Modeling of risk losses using size-biased data

E. Yashchin

In this paper we present a method for drawing inferences about the process of financial losses that are associated with the operations of a business. For example, for a bank such losses may be related to erroneous transactions, human error, fraud, lawsuits, or power outages. Information about the frequency and magnitude of losses is obtained through the search of a number of sources, such as printed, computerized, or Internet-based publications related to insurance and finance. The data consists of losses that were discovered in the search. We assume that the probability of a loss appearing in the body of sources and also being discovered increases with the magnitude of the loss. Our approach simultaneously models the process of losses and the process of populating the database. The approach is illustrated using data related to operational risk losses that are of special interest to the banking industry.

Introduction

Consider a business, such as a bank, that is interested in estimating the types of risks it faces. For example, banks have recently become very interested in estimating their exposure to *operational risk,* which includes almost all forms of risk except those related to financial markets and credit. A classification and explanation of these risks can be found in the description of the Basel II framework for international banking [1]. Of special relevance is Section V of this document, which provides a set of requirements to be met by a banking institution in order to prove to the Basel Committee that it "has an operational risk management system that is conceptually sound and is implemented with integrity." Furthermore, the framework specifies requirements for reporting losses and for self-assessment. It also offers one of three approaches for calculating operational risk losses, leaving a substantial degree of flexibility for the banking institution to account for business profiles of individual institutions.

Because the new regulations require banks to set aside resources to cover operational risk losses, the issue of risk estimation has become an important research subject. For example, methods for risk modeling, estimation, management, and hedging are considered in recent books [2–5]. An extensive summary of operational risk issues can be found in the January 2002 issue of the *Risk* journal; in particular, see [6]. A number of publications focus on methods emphasizing causal modeling and management of specific types of risks [7, 8]. Bayesian methods for risk modeling and estimation are discussed, for example, in [9, 10]. Statistical issues related to the estimation of losses are considered in [11, 12].

In this paper, we discuss an approach to modeling that is most appealing in the early phases of risk modeling, when reliable data is difficult to obtain and the existing data sources are known to be incomplete. Specifically, in the application that inspired this paper, we made use of a database that contained descriptions of operational losses suffered by various companies over a number of years. These losses were generally large, and entries related to the bank of interest itself were extremely rare. The relevant data can be found in [13]. This database was in the initial phase of construction and was thus known to be incomplete. We can safely assume that it referred to only a small fraction of losses suffered by various businesses.

The process of populating the database is typically focused on a certain *set of sources*. We assume that only losses that appeared or might have appeared in this set of sources are relevant. Our main problem of interest is how to use such a database to gain information about the stream of losses facing a given institution. A number of techniques have been used to increase the information content of databases containing rare events. For example, in many areas related to health and safety (such as the

0018-8646/07/$5.00 © 2007 IBM

chemical process industry) it is customary to report the "near-miss" incidents and use this information to enhance the inference related to risks, incident rates, and losses [14]. However, even with enhancements of this type, we expect that a substantial portion of risk-contributing events is likely to remain under-reported, especially in the early phase of populating the database. For successful inference under the stated conditions, one must be able to model the process of populating the database.

The data-collection process is probably the most challenging part of building an inference system related to operational losses. Of course, the issues of data quality and relevance are fundamental, and addressing these issues is a necessary condition for credible quantitative analysis of risks pertaining to a given institution. These issues have been discussed extensively in the literature [3]. To increase the amount of reliable data relevant to operational risk modeling, a number of data-sharing arrangements among banks have been established. However, the coverage of events related to operational risk is still quite limited, since a large number of such losses remain either unreported or not identified as being relevant for inclusion in a database. Furthermore, the relevance of the data that has already been collected tends to diminish with time because of factors such as inflation, major societal disruptions (such as the terrorism events of September 11, 2001), or new developments in the field of information technology. Therefore, data collection is an ongoing process and is itself exposed to a number of risks.

In this paper, we focus on one of these risks, namely, under-representation of losses due to the fact that sizable financial losses are generally more likely to appear in the set of sources, and be discovered, than losses of moderate or low magnitude. From the perspective of a given institution, this phenomenon occurs not only in the course of populating a database from sources that are external to this institution, but also in the process of populating internal databases. Though the Basel II framework mandates reporting of internal losses, it offers banks substantial flexibility on choosing reporting thresholds—and this in itself can create bias, even within the framework of a single institution. The problem of systematic biases in the data is of special importance in the environment in which data sources are assembled, at least partially, on the basis of automated text analysis of documents obtained via search of databases containing nonstructured data or searches of the Internet.

To further understand the concept of external and internal sources, consider an example of a particular bank. According to the Basel II regulations, this bank must maintain a record of all losses exceeding some threshold. The bank is not obliged to disclose this information to any external party, except for the Basel Committee auditors. This is referred to as an internal source. On the other hand, a large number of bank losses eventually appear in either publicly available sources or sources that are part of data-sharing consortia, and these are referred to as "external sources."

Similarly to the approach presented in this paper, one may approach the problem of estimating operational risk losses of medium and large magnitudes in three stages. The initial goal (stage 1) is to develop methods for characterizing the stream of losses related to the operations of both financial and nonfinancial institutions, observed worldwide, as well as the magnitudes of these losses; this involves drawing inferences about the hidden population of losses that are not represented in the database. The latter goal is achieved by introducing a concept called the discovery probability curve (DPC), which specifies the probability that a loss of a given size will enter the body of sources and be discovered. This probability curve can itself be subject to an estimation effort. Subsequently, in the second stage, one may estimate the fraction of these losses that is related to financial institutions. Finally, in the third stage, one can use this model, in conjunction with characteristics of the specific bank of interest, to estimate model parameters that relate to the stream of operations-related losses for this bank. Though we present the general structure for such causal modeling in the last section of this paper, the focus of the paper is on problems related to the first-stage methods for characterizing losses.

The proposed modeling of risks is useful in several respects. For example, it can be used by a bank, in conjunction with analysis of internal losses, as a basis for reserving capital needed to cover operational losses for a given period. Also, it can be used by insurance companies to assess the risk (or specific types of risks) related to the bank and to establish premiums. Although modern banks are usually self-insured with respect to operational losses, in the future some banks may prefer to mitigate the effects of these types of losses through insurance companies. For example, Financial Institution Operating Risk Insurance (FIORI) is currently being offered by the Swiss Re company, one of the world's largest reinsurers [15]. While some classes of risks may be good candidates for insurance coverage, other types of risks could be mitigated by service agreements and risk-sharing arrangements with other companies. In essence, we are considering here a situation faced by every newly emerging branch of insurance when data is sparse and expensive to collect and risks are poorly understood. The current literature related to actuarial science does not appear to provide an agreed-upon statistical methodology for the establishment of a new area of risk analysis or insurance. In this work, we attempt to formulate a framework that may be helpful in the development of such methodology.

In the next section, we describe our basic approach. Subsequently, we consider the problem of estimating model parameters and the use of goodness-of-fit tests for various aspects of the model. The fourth section considers models that are focused on losses that exceed a pre-specified threshold and discusses tail-based inference for such models in the presence of size-biased sampling. Finally, we discuss examples, generalizations, and directions of future research.

The basic approach

Estimation of the properties of hidden populations has been considered in the literature in relation to such areas as demography (e.g., population size estimation [16]), software reliability (estimation of the number of software defects hidden in programming code [17]), or nondestructive evaluation (inferences related to hidden defects [18]). The corresponding techniques are referred to in the statistical literature as *size-biased sampling*. What makes the present problem special is its strong actuarial aspect: The questions that are asked in the current context are much different from those asked in the areas mentioned above. These questions, in turn, determine the tools used in the statistical analysis.

We now introduce the approach for addressing questions arising in the context of risk estimation. Our basic assumptions are as follows:

- The process of losses is modeled as a homogeneous Poisson distribution with a rate of λ events per year.
- The underlying distribution of loss magnitudes is described by some density $f(x)$ that belongs to one of the families that are typically used to describe distribution of losses [5, 19]. For example, the Pareto, Weibull, or lognormal families can be considered good candidates.
- If a loss of magnitude x occurs, its probability of being discovered in the process of populating a database is $p(x)$, where p is a monotone function with a value from 0 to 1. In essence, we demand that $p(x)$ satisfy the properties of a cumulative distribution function (cdf). Some considerations that might be instrumental in selecting the suitable form of $p(x)$ are given below. We henceforth frequently refer to this function as the *discovery probability curve,* or DPC.

We note that in more complex applications, the rate λ, as well as the parameters associated with the distribution of losses and the DPC that corresponds to these losses, will depend on a set of factors, as is discussed in the last section.

