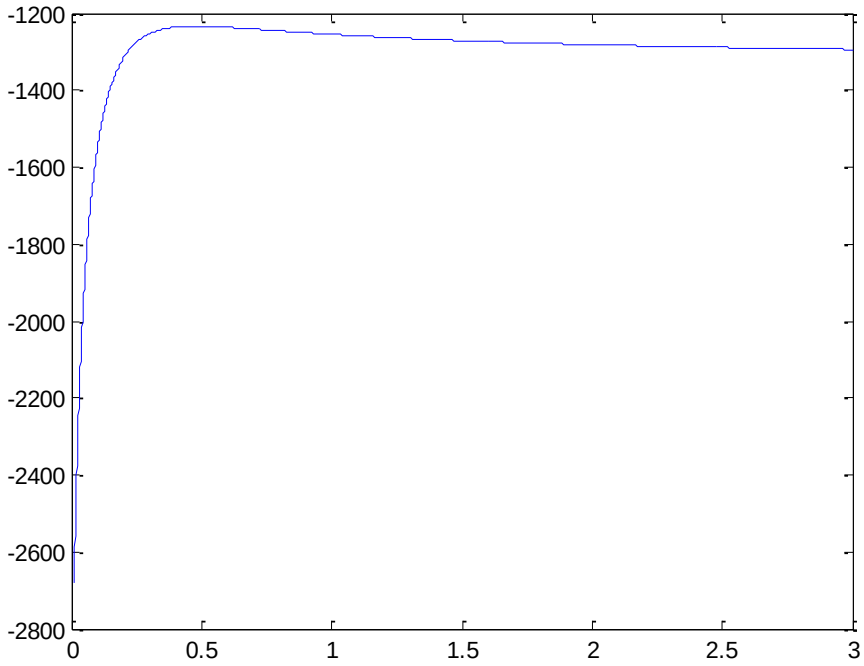
follows from [Eq. (1)] by replacing with . An estimate of is now obtained from the mode of (8). The quantile function from (7) becomes

(9)

Example (continued)

Considering the same data set as in the example of section 2, the lower threshold again is chosen as , and the upper threshold as . We therefore cater for under recovery of diamonds larger than 49cts. Under the truncated model, the Bayesian estimate of is . Figure 3 shows the posterior distribution of with mode 0.481.

Figure 3 Truncated Generalized Pareto model: Log of Posterior distribution for (lower threshold ; upper threshold)



The QQ-plot of the observed against the expected quantiles (Figure 4) indicates good model fit over the range of data to which the truncated Pareto Type model was fitted (between the lower threshold and the upper threshold).

Figure 4 Truncated Generalized Pareto model: QQ-plot of observed vs. expected quantiles

