

# MODELING INSURANCE CLAIMS WITH EXTREME OBSERVATIONS: TRANSFORMED KERNEL DENSITY AND GENERALIZED LAMBDA DISTRIBUTION

Uditha Balasooriya\* and Chan-Kee Low†

---

## ABSTRACT

In modeling insurance claims, when there are extreme observations in the data, the commonly used loss distributions often are able to fit the bulk of the data well but fail to do so at the tail. One approach to overcome this problem is to focus on the extreme observations only and model them with the generalized Pareto distribution, as supported by extreme value theory. However, this approach discards useful information about the small and medium-sized claims, which is important for many actuarial purposes. In this article we consider modeling large skewed data using a highly flexible distribution, the generalized lambda distribution, and the recently proposed semiparametric transformed kernel density estimation. Our results suggest that both these approaches are credible options for the investigator when modeling insurance claims data that typically contain large extreme observations. In addition, even at the extreme tails they perform well when compared with the generalized Pareto distribution.

---

## 1. INTRODUCTION

In practice, finding a good-fitting distribution for large data sets that contain some relatively large claim amounts, such as insurance claims, is not an easy task. Often actuaries find that standard models such as the lognormal, Weibull, and Pareto are able to fit the bulk of the data well, but they fail to capture adequately atypical extreme observations. Because these extreme observations are disproportionately important for rating or reserving purposes, some actuaries have analyzed them separately from the bulk of the data.

Extreme value theory (EVT) provides a theoretical foundation for statistical methods that analyze extremal observations. The theory shows that, for a broad class of distributions that include the normal, gamma, lognormal, Weibull, and many other commonly used distributions, *conditional excess values* over a sufficiently high threshold follow a generalized Pareto distribution (GPD). Among others, Rootzén and Tajvidi (1996), McNeil and Saladin (1997), and Cebrián, Denuit, and Lambert (2003) have applied the GPD to analyze actuarial problems.

This approach requires the determination of an appropriate threshold value. If too high a threshold is chosen, the number of observations above it could be too few to permit accurate estimates of upper quantile values. On the other hand, if the threshold is too low, the GPD may not apply to the moderate observations, resulting in biased estimates. As the moderate observations usually form a large proportion of the sample, this bias can be a serious problem.

For rating purposes, it is important that the information contained in all claims data be utilized. As pointed out by Bolancé, Guillen, and Nielsen (2003), “Actuaries are interested in having good estimates

---

\* Uditha Balasooriya, PhD, is Associate Professor in the Division of Banking and Finance/College of Business (Nanyang Business School), Nanyang Technological University, Singapore, auditha@ntu.edu.sg.

† Chan-Kee Low, PhD, is Associate Professor in the Division of Economics/School of Humanities and Social Sciences, Nanyang Technological University, Singapore, acklow@ntu.edu.sg.

for all the values in the domain range: small losses because they are very frequent, medium losses causing a dramatic increase of expenses (demanding liquidity) and large losses that may mean that reinsurance contracts should be reconsidered.” The excess-over-threshold model would be primarily useful for making inferences about the tail area of the loss distribution. Such information will be particularly useful for decisions concerning reinsurance, but not adequate for the direct pricing of insurance products. Ideally, one hopes to obtain a single density function that can fit the whole range of data well, without the need to leave out any sample information.

In this article we investigate two approaches to analyzing large data sets that contain extreme observations: fitting the highly flexible generalized lambda distribution (GLD) and a semiparametric transformed kernel density estimation. Our results show that these two approaches yield densities that fit the entire data range well. Further, at the extremes, they compare well with the excess-over-threshold approach of the EVT.