To illustrate the basic ideas of our approach, we briefly discuss the data given in Appendix A of [13]. This data

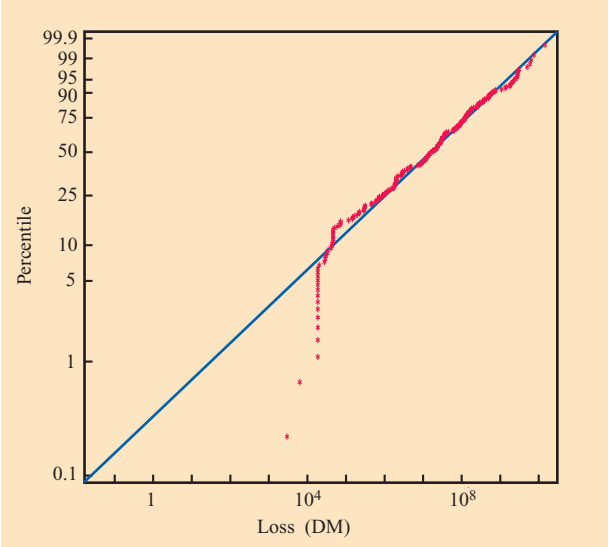

Figure 1

Observed losses (in Deutsche marks) from Appendix A in [13] plotted on the Weibull probability plot. The straight line corresponds to the parameter values ($\hat{c} = 0.32$, $\hat{b} = 4.9 \times 10^7$).

contains records of 226 losses assembled from public sources (e.g., news reports) and their degree of relevance to the banking industry. The losses are in Deutsche marks (DM). In 1998, when the database was populated, the exchange rate was approximately one U.S. dollar (1 USD) = 1.8 DM. Record No. 12 (a 177-DM loss by AVA, or Asesores de Valores, the institution that suffered the loss) is considered an outlier and it is not used in the analysis. Therefore, our data consists of 225 observations.

In our initial analysis, we subdivided the data randomly into two parts, the learning sample and the test sample. All of the methods discussed below were first applied to the learning sample and then validated on the test sample. In this paper, however, we show results only for the overall sample. Among several candidate distributions used by practitioners for describing the magnitude of losses (e.g., see [5, 19]), we considered the Weibull and Pareto distributions. Because our sample size was very small, the simplicity of these distributions was a strong factor in our decision to use them in order to avoid well-known data-analysis problems such as overfitting, with the resulting loss of predictive ability. In **Figure 1** we show the observed losses on the Weibull probability plot. The Weibull cdf $F(x)$ and density $f(x)$ are given by

$$F(x) = 1 - \exp[-(x/b)^c], \quad x > 0;$$
$$f(x) = (c/b)(x/b)^{c-1}\exp[-(x/b)^c], \tag{1}$$

and the estimated parameters of the "law" (i.e., the

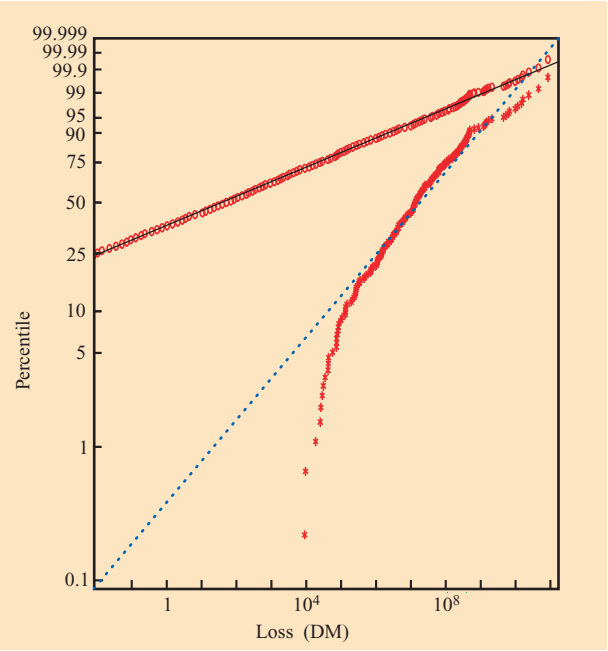

Figure 2

Simulated replica (asterisks) of the data obtained by combining Weibull-distributed losses (circles) with parameters $c = 0.133$, $b = 26{,}900$ (i.e., $u_1 = 10.2$, $u_2 = 7.5$, which corresponds to the solid straight line) and a logistic DPC with parameters $(v_1, v_2) = (14, 1.7)$. The dotted straight line corresponds to ($\hat{c} = 0.31$, $\hat{b} = 5.9 \times 10^7$).

Weibull cdf) are $\hat{c} = 0.32$ and $\hat{b} = 4.9 \times 10^7$. From Figure 1 it may initially appear that the distribution is consistent with this law, except that the smaller losses are missing, presumably because it is difficult for such losses to enter the set of sources and to be discovered.

Upon closer inspection, however, such a simplistic explanation becomes unsatisfactory. Suppose that the population of losses is indeed distributed in accordance with the above Weibull law. Then the fraction of operational losses below 10^8 DM $\approx$ \$55M, or 55 million U.S. dollars, in the overall population is estimated to be 75%, which appears to be much too small, given that the data collection effort was not limited *a priori* to sources that focus only on low-frequency and high-impact events.

Despite these reservations, the fact that the Weibull probability plot is linear in the upper tail (i.e., for large observed losses) suggests that the upper tail of the distribution may indeed be Weibull (albeit with different parameters) and, with a suitably chosen and plausible DPC, $p(x)$, we may obtain results that are consistent with the data.

To illustrate this point, we may switch to logarithmically transformed data. In the discussion that follows, we work primarily with the observations $y = \ln(x)$. If the losses are distributed in accordance with Equation (1), the cdf and density of log losses are

$$\tilde{F}(y) = 1 - \exp\{-\exp[(y - u_1)/u_2]\},$$
$$\tilde{f}(y) = u_2^{-1} \exp\{-\exp[(y - u_1)/u_2] + (y - u_1)/u_2\}. \quad (2)$$

[Note that throughout this paper, the tilde (~) indicates quantities associated with log losses.] It is easy to see that

$$u_1 = \ln(b), \quad u_2 = 1/c. \quad (3)$$

Now let us define the DPC in terms of a logistic curve:

$$\tilde{p}(y) = \{1 + \exp[-(y - v_1)/v_2]\}^{-1}. \quad (4)$$

We select the DPC parameters $(v_1, v_2) = (14, 1.7)$, using some prior expectation based on knowledge about the data-collection mechanism, and we estimate the Weibull loss distribution parameters on the basis of this selection, using methods described in subsequent sections. We next use the resulting estimated parameters $(u_1, u_2) = (10.2, 7.5)$, jointly with $(v_1, v_2) = (14, 1.7)$, to simulate the process of losses and discovery. The resulting plot is shown in **Figure 2**. One can see that the simulated loss data is quite similar to that presented in Figure 1. Now, for the underlying Weibull law, the fraction of operational losses below 10^8 DM in the overall population is estimated to be 0.96, which is more consistent with our expectations. Notice that the fitted distribution is defined for losses as small as 1 DM, indicating that this model extrapolates far beyond the range of the losses that were actually observed. This type of extrapolation does not normally interfere with the estimation of quantities important for decision-making, such as those associated with value-at-risk (VAR), which is a well-known category of risk metric [5]. An immediate question arises as to whether the particular distributions selected above are plausible and do not conflict with actual measurements. In the following sections, we develop methods for answering such questions and for fitting models of this type.

While the selection of the distribution of loss magnitudes can be motivated to some extent by models prevalent in the actuarial literature, the selection of the DPC is more problematic. On the basis of experience so far, it appears reasonable to require that the DPC curve satisfy the following relation: For every $\eta > 0$ there exists x_η so that

$$1 - p[(1 + \eta)x] \geq [1 - p(x)]^{1+\eta} \quad (5)$$

for all $x \geq x_\eta$. For example, let us select $\eta = 1$ and consider an event resulting in a loss of some high magnitude corresponding to a value of $2x$. Let us suppose

that this loss can be documented in one of two ways: 1) as a single loss of magnitude $2x$, or 2) as a pair of losses of magnitude x each, which are introduced independently into the body of sources. In the latter case, the event will be discovered if at least one of the pair of losses is discovered. The criterion in Equation (5) states that the probability of overlooking the event under the first scenario is greater than the probability of overlooking it under the second scenario.