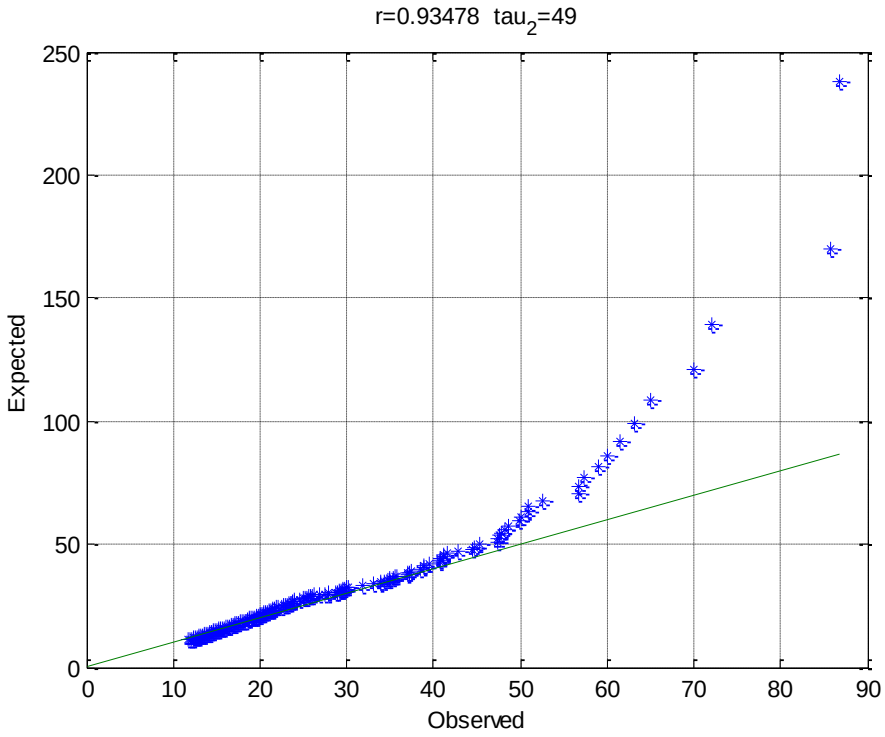
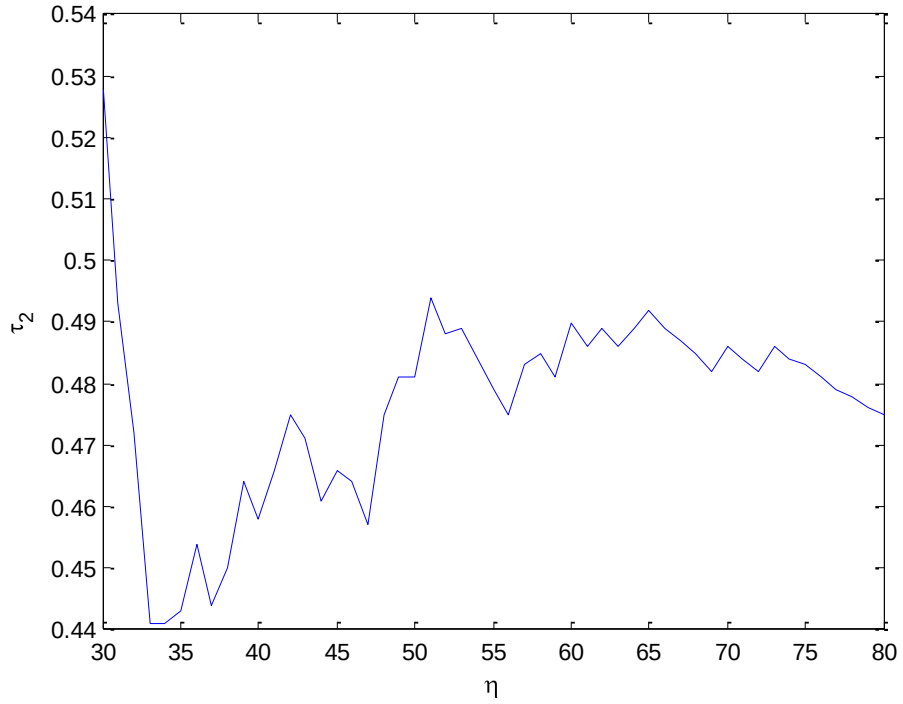


Figure 5 shows a plot of the estimated values for different choice of upper threshold . Figure 5 suggests that starts to stabilize around 0.48 for upper thresholds of 49 and greater. Therefore, choosing an upper threshold of 49 in this case seems appropriate.

Figure 5 Truncated Generalized Pareto model: Estimates of as a function of upper threshold



4 Estimating the number of diamonds not recovered

The expected number of diamonds larger than the upper threshold can be estimated either through a “plug-in” method, or though a prediction method, as discussed below in Section 4.1 and 4.2 respectively. The number of unrecovered diamonds above the upper threshold is then estimated as the difference between the expected and the observed number of diamonds.

4.1 Plug-in estimate

Let be the probability of a diamond weight being less than or equal to ,Furthermore, is the number of observed values in the sample below

(but greater than) and is the number of observed values above , so that is the total sample size. Since is unreliable (too small) because of potential under recovery of large diamonds, the total observed sample size is also unreliable (too small). The objective is to estimate the true as . Of course, if we have an estimate for , the estimate for is given as . The problem therefore reduces to the problem of estimating the total sample size of a binomial variate.