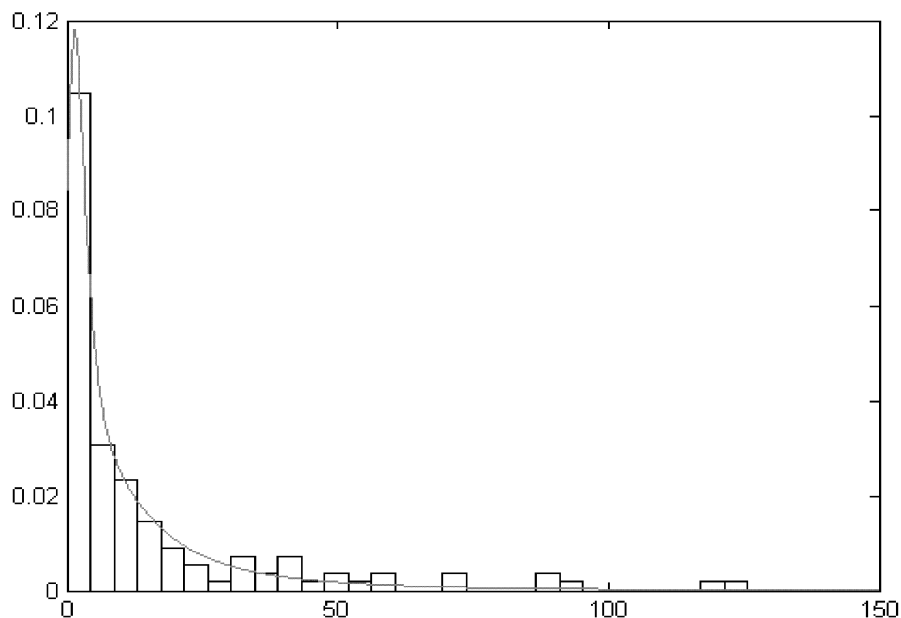
In Section 2 we briefly outline the semiparametric transformed kernel density estimation. The GLD introduced by Ramberg et al. (1979) is discussed in Section 3. In Section 4 we present the results of a Monte Carlo experiment that compares several cumulative distribution transformations used for the semiparametric kernel density estimation. In Section 5 we apply the proposed approaches to the Society of Actuaries’ medical claims data and compare our results with the excess-over-threshold approach. A general discussion of our findings and concluding remarks are given in Section 6.

## 2. THE TRANSFORMED KERNEL DENSITY ESTIMATION

In kernel density estimation, the shape of the estimated density is determined by the data, and in principle, given a sufficiently large data set, the technique is capable of estimating an arbitrary density  $f$  fairly accurately. It is a nonparametric method that does not make any distributional assumption about the underlying density. Kernel density estimation has attracted the attention of many researchers; a good introduction to the subject is given in Silverman (1986).

Let  $X_1, X_2, \dots, X_n$  be a random sample from an unknown density function  $f_X$ . Then the kernel density estimator of  $f_X$  is given by

Figure 1  
Plot of Transformed Kernel Density of Traffic Data



$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K_h \left( \frac{x - X_i}{h} \right),$$

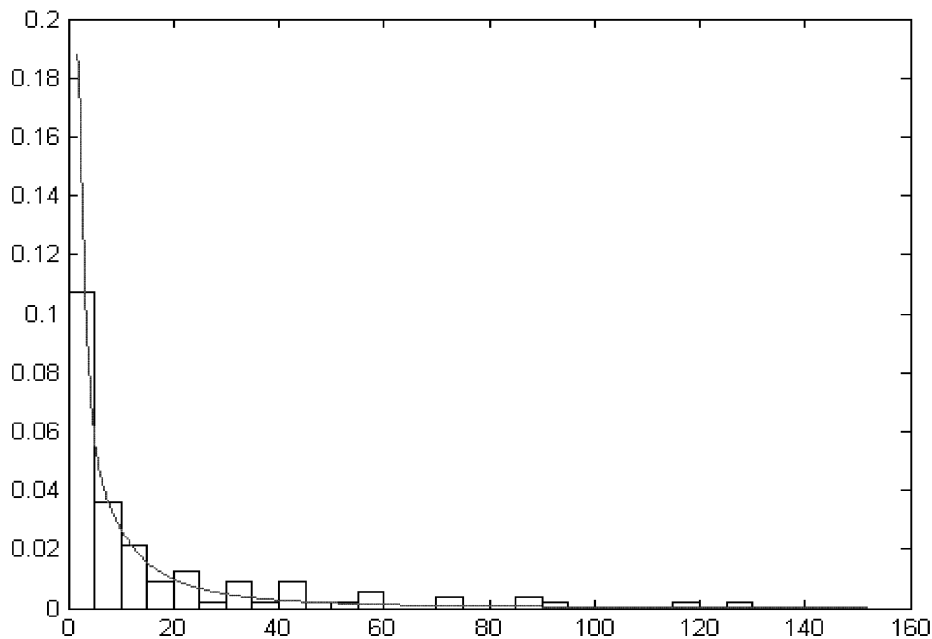
where the kernel function  $K_h$  is generally a *unimodal probability density function* and  $h(> 0)$  is a smoothing parameter often called the *bandwidth*.