Of special interest are DPC curves that satisfy Equation (5) for every $\eta > 0$ and $x > 0$. After performing algebraic manipulation, one can show that two curves of this type are

$$p_1(x) = 1 - \left\{1 + [(x-\zeta)/h]^r\right\}^{-1}, \quad x > \zeta, \tag{6}$$

with shift, scale, and shape parameters ($\zeta > 0$, $h > 0$, $0 < r \leq 1$), and

$$p_2(x) = 1 - \exp\left\{-[(x-\zeta)/h]^r\right\}, \quad x > \zeta \tag{7}$$

(i.e., a shifted Weibull distribution) with the shape parameter $0 < r \leq 1$. In what follows, we assume that $\zeta = 0$. Under this assumption, the function $p_1(x)$ results, in terms of log losses y, in the logistic distribution in Equation (4) with $v_1 = \ln(h)$, $v_2 = 1/r$. The function $p_2(x)$ results in the distribution

$$\tilde{p}_2(y) = 1 - \exp\left\{-[\ln(2)] \times \exp[(y - v_1)/v_2]\right\}, \tag{8}$$

with parameters $v_1 = \ln[\ln(2)/r] + \ln(h)$ and $v_2 = 1/r$. The correction factor $\ln[\ln(2)] \approx -0.3665$ is introduced in order to ensure that v_1 is the median point of the DPC $\tilde{p}_2(y)$.

Note that distributions (7) and (8) both belong to a *location-scale family* with location and scale parameters v_1 and v_2, respectively. Our discussion is limited to the case of the logistic DPC $\tilde{p}(y)$ given in Equation (4).

The estimation problem

The density of log losses is represented by $\tilde{f}(y|\boldsymbol{u})$, and $\tilde{p}(y|\boldsymbol{v})$ is the DPC; $\boldsymbol{u}$ and $\boldsymbol{v}$ are the corresponding vectors that describe the density of log losses and DPC, respectively. The density of a log loss y that is conditional on this loss appearing in the body of sources and being discovered in a source is given by

$$\tilde{f}_c(y|\boldsymbol{u},\boldsymbol{v}) = \left[\tilde{f}(y|\boldsymbol{u})\tilde{p}(y|\boldsymbol{v})\right]/C(\boldsymbol{u},\boldsymbol{v}), \tag{9}$$

where the mean value of the discovery probability, represented by the normalizing constant $C(\boldsymbol{u}, \boldsymbol{v})$, is given by

$$C(\boldsymbol{u},\boldsymbol{v}) = \int_{-\infty}^{\infty} \tilde{f}(y|\boldsymbol{u})\tilde{p}(y|\boldsymbol{v})dy. \tag{10}$$

Suppose that the overall number of losses recorded in the set of sources is N and the actual number of discovered losses is k; the corresponding log losses are $y_1, y_2, \cdots, y_k$. Our challenge is to estimate the parameters $\boldsymbol{u}$, $\boldsymbol{v}$ and N.

Likelihood-based estimation

The log-likelihood of the observed data is given by

$$\begin{aligned} L(\boldsymbol{u}, \boldsymbol{v}, N|y_1, y_2, \cdots, y_k) &= \ln\binom{N}{k} + k \times \ln C(\boldsymbol{u}, \boldsymbol{v}) + (N-k) \times \ln[1 - C(\boldsymbol{u}, \boldsymbol{v})] \\ &\quad + \sum_{i=1}^{k} \ln\left[\tilde{f}_c(y_i|\boldsymbol{u}, \boldsymbol{v})\right], \end{aligned} \tag{11}$$

where the first three terms (that is, all items between the = symbol and the summation) are related to the binomial probability of discovering k losses, and the last term represents magnitudes of log losses, conditional on being discovered. The inference can now be based on this log-likelihood. When nothing else is known about the parameters, one can derive the maximum likelihood estimators (MLEs) by finding the parameters that maximize Equation (11). In this paper, we do not perform such a likelihood analysis; instead, we work with a somewhat simplified form of the likelihood that arises when it is known *a priori* that N is large and C is small. The presented approach adequately represents the main ideas and is sufficiently accurate to address the problems that motivated this research. Analysis of the exact likelihood in Equation (11) can be performed in a similar way.

When it is known *a priori* that N is large and only a small fraction of the losses have been discovered, one can approximate the binomial term in Equation (11) by the corresponding Poisson term. The approximate log-likelihood becomes

$$\begin{aligned} L_1(\boldsymbol{u}, \boldsymbol{v}, N|y_1, y_2, \cdots, y_k) &\approx \ln\left\{\frac{[NC(\boldsymbol{u},\boldsymbol{v})]^k e^{-NC(\boldsymbol{u},\boldsymbol{v})}}{k!}\right\} \\ &\quad + \sum_{i=1}^{k} \ln\left[\tilde{f}_c(y_i|\boldsymbol{u},\boldsymbol{v})\right]. \end{aligned} \tag{12}$$

In the process of maximum likelihood estimation, one can take advantage of the fact that for given $\boldsymbol{u}$ and $\boldsymbol{v}$ the likelihood is maximized when

$$N = \hat{N} = k/C(\boldsymbol{u}, \boldsymbol{v}), \tag{13}$$

indicating that one can expect to obtain estimates of good quality based on the conditional distribution of the observed losses only. For values of C that are not very small, the estimate in Equation (13) is still quite sensible, and it can be substituted into Equation (11) to yield (after using Stirling's expansion and some algebra) an approximate likelihood

$$L_2(\boldsymbol{u}, \boldsymbol{v}, N = \hat{N} | y_1, y_2, \cdots, y_k)$$
$$= \ln\left(\frac{k^k e^{-k}}{k!}\right) + \sum_{i=1}^{k} \ln\left[\tilde{f}_c(y_i|\boldsymbol{u}, \boldsymbol{v})\right] - 0.5 \ln[1 - C(\boldsymbol{u}, \boldsymbol{v})] - \frac{C^2(\boldsymbol{u}, \boldsymbol{v})}{12k[1 - C(\boldsymbol{u}, \boldsymbol{v})]} + O\left[\frac{C^3(\boldsymbol{u}, \boldsymbol{v})}{k^2}\right]. \quad (14)$$

Such an estimation technique is similar to the so-called "profile likelihood" methodology. In most practical applications, the second term dominates the last three, and we can, once again, obtain parameter estimates of good quality based on this term only. We restrict ourselves to using estimates based on the second term of Equation (12).

After substituting Equation (13) into Equation (12), we can obtain the estimates by solving the gradient equations,

$$\begin{cases} \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = k \times \dfrac{\nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v})}{C(\boldsymbol{u}, \boldsymbol{v})}, \\ \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{v}} \tilde{p}(y_i|\boldsymbol{u})}{\tilde{p}(y_i|\boldsymbol{u})} = k \times \dfrac{\nabla_{\boldsymbol{v}} C(\boldsymbol{u}, \boldsymbol{v})}{C(\boldsymbol{u}, \boldsymbol{v})}. \end{cases} \quad (15)$$

The equations (15) can be solved by using a simple iterative scheme that starts from some initial values $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)})$ and proceeds using the following estimation procedure:

- *Step 1:* For the current values $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^{(i)}, \boldsymbol{v}^{(i)})$, compute $C(\boldsymbol{u}, \boldsymbol{v})$ and its gradient vectors by $\boldsymbol{u}$ and $\boldsymbol{v}$, $\nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v})$ and $\nabla_{\boldsymbol{v}} C(\boldsymbol{u}, \boldsymbol{v})$.
- *Step 2:* Substitute the resulting values in the right-hand side of Equations (15) and solve the two groups of equations separately. Assign the solutions to $[\boldsymbol{u}^{(i+1)}, \boldsymbol{v}^{(i+1)}]$.
- *Step 3:* Iterate Step 1 and Step 2 until the convergence occurs. Accept the result if it passes tests for local optimality, sanity (i.e., plausibility) and goodness of fit, as described later.

It is important to note that the above procedure does not guarantee that convergence will occur. Furthermore, the tests in Step 3 are essential because even if convergence occurs, the limiting point is not guaranteed to be a local maximum of the approximate log-likelihood in Equation (12). Finally, the limiting point is not guaranteed to correspond to a global maximum. This correspondence would guarantee asymptotically optimal behavior of the resulting estimates, given that we are dealing primarily with densities that conform to the so-called *regularity conditions* [20, Sec. 6] and smooth DPCs. Our experience with losses distributed according to Weibull or Pareto distributions, in conjunction with a logistic DPC in Equation (4), suggests that the above estimation procedure is reliable for these distribution families. We did not observe the procedure to fail, using either real or simulated data. In these cases, we also did not see evidence of multiple maxima, despite the fact that the likelihood function is definitely not log-concave.

The tests for "sanity" mentioned in the above procedure are needed because a) it may be quite difficult to foresee the implications of mis-specifying the shape and/or parameters of the DPC, especially for small sample sizes, and b) the solution of Equation (15) maximizes the approximate log-likelihood in Equation (12) and not the exact likelihood, Equation (11). Therefore, if for the resulting estimates $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ the value $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ is not small enough to justify the Poisson approximation used to obtain Equation (15), this solution should be considered suspicious. In such situations, one can expect that solving the equations based on L_2 given by Equation (14), or even solving the exact profile likelihood equations, will also result in a relatively large value of $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$. In many practical situations such an estimate would be considered implausible, since one would generally expect that the process of populating a database is capable of exploring only a small fraction of the body of sources, and that even within this fraction most of the losses would remain undetected.