The number of diamonds below the upper threshold follows a Binomial distribution . Thus — for given and , but for a given estimate we can estimate as —. Thus an estimate for is given by

— — —— (10)

In Equation (10), is estimated as follows: first, we estimate from the data between the lower and upper thresholds using the truncated Generalized Pareto model, as described in section 3. The estimate for is then plugged into Equation (1), to obtain the estimate for . Finally, the estimate is plugged into Equation (10) to obtain . We refer to [Eq. (10)] as the “plug-in” estimate of .

Using [Eq. (10)] the estimated number of diamonds greater than 49cts is while the observed number of diamonds greater than 49cts is only . Therefore, the plug-in estimate suggests that there are about 8 more diamonds greater than 49cts to be recovered.

4.2 Predictive estimate

In this section the posterior prediction quantile function is used to predict the quantile, that is, the weight of a diamond, for a given small tail probability.

The posterior quantile function is given by

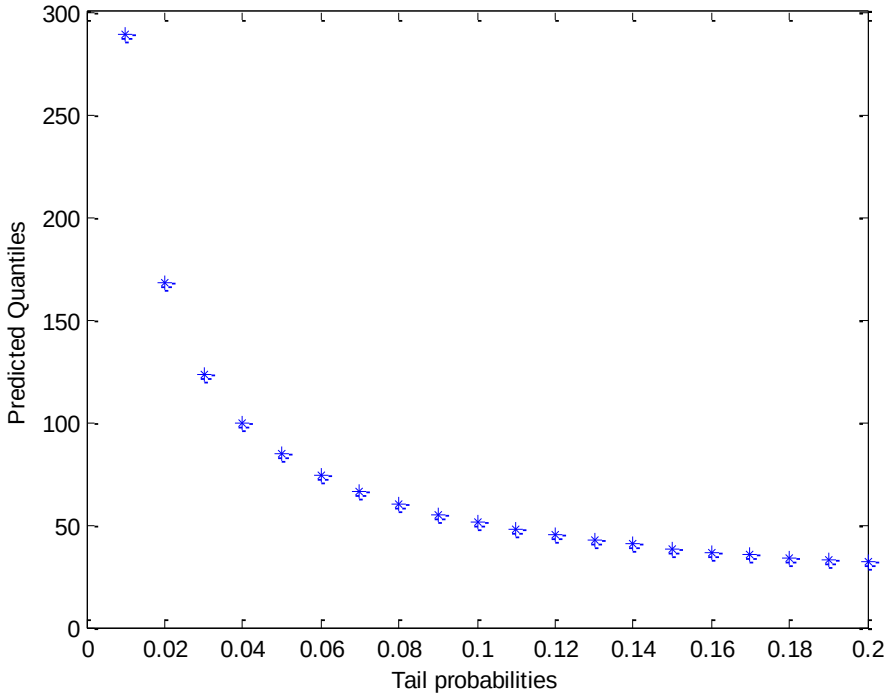
(11)

where denotes the tail probability and the quantile function [Eq. (5)]. Since integral [Eq. (11)] is impossible to evaluate analytically, is estimated as

– (12)

Eq. (12) is evaluated by simulating values of from its posterior distribution [Eq. (8)]. Note that the first and second moments of the Generalized Pareto Type distribution are only defined for and – respectively. Therefore, values of are simulated from the truncated posterior distribution, where (upper) truncation limit is less than or equal to 1. Figure 6 shows the predicted quantiles for various tail probabilities, estimated from 10 000 simulations where was simulated with an upper truncation limit of the posterior distribution of 1.

Figure 6 Truncated Generalized Pareto model: Predicted quantiles at various tail probabilities



Similarly, by simulating values from its posterior distribution [Eq. (8)], we can estimate the predicted tail probability for a future diamond weight as

– . (14)

Based on the data set used in sections 3 and 4 the posterior predicted tail probability of observing a diamond 49cts or larger is estimated as 0.0618 (10 000 simulations of with an upper truncation limit of the posterior distribution of 0.8).