One difficulty with kernel density estimation is the choice of a bandwidth. For long-tailed distributions, there is a tendency for spurious noise to appear in the tail area of the distribution. If a large bandwidth is chosen to deal with this, then essential details in the main part of the distribution are masked. On the other hand, if a narrower bandwidth is chosen to preserve details in the bulk of the distribution, uneven humps typically arise in the tails of the distribution. Some early works on bandwidth selection include that of Bowman (1984), Stone (1984), Park and Marron (1990), Chiu (1991, 1992), Jones, Marron, and Park (1991), and Sheather and Jones (1991).

To address the problem associated with a fixed bandwidth, Wand, Marron, and Ruppert (1991) proposed to transform the data so that a global bandwidth may be more appropriate to the transformed data over the entire data range. They have shown that by applying kernel smoothing on transformed data, one could improve the accuracy of estimation by six or seven times over the classical kernel density estimation on the original data. In general, this transformation reduces the skewness of the data and thereby improves the kernel density estimation. In the literature various transformations have been proposed. For example, Clements, Hurn, and Lindsay (2003) proposed the Möbius-like transformation that transforms positive support into a finite sample space,  $[-1, 1]$ ; Bolancé, Guillen, and Nielsen (2003) discussed a modification to the transformation method of Wand, Marron, and Ruppert (1991), which improves the efficiency of kernel density estimation for highly skewed distributions. More recently, Bolancé et al. (2005) employed a modified Champernowne cumulative density function (*cdf*) transformation that maps the original data into a finite  $[0, 1]$  range. When compared to the shifted power transformation of Wand, Marron, and Ruppert (1991) and the Möbius-like mapping of Clements, Hurn, and Lindsay (2003), *cdf* transformation is straightforward and computationally less expensive. Further, Monte Carlo studies of Bolancé et al. (2005) have shown that for some selected underlying distributions, the kernel density estimate with *modified Champernowne cdf transformation*, given by

Figure 2

**Estimated Transformed Kernel Density Using Truncated and Censored Data**



$$T_{\alpha, M, c}(x) = \frac{(x + c)^\alpha - c^\alpha}{(x + c)^\alpha + (M + c)^\alpha - 2c^\alpha}, \quad x \geq 0,$$

where  $\alpha > 0$ ,  $M > 0$ , and  $c \geq 0$  are parameters, is superior for heavy-tailed situations when compared to the estimators of Bolancé, Guillen, and Nielsen (2003) and the Möbius-like mappings of Clements, Hurn, and Lindsay (2003).

The transformed kernel density estimation proceeds as follows. Let  $X_1, X_2, \dots, X_n$  be a random sample drawn from an unknown distribution with *cdf*  $F$  and probability density function (*pdf*)  $f$ , and  $T(x; \theta_1, \dots, \theta_m)$  be a transformation function that is the *cdf* of a probability distribution.

*Step 1:* Estimate the parameters,  $\theta_1, \dots, \theta_m$ , of  $T$  by maximizing

$$L = \prod_{i=1}^n T'(X_i; \theta_1, \dots, \theta_m),$$

where  $T'(x; \theta_1, \dots, \theta_m) = d/dx [T(x; \theta_1, \dots, \theta_m)]$ .

*Step 2:* Transform the data  $X_i, i = 1, 2, \dots, n$  by the estimated *cdf*,

$$Y_i = T(X_i; \hat{\theta}_1, \dots, \hat{\theta}_m).$$

*Step 3:* Obtain the kernel density for the transformed data,  $Y_i, i = 1, 2, \dots, n$ :

$$\hat{f}_Y(y) = \frac{1}{nh} \sum_{i=1}^n K_h \left( \frac{y - Y_i}{h} \right).$$

Because  $0 \leq Y_i \leq 1$ , the support of the estimated density,  $\hat{f}_Y(y)$ , has to be in the range  $[0, 1]$ . Boundary correction is required for each kernel function that falls outside the range. In this article we use a common approach, which is to compute the area of each kernel function that lies within  $[0, 1]$  and normalize the kernel density to have a unit area by dividing the kernel function by this area.