Therefore, failure of the equations (15) to produce a value of discovery probability, $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$, that is small enough to be compatible with one's expectation indicates that the estimation based on a full optimization approach may be inadequate, and some additional restrictions on parameters are necessary.

Once the estimates $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ have been obtained, the estimate $\hat{N}$ is obtained by substituting these values into Equation (13).

Constrained estimation and inference

Many problems related to the model described above involve maximization of the likelihood function in the presence of some constraints on the parameters. For example, after obtaining the ML estimates, one may decide that the resulting value of $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ is too high to be plausible, and carry out estimation under the constraint

$$C(\boldsymbol{u}, \boldsymbol{v}) = c_0, \quad (16)$$

where c_0 is chosen to represent the highest value of $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ that one is comfortable using. The estimation can be carried out by introducing a Lagrange multiplier β associated with this constraint and finding the stationary point of the Lagrangian,

$$L_\beta(u, \boldsymbol{v}, N|y_1, y_2, \cdots, y_k) = L(\boldsymbol{u}, \boldsymbol{v}, N|y_1, y_2, \cdots, y_k) - \beta[\ln C(\boldsymbol{u}, \boldsymbol{v}) - \ln c_0], \quad (17)$$

by solving the gradient equations

$$\begin{cases} \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = (k+\beta) \times \dfrac{\nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v})}{C(\boldsymbol{u}, \boldsymbol{v})}, \\ \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{v}} \tilde{p}(y_i|\boldsymbol{u})}{\tilde{p}(y_i|\boldsymbol{u})} = (k+\beta) \times \dfrac{\nabla_{\boldsymbol{v}} C(\boldsymbol{u}, \boldsymbol{v})}{C(\boldsymbol{u}, \boldsymbol{v})}, \\ C(\boldsymbol{u}, \boldsymbol{v}) = c_0. \end{cases} \quad (18)$$

The above equations can be solved by repeating, for various values of β, the process similar to the procedure described in the previous section until a value of β is found for which the constraint in Equation (16) is satisfied. The details of this algorithm are omitted for brevity.

Constrained optimization may also be used in conjunction with the likelihood analysis when one is willing to assume that some components of the parameters are known. This leads to a reduced system in Equation (15) that contains only the equations corresponding to unknown parameters. This system can be solved by using a procedure of the type described in the previous section. For example, under the assumption that the vector $\boldsymbol{v}$ that characterizes the DPC is known and equal to $\boldsymbol{v}_0$, the estimation process involves solving the system

$$\sum_{i=1}^{k} \frac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = k \times \frac{\nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v}_0)}{C(\boldsymbol{u}, \boldsymbol{v}_0)}. \quad (19)$$

Constrained estimation also plays an important role in inference related to the parameters of interest. For example, let us assume that $\boldsymbol{v}$ is known and equal to $\boldsymbol{v}_0$, and one is interested in testing the hypothesis H_0: $C(\boldsymbol{u}, \boldsymbol{v}_0) = c$ against the alternative $C(\boldsymbol{u}, \boldsymbol{v}_0) < c$, at the significance level γ. To achieve this goal, we can compute the maximum value of the log-likelihood under the constraint $C(\boldsymbol{u}, \boldsymbol{v}_0) = c$ (where we denote the constrained and unconstrained estimates by $\hat{\boldsymbol{u}}_c$ and $\hat{\boldsymbol{u}}$, respectively) and reject H_0 if $C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0) < c$ and if

$$\Psi_c(c) \stackrel{\text{def}}{=} 2\{L[\hat{\boldsymbol{u}}, \boldsymbol{v}_0, k/C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0)] - L[\hat{\boldsymbol{u}}_c, \boldsymbol{v}_0, k/C(\hat{\boldsymbol{u}}_c, \boldsymbol{v}_0)]\} > \chi^2_{1-\gamma}(1). \quad (20)$$

[Note that Equation (20) is both a definition and a condition, and that "(1)" refers to one degree of freedom of the chi-square distribution.] Furthermore, confidence bounds are obtained simply by collecting values that are not rejected by the corresponding test. For example, the value of c in the domain $c > C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0)$ for which the inequality in Equation (20) becomes an equality represents a $(1-\gamma) \times 100\%$ upper confidence bound for $C(\boldsymbol{u}, \boldsymbol{v}_0)$. As usual, a two-sided $(1-\gamma) \times 100\%$ confidence interval is obtained by combining lower and upper $(1-\gamma/2) \times 100\%$ confidence bounds. It is important to note that the threshold $\chi^2_{1-\gamma}(1)$ is based on the asymptotic theory, and its adequacy for small sample sizes has to be confirmed, for example, by using a simulation study.

Likelihood-based inference about $C(\boldsymbol{u}, \boldsymbol{v})$ does not lead directly to inference about N. In particular, if $(\underline{C}, \overline{C})$ is the $(1-\gamma) \times 100\%$ confidence interval for C, then $(k/\overline{C}, k/\underline{C})$ does not provide enough coverage to serve as the $(1-\gamma) \times 100\%$ confidence interval for N; however, these bounds are useful as initial points in the numeric procedure described below. To test the hypothesis that $N = n$ against the alternative $N < n$, we must compute the maximum value of the log-likelihood function under the constraint $N = n$. As can be seen from Equation (12), this goal can be achieved by solving the gradient equations

$$\begin{cases} \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{u}} \tilde{f}(y_i|\boldsymbol{u})}{\tilde{f}(y_i|\boldsymbol{u})} = n \times \nabla_{\boldsymbol{u}} C(\boldsymbol{u}, \boldsymbol{v}), \\ \sum_{i=1}^{k} \dfrac{\nabla_{\boldsymbol{v}} \tilde{p}(y_i|\boldsymbol{u})}{\tilde{p}(y_i|\boldsymbol{u})} = n \times \nabla_{\boldsymbol{v}} C(\boldsymbol{u}, \boldsymbol{v}) \end{cases} \quad (21)$$

by using a suitably modified estimation procedure of the type described in the previous section. Denote the constrained estimates by $(\hat{\boldsymbol{u}}_n, \hat{\boldsymbol{v}}_n)$ and the score associated with N by

$$\Psi_N(n) = 2[L(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \hat{N}) - L(\hat{\boldsymbol{u}}_n, \hat{\boldsymbol{v}}_n, n)]. \quad (22)$$

Then the hypothesis is rejected if $\hat{N} > n$ and $\Psi_N(n) > \chi^2_{1-\gamma}(1)$. The lower $(1-\gamma) \times 100\%$ confidence bound for N is then the value of $n < \hat{N}$ for which $\Psi_N(n) = \chi^2_{1-\gamma}(1)$. The upper bound is similarly obtained.

Goodness-of-fit tests

The fact that we have successfully obtained estimates of the basic parameters does not have great utility unless the data is compatible with our model. In this section, we discuss methods that enable one to make a judgment about such compatibility. We consider two situations. In the first one, we assume that the population of losses, whether or not it fits the model, remains homogeneous. In other words, we cannot readily identify subpopulations (SPs) for which the underlying model parameters can be suspected to be different. In the second situation, we have reasons to suspect nonhomogeneity and must test whether this is indeed the case.

Homogeneous population

Consider the case in which the estimated model does not fit the population of losses. When this case is associated with the choice of a wrong model rather than with the

presence of subpopulations, one can use a number of graphical and analytical tools to test the adequacy of the model. One important graphical tool is the probability plot. Denote the ordered observations (log losses) by $y_{(1)}, y_{(2)}, \cdots, y_{(k)}$. Denote the cdf of the observations, conditional on discovery, by

$$\tilde{F}_c(y|\boldsymbol{u},\boldsymbol{v}) = \int_{-\infty}^{y} \tilde{f}_c(t|\boldsymbol{u},\boldsymbol{v})dt = \int_{-\infty}^{y} \left[\tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})dt\right] / C(\boldsymbol{u},\boldsymbol{v}). \quad (23)$$

Suppose that the estimates of the parameters are $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$. One form of a probability plot is obtained by plotting, for $i = 1, 2, \cdots, k$, the points $[i/(k+1), \tilde{F}_c(y_{(i)}|\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})]$. The failure of these points to form a straight line with slope 1 is an indication of a lack of fit. Some standard tests, such as the Kolmogorov–Smirnov test or the Anderson–Darling test, can be used to test for the significance of the observed lack of fit. In cases in which parameters are estimated on the basis of the same data that is used in goodness-of-fit tests, we recommend the use of appropriately adjusted significance levels [21]. Another form of the probability plot is sometimes useful in models involving special parametric structure, such as location-scale equivariance. This form is obtained by computing the scores $s_i = \tilde{F}_c^{-1}[i/(k+1)]$ and plotting the points $(y_{(i)}, s_i)$, $i = 1, 2, \cdots, k$.