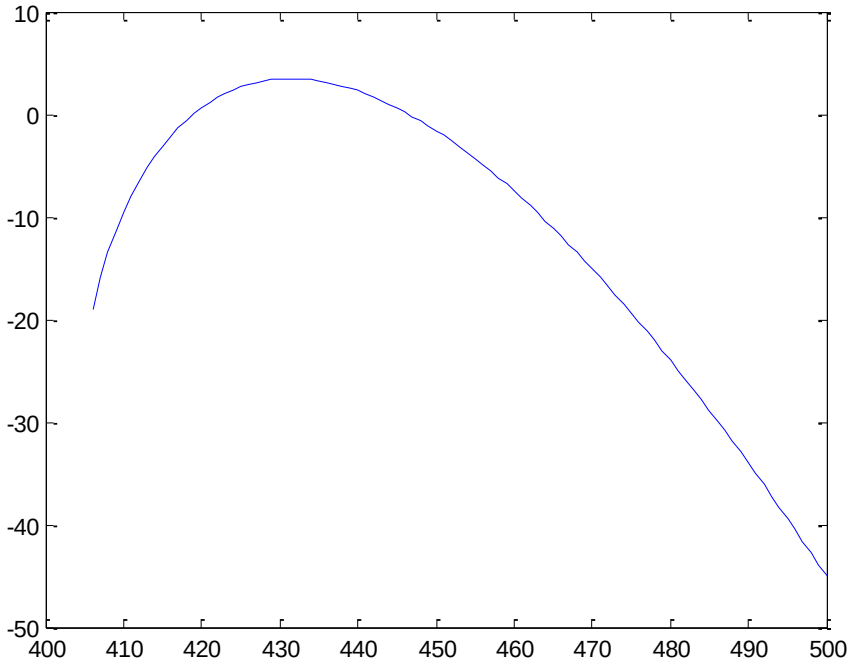
As pointed out in section 4.1, in order to estimate the number of diamonds above the upper threshold, the total sample size needs to be estimated. Again this becomes a Binomial estimation problem where the probability of success, , obtained from [Eq.

(14)], is 1 – 0.0618 = 0.9382, and the number of successes is for the example data set. If a non-informative uniform prior for defined on the parameter space is chosen, the posterior for is

(15)

Figure 7 shows the plot of the posterior of for and . From this posterior, the estimate of , taken as the mode, is . Thus the estimated number of diamonds greater than 49cts is , while only 17 of these diamonds were recovered. Therefore we estimate that another 9 diamonds could be mined which is consistent with the conclusion of Section 4.1.

Figure 7 Truncated Generalized Pareto model: Log of posterior of given and



We note that the predicted tail probability [Eq. (14)] is very sensitive to the chosen upper truncation limit for the posterior of . For example, if the upper bound is chosen as 1 instead of 0.8 the predicted tail probability is 0.0952 and the estimated number of diamonds greater than 49cts is 42. Table 1 shows the predicted tail probabilities and the estimated sample sizes greater than for different upper truncation limits for the posterior of .

Table 1 Estimated tail probability and total sample size for different upper truncation limits for the posterior of

upper bound		
0.3	0.0038	410
0.5	0.0211	414
0.8	0.0621	431
1	0.0954	448

Conclusions

The distribution of the weights of large diamonds can be modelled through a Generalized Pareto Type distribution fitted to the tail of the data. When it is known, or suspect, that large diamonds above a certain weight are under recovered in the mining process, the number of diamonds lost to tailings can be estimated by fitting a truncated Generalized Pareto Type distribution. In the latter case, both a lower threshold and an upper threshold for the distribution can relatively easily be incorporated into the model. The posterior predictive quantile function can be used to predict quantiles at various tail probabilities. The number of diamonds greater than the upper threshold can be estimated either through a plug-in or a prediction method.

References

Beirlant J, Goedgebeur Y, Segers J, Teugels J (2004) Statistics of extremes. Theory and applications. Wiley, Chichester

Verster A, De Waal DJ (2009) Approximating the Generalized Burr-Gamma with a Generalized Pareto-type of distribution. Technical Report 2/97, University of the Free State, Bloemfontein http://www.uovs.ac.za/faculties/documents/04/117/TechnicalReports/Teg397.pdf

Zellner A (1977) Bayesian analysis in econometrics and statistics. Edward Elgar, Lyme US.

Table 1 Estimated tail probability and estimated total sample size for different upper truncation limits for the posterior of

Truncation limit for		
0.5	0.0205	414
0.6	0.0335	420
0.7	0.0469	425
0.8	0.0620	432
0.9	0.0785	440
1	0.0952	448