*Step 4:* Obtain the kernel density for the original data,  $X_i, i = 1, 2, \dots, n$  by back transformation:

$$\hat{f}_X(x) = \hat{f}_Y(T(x; \hat{\theta}_1, \dots, \hat{\theta}_m)) |T'(x; \hat{\theta}_1, \dots, \hat{\theta}_m)|.$$

The expression for  $\hat{f}_X(x)$  in terms of the kernel function is

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K_h \left( \frac{T(x; \hat{\theta}_1, \dots, \hat{\theta}_m) - T(X_i; \hat{\theta}_1, \dots, \hat{\theta}_m)}{h} \right) T'(x; \hat{\theta}_1, \dots, \hat{\theta}_m).$$

Table 1

**Observed and Estimated Relative Frequencies of Transformed Kernel Density with Truncated and Censored Traffic Data**

Interval	Observed Relative Frequency	Estimated Relative Frequency
1.5–6.5	0.465	0.454
6.5–11.4	0.140	0.155
11.4–16.4	0.114	0.088
16.4–21.3	0.035	0.055
21.3–26.3	0.035	0.037
26.3–31.2	0.026	0.026
31.2–36.2	0.026	0.020
36.2–41.1	0.026	0.015
41.1–46.1	0.026	0.012
46.1–50.0	0.009	0.004
> 50	0.096	0.129

We illustrate the above procedures using a two-parameter generalized Pareto *cdf* transformation,  $T(x; k, \sigma) = 1 - (1 + kx/\sigma)^{-1/k}$ , on the traffic data set of Bain and Engelhardt (1980). The data set contains 128 observations on times, in seconds, between vehicles at a particular location on a road. Step 1 of the procedures yields the estimates  $\hat{k} = 0.6211$  and  $\hat{\sigma} = 7.4556$ . Step 2 transforms  $X_i, i = 1, 2, \dots, 128$ , using the function  $Y_i = 1 - (1 + 0.6211X_i/7.4556)^{-1/0.6211}$ , and Step 3 fits a commonly used normal kernel density to  $Y_i, i = 1, 2, \dots, 128$ . As observed by Wand and Jones (1995, p. 13) the choice of a kernel function is not particularly important. However, the choice of a bandwidth is very important. In this study the bandwidth is chosen to minimize the asymptotic mean integrated squared error (see Bowman and Azzallini 1997, p. 31). Finally, Step 4 back-transforms the fitted kernel to obtain the estimated density of the original data. Figure 1 shows the fitted transformed kernel density superposed on the histogram of the original observations.

Insurance policies often have policy excesses and policy limits that lead to insurance claims data being left-truncated and right-censored. To fit a transformed kernel density to such data, one needs only minor modifications to Steps 1, 2, and 4. Suppose we have the following left-truncated and right-censored random sample:

$$X_1, X_2, \dots, X_m, \underbrace{X^*, X^*, \dots, X^*}_{c \text{ times}},$$

which consists of  $m$  fully observed values and  $c$  censored values. The sample is truncated below at  $X^o < \min(X_1, \dots, X_m)$ . In Step 1 the function to be maximized will be given by

$$L_c = [1 - T(X^*; \theta_1, \dots, \theta_m)]^c \prod_{i=1}^m \frac{T'(X_i; \theta_1, \dots, \theta_m)}{1 - T(X^o; \theta_1, \dots, \theta_m)}.$$

In Step 2 the transformation function is given by

$$\frac{T(X_i; \hat{\theta}_1, \dots, \hat{\theta}_m) - T(X^o; \hat{\theta}_1, \dots, \hat{\theta}_m)}{1 - T(X^o; \hat{\theta}_1, \dots, \hat{\theta}_m)}.$$

We note that, if the fraction of observations censored is large, a hump may result in the fitted kernel density around the censoring point. One way to avoid this is to fit kernels to fully observed data points only. This approach may not lead to serious estimation error at the tail of the distribution since the kernel density is fitted to transformed data within the  $[0, 1]$  range. In Step 4 the back transformation formula is given by

$$\hat{f}_X(x) = \hat{f}_Y(T(x; \hat{\theta}_1, \dots, \hat{\theta}_m)) \left| \frac{T'(x; \hat{\theta}_1, \dots, \hat{\theta}_m)}{1 - T(X^o; \hat{\theta}_1, \dots, \hat{\theta}_m)} \right|.$$

To illustrate these procedures, suppose the traffic data are left-truncated at 1.5 and right-censored at 50. Truncation removes 16 unobserved values, and the censoring causes 11 observations to be censored at 50. Figure 2 shows the transformed kernel density fitted to the left-truncated and right-censored data superposed on the histogram of the original observations. Table 1 compares the observed with the expected frequencies estimated by the transformed kernel density. These results indicate that the transformed kernel density performs well even when data are truncated and censored.