Another useful method is to compare the log-likelihoods corresponding to individual losses with respect to the expected values. Denote the mean and variance of a single log-likelihood term by

$$E(\boldsymbol{u},\boldsymbol{v}) = \frac{\int_{-\infty}^{\infty} \tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v}) \times \ln\left[\tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})\right] dt}{C(\boldsymbol{u},\boldsymbol{v})} - \ln C(\boldsymbol{u},\boldsymbol{v}),$$

$$V(\boldsymbol{u},\boldsymbol{v}) = \frac{\int_{-\infty}^{\infty} \tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v}) \times \ln^2\left[\tilde{f}(t|\boldsymbol{u})\tilde{p}(t|\boldsymbol{v})\right] dt}{C(\boldsymbol{u},\boldsymbol{v})} - [E(\boldsymbol{u},\boldsymbol{v}) + \ln C(\boldsymbol{u},\boldsymbol{v})]^2. \quad (24)$$

Now assume that the model is correct and that the components of the model conform to well-known regularity conditions that ensure the conventional asymptotic properties of the estimators $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$. Then, for sufficiently large k, the normalized discrepancy

$$Z = \frac{\left\{\sum_{i=1}^{k} \ln\left[\tilde{f}_c(y_i|\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})\right]\right\} - kE(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})}{\sqrt{kV(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})}} \quad (25)$$

can be treated as a realization of a standard normal random variable. We can therefore reject, at the level of significance γ, the hypothesis that the observed losses come from the postulated model if $|Z| > z_{1-\gamma/2}$. For smaller sample sizes one may need to introduce more general distributions (e.g., possibly a Student's t-distribution) in order to describe the stochastic behavior of Equation (25) and to obtain suitable rejection threshold values. It is important to note, however, that if the distribution of log losses belongs to a location-scale family, and the DPC is a function of $(y - v_1)/v_2$, the distribution of Equation (25) is characterized by the pair of values, $a_1 = (u_1 - v_1)/v_2$ and $a_2 = u_2/v_2$. This fact greatly simplifies the work needed to establish goodness of fit, since the quantiles of the test statistic can be pre-computed (or tabulated) in the form of a three-way table containing a collection of tail quantiles of Equation (25) for every pair of values (a_1, a_2). This point is also relevant for other goodness-of-fit tests considered in this paper, such as Kolmogorov–Smirnov and Anderson–Darling tests.

One can make use of a number of additional goodness-of-fit tests to be found in [21]. When applying such tests, researchers should be aware that it is important to remember that practically every model of a fixed level of complexity will be rejected when the sample size becomes sufficiently large. As noted by statistics expert George E. P. Box, "All models are wrong, but some are useful." Therefore, model rejection typically leads one to examine aspects of the model that contradict the data. One may decide that the presence of subpopulations has led to model rejection, and then one may switch to a more complex model. On the other hand, one may find the violations that led to rejection of the model to be of little practical significance, and thus one may declare the model under consideration to be useful despite these violations.

Nonhomogeneous population

In the process of data collection, one generally tries to ensure a high degree of data homogeneity in order to prevent biases related to influential unidentified subpopulations, by performing an appropriate identification and classification of the losses. It is particularly important to prevent "hidden factors" from increasing the variability of the data or creating trends that must be addressed via segmentation or some other form of fragmentation of the data set. Such fragmentation could greatly increase the complexity of the model and thus lead to loss of statistical power, for example, as measured by the predictive ability of the model.

In many situations, however, subpopulations may occur naturally; under such conditions, one must decide whether a given subset of data should be treated as one coming from a homogeneous population or as one coming from a population that contains several subpopulations. Consider, for example, the case in which

the data set contains losses corresponding to two types of businesses: banking and others. If we disregard the distinction between the subpopulations and apply one of the tests described above, we might reach a conclusion that some given model adequately represents the observed losses. However, fitting two separate models to subpopulations of interest may explain the data much better. Suppose, for example, that we have identified m subpopulations $P_1, P_2, \cdots, P_m$ for which we suspect that the parameters of the underlying population of losses are different, but the DPCs are the same and are assumed to be known. One possible test for homogeneity could be based on the following statistic:

$$T = 2\left[\sum_{j=1}^{m} L(\hat{\boldsymbol{u}}_j, \boldsymbol{v}, \hat{N}_j | \boldsymbol{y}_j) - L(\hat{\boldsymbol{u}}, \boldsymbol{v}, \hat{N} | \boldsymbol{y})\right], \tag{26}$$

where

- $\boldsymbol{y}_j$ is the subsample of losses corresponding to the jth subpopulation.
- $\hat{\boldsymbol{u}}_j$ is the vector of estimated parameters based on the data for the jth subpopulation only.
- $\hat{N}_j$ is the estimated number of losses in the jth subpopulation. The estimation is based on y_j only.
- $L(\hat{\boldsymbol{u}}_j, \boldsymbol{v}, \hat{N}_j | \boldsymbol{y}_j)$ is the maximum log-likelihood based on the data corresponding to the jth subpopulation only.
- $\boldsymbol{y}$ is the overall sample.
- $\hat{\boldsymbol{u}}$ is the vector of estimated parameters based on the complete sample.
- $\hat{N}$ is the estimated overall number of losses.
- $L(\hat{\boldsymbol{u}}, \boldsymbol{v}, \hat{N} | \boldsymbol{y})$ is the maximum log-likelihood based on the complete sample.

If the population is homogeneous and the sample sizes of subpopulations are sufficiently large, the statistic T should have a chi-square distribution, with the number of degrees of freedom equal to the product of m and the number of parameters in which the subpopulations differ from one another. For example, if the distributions of log losses corresponding to different populations can differ in both location and scale, the number of degrees of freedom is $2m$. We reject the homogeneity hypothesis at the level of significance γ if T exceeds the $(1-\gamma)\times 100\%$th quantile of the chi-square distribution mentioned above.

It is not difficult to generalize the above test for the case in which the DPC parameters for various subpopulations can also be different.

Tail-based inference

As mentioned in the previous section, one can still make use of the model under consideration even if the goodness-of-fit tests, which are based on the complete data set, suggest its rejection. Consider a situation in which the company must estimate the reserves needed to cover the overall losses in the coming year. Consider two types of losses: small losses (not exceeding some prescribed level A) and large losses (greater than A). The company has enough internal information to estimate the magnitude and frequency of small losses. Larger losses, however, are rarely observed within the company, providing no solid basis for statistical estimation. It is then natural to perform the data analysis under the working assumption that the distribution of large losses pertaining to the business of the company can be estimated on the basis of observed losses suffered by "similar" companies. The company performs a search of the body of sources in order to collect information on such losses. Suppose that most of the discovered losses are greater than A, and our attempt to fit a model involving, for example, Weibull losses and logistic DPC fails; nonetheless, a possibility exists that this model will fit suitably transformed data if we limit our attention to the population of losses that are greater than A. For example, such a model could well fit some form of excess loss data, such as $(x_i - A)$ or $\ln(x_i/A - 1)$.

Estimation in the domain $x > A$ is of primary interest in the field of insurance. Suppose that the company intends to insure itself against losses exceeding A (here A could also represent the deductible demanded by the insurance company). From the point of view of the insurance company, losses below A are of no interest, and its risk analysis can be performed solely on the basis of a distribution that fits the data only in the domain $x > A$ (i.e., in the "tail area" of the data for large observed losses).

Instead of fitting some distribution to some form of excess loss data as suggested above, we may use an alternative approach inspired by the asymptotic theory of sample extremes [22]. One of the main subjects of this theory is the analysis of distributions that have a Pareto tail index, that is,

$$1 - F(x) \sim x^{-a} L(x), \quad \text{as } x \to \infty. \tag{27}$$

Here, $L(x)$ is some *slowly varying* function, that is, a function that satisfies the relation

$$L(tx)/L(x) \to 1 \quad \text{as } x \to \infty, \tag{28}$$

for every $t > 0$. Many examples exist of this class of distributions with a Pareto tail index, and the class includes many of the distributions used by practitioners to model losses. When A is a large number, as is the case in insurance applications or the problem of operational risk estimation described at the beginning of the section, the distribution of the data in the domain $x > A$ is given by

$$\begin{aligned} F(x | x > A) &= 1 - (x/A)^{-a}[L(x)/L(A)] \\ &\approx 1 - (x/A)^{-a}, \quad x > A. \end{aligned} \tag{29}$$

Table 1 Losses (in DM) corresponding to selected values of DPC for Weibull–logistic (WL) and Pareto–logistic (PL) models. (The column heads indicate DPC percentiles corresponding to probabilities of loss discovery. For example, the column corresponding to DPC = 0.5 shows the losses for which the probability of discovery is 0.5.)

Model	*Percentiles:* *0.1*	*0.25*	*0.5*	*0.75*	*0.9*	*0.99*
WL, $\boldsymbol{u} = (17.1, 3.74)$, $\boldsymbol{v} = (8.78, 0.34)$	3,100	4,500	6,500	9,400	14,000	31,000
WL, $\boldsymbol{u} = (10.2, 7.5)$, $\boldsymbol{v} = (14, 1.7)$	28,000	1.9×10^5	1.2×10^6	1.8×10^6	5.0×10^7	3.0×10^9
PL, $\boldsymbol{u} = (14, 1.14)$, $\boldsymbol{v} = (19.7, 0.98)$	4.2×10^7	1.2×10^8	3.2×10^8	1.1×10^9	3.2×10^9	3.4×10^{10}
PL, $\boldsymbol{u} = (14, 1.97)$, $\boldsymbol{v} = (17, 1)$	2.7×10^6	8.1×10^6	2.4×10^7	7.2×10^7	2.2×10^8	2.4×10^9

This approximation, suggested by Equation (28), can be justified in many practical situations. In the simplest case, in which the distribution of losses in the range of interest $x > b$ is a two-parameter Pareto,

$$F(x) = 1 - (x/b)^{-a}, \quad x > b, \tag{30}$$

i.e., $L(x) \equiv 1$, the approximation in Equation (29) reduces to equality; in other words, $F(x|x > A) = 1 - (x/A)^{-a}$ exactly, not approximately. In terms of logarithms, the distribution becomes shifted exponential, that is,

$$\tilde{f}(y|y > u_1) = 1 - \exp[-(y - u_1)/u_2],$$

$$\tilde{f}(y|y > u_1) = u_2^{-1}\exp[-(y - u_1)/u_2], \quad y > u_1, \tag{31}$$

where

$$u_1 = \ln(A), \quad u_2 = 1/a. \tag{32}$$

The above argument illustrates the point that in the tail area, the location-scale distribution families once again lead to relatively tractable models; however, as in Equation (31), the location parameter typically turns out to be the left endpoint of the corresponding distribution.

The model in Equation (29) is a special case of the generalized Pareto distribution (GPD), which can be represented in the form

$$F(x|x > A) = \begin{cases} 1 - [1 + \xi(x - A)/\beta]^{-1/\xi}, & \xi \neq 0,\ x > A, \\ 1 - \exp[-(x - A)/\beta], & \xi = 0,\ x > A, \end{cases} \tag{33}$$

where $\xi \neq 0$ and $\beta = A\xi$. The latter model covers more general tail behavior, since it can be adapted to the case in which the tail of $F(x)$ cannot be represented in the form shown in Equation (27) [23–25]. Although this model does not generally represent a location-scale family, we advise researchers to explore the model once the amount of available data can support the added complexity.

In general, the case in which the log losses a) are treated as left-censored (i.e., only losses above some threshold are available), b) are assumed to come from the distribution $\tilde{f}(y|y > u_1; \boldsymbol{u})$, and c) are observed in accordance with some DPC $\tilde{p}(y_i|\boldsymbol{v})$, the problem of inference is similar to that described in the sections on likelihood-based estimation and goodness-of-fit tests. From a practical standpoint, the analysis described in these two sections is frequently simpler because in many applications u_1 can be treated as known, and thus we have fewer parameters to estimate.

Examples

For purposes of illustration, we may consider the data in Appendix A of [13]. To demonstrate the application of the described methods, we consider two cases: In the first case we fit the Weibull–logistic (WL) scheme to the entire distribution of losses contained in the set of sources. In the second case, we focus exclusively on large losses (i.e., those exceeding some "deductible" or other boundary of interest), disregarding the possible lack of fit for the distribution as a whole. We then apply the Pareto–logistic (PL) model to the data. As can be seen, the PL model provides a better fit to the data than the WL model.

Global Weibull–logistic model

In this section, we assume that the underlying distribution of losses is Weibull, i.e., that the log losses are distributed in accordance with Equation (2) and that the DPC is represented by the logistic equation in Equation (4). In the first phase, let us estimate the parameters $(\boldsymbol{u}, \boldsymbol{v}, N)$ without imposing any restrictions on them. Maximization of the log-likelihood Equation (12) leads to the estimates $\hat{\boldsymbol{u}} = (17.1, 3.74)$ and $\hat{\boldsymbol{v}} = (8.78, 0.34)$ which imply, by Equation (10), that $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.90$ and, by Equation (13), that $\hat{N} = 225/0.90 = 250$. In other words, the "best" explanation of the data offered by an unconstrained model is due to the fact that the losses correspond to the Weibull distribution with a very large scale parameter, $e^{17.1} = 2.67 \times 10^7$ and not a very small shape parameter, $c = 1/3.74 = 0.27$. Thus, the underlying set of sources does not contain many small and moderate losses. The loss corresponding to the probability of discovery 0.5 is $[e^{8.78} = 6{,}500]$, and losses corresponding to the probability of discovery 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99 are given in the first line of **Table 1**. The unconstrained approach

essentially suggests that all losses, except those less than 30,000, are discovered and represented in the available data.

From a practical standpoint, this explanation, of course, does not make sense, because our data set contains just 225 recorded losses; if all of the losses above 30,000 were discoverable, the resultant data set would have been much larger. This example illustrates what may happen if one does not consider *a priori* the plausible values of the parameters, and if one relies on statistical estimation to discover them. In the next step, assume that one has reason to believe that the DPC parameters can be treated as known and equal to $\boldsymbol{v}_0 = (14, 1.7)$. This suggests that the loss corresponding to the probability of discovery 0.5 is 1.2 million Deutsche marks, and losses corresponding to other values of the probability of discovery are given in the second row of data in Table 1. Use of the constrained estimation procedure described above in the section on constrained estimation and inference leads to the estimate $\hat{\boldsymbol{u}} = (10.2, 7.5)$, which appears more sensible from a practical point of view. The large losses are now explained less by a *large scale* parameter, as in the unconstrained case, than by a *smaller shape* parameter. This suggests that the bulk of the losses in the body of sources are still undiscovered: The estimated proportion of discovered losses is $C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0) = 0.20$; consequently, the estimated number of losses recorded in the set of sources is $\hat{N} = 225/0.20 = 1{,}125$.

As noted in the Introduction, a simulated sample from this model is shown in Figure 2. Even in the unconstrained case, we have no assurance that the model corresponding to estimated parameter values will fit the data. Once constraints are imposed, it is quite possible that the model will fit poorly, and careful examination of goodness-of-fit issues is appropriate. We apply some of the techniques described in the sections discussing estimation and goodness-of-fit tests to measure the extent to which the constraint $\boldsymbol{v}_0 = (14, 1.7)$ is compatible with the data. At this point, let us assume, for the sake of argument, that the analyst had prior knowledge not only about $\boldsymbol{v}$, but also about $\boldsymbol{u}$, and that the values $\hat{\boldsymbol{u}} = (10.2, 7.5)$ were in fact anticipated *a priori*. Such an assumption can be justified, for example, in situations in which the model is estimated on the basis of the learning sample and then applied to the test sample, as mentioned early in this paper in the section on the basic approach. Under this assumption, we could test the data for conformance with a fully specified model. In particular, let us consider the probability plot in **Figure 3**. The maximal deviation from the straight line, corresponding to the hypothesized model, is 0.09. In accordance with common statistical practice, we would reject the hypothesis that the model fits the data if this deviation exceeds the critical 5% value for the Kolmogorov–Smirnov statistic. Since this critical value is known to be $1.36/\sqrt{225} = 0.09$ (see Table 54 in [26]), we have insufficient evidence that the model does not fit the data. However, one could apply an alternative (Anderson–Darling) goodness-of-fit test that tends to be more sensitive with respect to deviations in the tails. The value of the Anderson–Darling statistic is 3.05, which corresponds to the 2.5% percentile of its distribution under the assumption that the model is adequate; therefore, the fit in the tail area is definitely problematic.

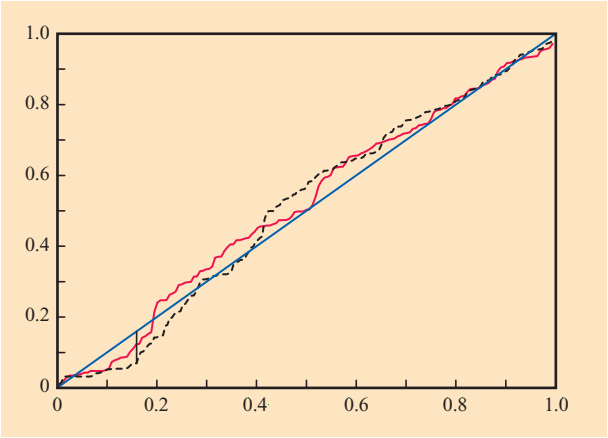

Figure 3

Probability plots. Weibull–logistic (WL) model with parameters $\hat{\boldsymbol{u}} = (10.2, 7.5)$, $\boldsymbol{v} = (14, 1.7)$ (dashed line), and Pareto–logistic (PL) model with parameters $\hat{\boldsymbol{u}} = (14, 1.97)$, $\boldsymbol{v} = (17, 1)$ in the domain $y > 14$ (wavy solid line). The maximal Kolmogorov–Smirnov deviation is shown as a short vertical line for the WL model. The x and y axes refer to probabilities, and the coordinates of points on the graph are $i/(k+1)$, $\hat{F}_c(y_{(i)} \mid \hat{\boldsymbol{u}}, \boldsymbol{v})$.

To apply a test based on the likelihood function, note that the log-likelihood of the data in the constrained model is −605. The formulas in Equation (24) suggest that for data coming from the model with parameters $\hat{\boldsymbol{u}} = (10.2, 7.5)$ and $\hat{\boldsymbol{v}} = (14, 1.7)$, the average score per observed loss is $E(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = -2.63$ and the variance is $V(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.55$. This suggests that the value of the log-likelihood observed under the estimated model is approximately normal, with mean and standard deviation $-2.62 \times 225 = -591$ and $\sqrt{0.55 \times 225} = 11.1$, respectively. The value −605 is within 1.27 standard deviations from the mean, which corresponds to the p-value of 0.1 for the one-sided goodness-of-fit test; thus, this test does not lead to rejection of the model.

One may notice that even under the assumption that the model is completely specified, the constrained model is barely acceptable, and it does not even have some features that a practitioner may desire. In particular, given that the provided data is the result of a limited search effort, the probability 0.1 of discovering a loss of

magnitude 28,000 in the body of sources appears to be too high, and the overall probability of discovery 0.2 also appears to be too high. However, our analysis shows that an attempt to obtain a much better model of the Weibull–logistic type for the available data set is not successful: Models that appear more attractive from the practical standpoint unfortunately do not fit the data, especially for lower loss values. The fit in the right tail (i.e., for higher loss values) is also problematic, as can be seen from Figure 2, in which the right tail bends slightly in comparison with the depiction in Figure 1. The difficulties with the model describing the data are primarily related to the fact that at the time the data became available to us, the efforts of populating the database were in the very initial stage and gave rise to very uneven coverage. Furthermore, some values of the data have a much higher probability than the neighboring data, which exposes the fact that a Weibull model is *a priori* just a convenient mathematical approximation. For example, a typical small fine imposed by a judge against an operational-risk-related violation, and reported in the press, is much more likely to be \$10,000 (DM 18,497 in the data set) than \$9,000. Though such partial grouping (e.g., in which loss values tend to cluster around some "round" quantities such as \$10,000) does not prevent the estimation process from producing useful results, it is advisable to define, for every individual study, the effect of grouping on both estimation and goodness-of-fit tests.

Finally, we test whether, under the assumption that the model is WL with $\hat{\boldsymbol{v}} = (14, 1.7)$, the population of losses classified as being of "high" or "medium" relevance (subpopulation 1, denoted SP1) differs significantly from the population classified as being of "low" relevance (subpopulation 2). Let us fit two separate models for the two subsamples. The estimated population parameters (based on a sample size of 140) for SP1 are $\hat{\boldsymbol{u}}_1 = (9.8, 7.9)$, and the maximal value of the log-likelihood is -385. (The model cannot be rejected by a goodness-of-fit test, but the quality of the fit is marginal.) The parameters for SP2, based on a sample size of 85, are $\hat{\boldsymbol{u}}_2 = (10.7, 7.0)$. The maximal value of the log-likelihood is -220, and the fit is very good. As indicated earlier, the maximum log-likelihood value for the complete data set was -605. To test whether the complete data set is explained better by two separate models, one for SP1 and the other for SP2, rather than by a single model, we must compute T. Because we have two subpopulations that differ in two parameters, T should be compared to $\chi^2_{0.95}(4) = 9.49$. In our case, $T = 2 \times (-385 - 220 + 605) = 0$, indicating that we have no evidence, using the given DPC, to conclude that there is a significant difference between the loss distributions corresponding to SP1 and SP2.

Note that the assumption that the model is fully specified is quite consequential, and in many practical situations one will have to assess the validity of the model on the basis of "in-sample" data. In other words, when using "in-sample" data we estimate the model on the basis of a given sample, and then use the same sample to test goodness of fit. In such situations, significance values (or p-values) for the tests mentioned above must be adjusted by accounting for the fact that the model parameters $\boldsymbol{u}$ were estimated from the same data that was used in goodness-of-fit tests, resulting in a fit that appears to be better than what would be expected under a fully specified model. To obtain the significance values, one can conduct simulations in order to study the behavior of the test statistics under the assumption that the true values of the parameters are equal to the estimated values. In every simulation run, one obtains a new set of losses, estimates the model parameters, substitutes these estimates into a given goodness-of-fit statistic, and computes the discrepancy. This type of technique is called the "parametric bootstrap" [27]. In particular, for the above WL model with $\hat{\boldsymbol{u}} = (10.2, 7.5)$ and $\boldsymbol{v}_0 = (14, 1.7)$, the significance levels of the Kolmogorov–Smirnov and Anderson–Darling tests are estimated to be 0.001 and <0.001, respectively, indicating that an incompletely specified model of this type (i.e., $\boldsymbol{u}$ estimated solely on the basis of the same data that is used to test goodness of fit) would be promptly rejected.

Tail Pareto–logistic (PL) model

We may now consider the situation from the perspective of the insurance company and assume that only losses exceeding $A = 1.2$M (the deductible) are of interest. We further assume that the distribution of losses follows a two-parameter Pareto distribution. As noted in the previous section, this implies that the log losses are distributed in accordance with Equation (31), with $u_1 = \ln(1.2 \times 10^6) = 14$. The only parameters of interest are $\boldsymbol{v}$, the scale u_2, and N. The relevant data is now reduced from 225 to 163 losses that exceed 1.2M.

First, let us apply the unconstrained model. Maximization of the likelihood based on *conditional* density functions leads to the estimates $\hat{\boldsymbol{u}}_2 = 1.14$ and $\hat{\boldsymbol{v}} = (19.7, 0.98)$. Therefore, $C(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = 0.02$, and the total number of losses contained in the set of sources that exceed 1.2M is estimated to be $\hat{N} = 163/0.02 = 8{,}150$. This model suggests that the loss that has the probability of 0.5 to enter into the set of sources and be discovered is 3.7×10^8. Losses corresponding to various values of the DPC are shown in the third row of data in Table 1.

Once again, we have a reason to be disappointed by the results of the automatic search for the model parameters. Under the DPC corresponding to this model (see Table 1), losses of very high magnitude have an

uncomfortably high chance of being overlooked in the process of populating the database. One can then choose to override the estimated $\hat{\boldsymbol{v}}$ by introducing a constraint that the parameters of the logistic DPC are $\hat{\boldsymbol{v}}_0 = (17, 1)$. Under this constraint, the magnitude of a loss that has a 0.5 probability to be discovered is 2.4×10^7. Other values (see Table 1) also appear to be more reasonable to a decision maker. The resulting estimate of the scale corresponds to $u_2 = 1.97$. The estimated overall discovery probability is now much larger, namely $C(\hat{\boldsymbol{u}}, \boldsymbol{v}_0) = 0.30$, leading to $163/0.3 = 543$ as the estimate of the total number of losses of magnitude exceeding 1.2M in the time period of interest.

To check whether the resulting PL model fits the data for losses exceeding 1.2M, with parameters $\hat{\boldsymbol{u}} = (14, 1.97)$ and $\boldsymbol{v}_0 = (17, 1)$, we once again consider two cases. In case (a), the parameter $\boldsymbol{u}_0 = (14, 1.97)$ is assumed to be provided externally (for example, based on the training data set); in case (b) it is assumed to be estimated from the data, as illustrated above.

In case (a), we first examine the probability plot in Figure 3. Both the Kolmogorov–Smirnov and Anderson–Darling tests suggest that the fit is good. The mean and variance of the log-likelihood score corresponding to a single measurement are $E(\boldsymbol{u}_0, \boldsymbol{v}_0) = -2.19$ and $V(\boldsymbol{u}_0, \boldsymbol{v}_0) = 0.55$. Therefore, the mean and standard deviation of the log-likelihood under the assumption of the above model are $-2.19 \times 163 = -356$ and $\sqrt{0.55 \times 163} = 9.4$. The maximal log-likelihood computed for the available data under the constraint $\boldsymbol{v} = \boldsymbol{v}_0 = (17, 1)$ is -353, which is in agreement with the mean and standard deviation computed above. In case (b), the goodness of fit is still acceptable: For example, the significance levels of the Kolmogorov–Smirnov and Anderson–Darling tests are estimated to be 0.12 and 0.05, respectively.

Conclusions and directions of future research

The process of data collection for the estimation of the intensity and stochastic characteristics of losses is typically conducted under conditions that induce biases. The presence of these biases requires the use of special statistical techniques, some of which originate in other fields in which data bias is also a problem. The framework of size-biased sampling appears to be useful for identifying and correcting biases related to the sizes of the losses, and for early identification of models that are suitable for characterizing the process of losses. In particular, this framework may be used to obtain the basic "building blocks" characterizing various categories of losses from which a more comprehensive model can eventually be obtained.

The limited scope of the data set that served as the basis of this research enables us to answer only a few basic questions by assuming a given model with some very simple data categorization. One would need a much more elaborate data set in order to address more complex questions. For example, consider the problem of building a model in order to estimate operational risk losses for a given enterprise. A database suitable for such estimation would consist of a list of losses, and for each record we would have not only loss magnitude and relevance, but also such entries as industry, number of employees, type of loss, and market value of the bank. A promising strategy for risk estimation suggests that we

- Establish, for each factor, whether it affects a) the intensity (i.e., rate) of losses λ, b) (u_1, u_2), or c) (v_1, v_2). (Some effects can be reasonably postulated; for example, for some types of losses the rate of losses could be assumed to be roughly proportional to the number of employees. Such *a priori* relationships can simplify the subsequent analysis considerably. Their validity can be tested by using post-estimation goodness-of-fit procedures.)
- Estimate the relationship between factors and the basic model parameters, λ, (u_1, u_2), (v_1, v_2).
- For a given enterprise P, evaluate, on the basis of the above model, the corresponding parameters, λ_P, (u_{1P}, u_{2P}), (v_{1P}, v_{2P}).
- Evaluate risks (for example, value-at-risk, or VAR) related to P on the basis of the estimated parameters.

The proposed framework may be used not only for modeling the process of external losses, but also for the analysis of internal losses pertaining to a given institution of interest. Although one can expect that the internal databases (i.e., databases that contain information that comes from inside a company) are much better structured and maintained than databases based on external searches, biases related to the size of losses are likely to remain a factor of concern. The bias-causing mechanisms for various classes of internal losses, however, are likely to be of a different nature than those related to external data sources. Furthermore, bias in internal data may be a lesser factor in determining deleterious exposure to tail events and VAR than biases related to external data sources, possibly necessitating modifications in the statistical inference procedure.

Another possible extension of the proposed approach is related to an inference system based on three sources of data: internal losses, external losses coming from a shared database pertaining to the same type of business (for example, banking institutions only), and losses discovered by searches of external sources that may be related to a much broader class of institutions.

In our analysis we emphasized the so-called "frequentist" inference techniques, whereby we

considered the parameters as fixed (although possibly unknown) quantities. However, under the described conditions, we might benefit from the alternative statistical methodology, namely, Bayesian inference, which considers the parameters as random variables having some *prior* distribution. For example, given the very small sample size, we typically considered the parameters v of the DPC as known quantities. By using a Bayesian approach, one could associate some degree of uncertainty with these parameters. Bayesian inference may also be useful in aggregating information obtained from several databases, as described above.

Finally, it is important to note that the presented approach takes into account only one source of bias, namely under-reporting related to the size of the loss. However, in practice the process of populating a database may involve other biases that are equally serious. One such bias may be related to the multivariate nature of the process of losses that could have a profound effect on VAR estimation. For example, an event represented in a database as a computer failure may actually be caused by a massive power outage, affecting a larger class of losses and magnitudes that are not represented in a database. Therefore, accurate modeling of the risk exposure may benefit from the use of a larger range of databases, necessitating a more complex modeling effort.

Acknowledgments

I deeply appreciate the help of Dr. Mark Laycock and Dr. Stephen Witter, both from Deutsche Bank, who provided comments and criticisms on the original version of the manuscript. I also thank Drs. Jonathan Hosking, Dirk Siegel, and Katie Richards for useful discussions on this subject. I thank the five anonymous referees for their valuable suggestions and criticisms.

References

1. Basel Committee on Banking Supervision, *Convergence of Capital Measurement and Capital Standards*, Bank for International Settlements, Basel, Switzerland, 2006.
2. M. G. Cruz, *Modeling, Measuring and Hedging Operational Risk*, Wiley, New York, 2002.
3. D. N. Chorafas, *Operational Risk Control with Basel II: Basic Principles and Capital Requirements*, Elsevier, Amsterdam, The Netherlands, 2004.
4. B. Engelmann and R. Rauhmeier, *The Basel II Risk Parameters*, Springer, Berlin–Heidelberg, Germany, 2006.
5. H. H. Panjer, *Operational Risks: Modeling Analytics*, Wiley, New York, 2006.
6. C. Alexander, "Rules and Models," *Risk* **15**, No. 1, 18–20 (2002).
7. F. Cheng, D. Gamarnik, N. Jengte, W. Min, and B. Ramachandran, "Modeling Operational Risks in Business Processes," *Research Report RC-23672*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 2004.
8. C. Supatgiat, C. Kenyon, and L. Heusler, "Cause-to-Effect Operational Risk Quantification and Management," *Risk Manage*. **8**, 16–42 (2006).
9. P. Giudici and A. Bilotta, "Modeling Operational Losses: A Bayesian Approach," *Qual. & Reliabil. Eng. Intl.* **20**, 407–417 (2004).
10. K. Adusei-Poku, "Operational Risk Management—Implementing a Bayesian Network for Foreign Exchange and Money Market Settlement," Doctoral Thesis, University of Göttingen, Germany, 2005.
11. A. Roehr, "Modeling Operational Losses," *Algo Res. Quart.* **5**, No. 2, 53–64 (2002).
12. V. Chavez-Demoulin, P. Embrechts, and J. Neslehova, "Quantitative Models for Operational Risk: Extremes, Dependence and Aggregation," *J. Banking Finance* **30**, No. 10, 2635–2658 (2006).
13. E. Yashchin, "Modeling of Risk Losses Based on Incomplete Data," *Research Report RC-23676*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 2005.
14. J. R. Premister, U. Oktem, P. R. Kleindorfer, and H. Kunreuther, "Near-Miss Incident Management in the Chemical Process Industry," *Risk Anal.* **23**, No. 3, 445–459 (2006).
15. G. Van den Brink, *Operational Risk: The New Challenge for Banks*, Palgrave, New York, 2002.
16. D. K. Rosenberg and W. S. Overton, "Estimation of Animal Abundance When Capture Probabilities are Low and Heterogeneous," *J. Wildlife Manage.* **59**, 252–261 (1995).
17. B. Littlewood, "Predicting Software Reliability," *Phil. Trans. Roy. Soc. Lond. A* **327**, 513–527 (1989).
18. B. D. Olin and W. Q. Meeker, "Applications of Statistical Methods to Nondestructive Evaluation," *Technometrics* **38**, No. 2, 95–112 (1996).
19. T. Alderweireld, J. Garcia, and L. Leonard, "A Practical Operational Risk Scenario Analysis Quantification," *Risk* **19**, No. 2, 93–95 (2006).
20. E. L. Lehmann, *Theory of Point Estimation*, Wiley, New York, 1983.
21. R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit Techniques*, Marcel Dekker, New York, 1986.
22. J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, Robert E. Krieger Publishing Co., Malabar, FL, 1987.
23. A. A. Balkema and L. de Haan, "Residual Lifetime at Great Age," *Ann. Probabil.* **2**, 792–804 (1972).
24. J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of Extremes: Theory and Applications*, Wiley, New York, 2004.
25. K. Bocker, "Operational Risk: Analytical Results When High-Severity Losses Follow a Generalized Pareto Distribution (GPD)—A Note," *J. Risk* **8**, No. 4, 117–120 (2006).
26. E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 2, Biometrika Trust, London, U.K., 1976.
27. A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*, Cambridge University Press, U.K., 1997.

Received September 15, 2006; accepted for publication October 14, 2006; Internet publication May 29, 2007

Emmanuel Yashchin *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (yashchi@us.ibm.com)*. Dr. Yashchin is a Research Staff Member in the Department of Mathematical Sciences at the IBM Thomas J. Watson Research Center. He joined IBM in 1983. He received a Diploma in applied mathematics from Vilnius State University (U.S.S.R.) in 1974 and a M.Sc. degree in operations research and a D.Sc. degree in statistics from the Technion–Israel Institute of Technology in 1977 and 1981, respectively. In 1982, he was a Visiting Assistant Professor at Iowa State University. From 1996 to 2002, Dr. Yashchin was Manager of the Statistics Group at the IBM Thomas J. Watson Research Center. His research interests include quality control, reliability, statistical modeling, risk analysis, and operations research.