### 3. GENERALIZED LAMBDA DISTRIBUTION

The probability density function of the generalized lambda distribution with parameters  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  is given by

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3-1} + \lambda_4 (1-y)^{\lambda_4-1}}, \quad \text{at } x = Q(y) \tag{3.1}$$

with the quantile function  $Q(y)$  given by

$$Q(y) = \lambda_1 + \frac{y^{\lambda_3} - (1 - y)^{\lambda_4}}{\lambda_2}, \quad 0 \leq y \leq 1, \quad (3.2)$$

where  $\lambda_1$  and  $\lambda_2$  are location and scale parameters, respectively, and  $\lambda_3$  and  $\lambda_4$  are shape parameters (skewness and kurtosis, respectively). This distribution was first introduced by Tukey (1960) and later generalized to the four-parameter case by Ramberg and Schmeiser (1974). It can produce a wide variety of curve shapes including that of many standard symmetric and skewed distributions. Figure 3 shows a few plots of the GLD distribution corresponding to different parameter values. The GLD fit has been applied successfully in a variety of disciplines. These include modeling of quantile response in bioassay and economics (Mudholkar and Phatak 1984; Pregibon 1980), meteorology (Öztürk and Dale 1982), and engineering and quality management (Silver 1977; Dudewicz 1999).

As the estimation of the parameters is quite involved, Ramberg et al. (1979), Dudewicz (1999), and Karian and Dudewicz (2000) have proposed two methods based on moments and percentiles. Similar to the method of moments, the percentile-based approach equates sample summary measures based on percentile values with the corresponding population values to obtain the parameter estimates. In the application to the Society of Actuaries (SOA) medical claims database we use the method of percentiles for estimating the parameters of the GLD. As pointed out by Karian and Dudewicz (2000), the percentile approach has several advantages over the method of moments. It is simpler to apply and it provides more accurate estimates than the method of moments. Further, there are GLD distributions that have fewer than four moments, in which case the method of moment is not applicable. When data are truncated and/or censored, the percentile method remains valid. The current tabulated values in Karian and Dudewicz (2000) are directly applicable for curtailment not exceeding 10% of the number of observations at each extreme end of the distribution.

#### 4. SIMULATION STUDY ON TRANSFORMED KERNEL ESTIMATION

In this section we report the results of a simulation study to investigate the relative performance of some selected *cdf* transformations used in the transformed kernel density estimation.

Figure 3  
Density Plots of Selected GLD Distributions

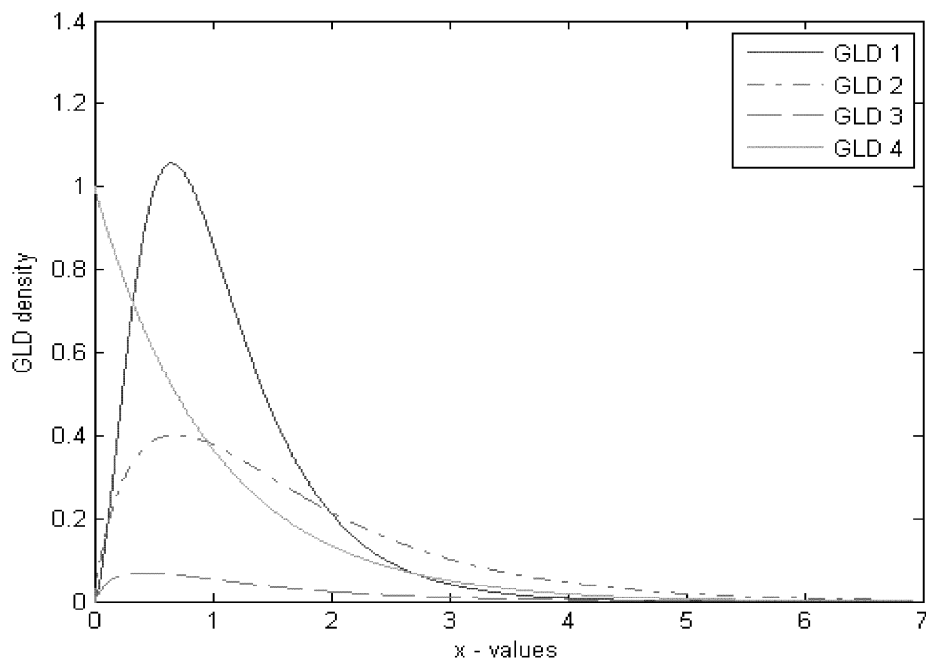
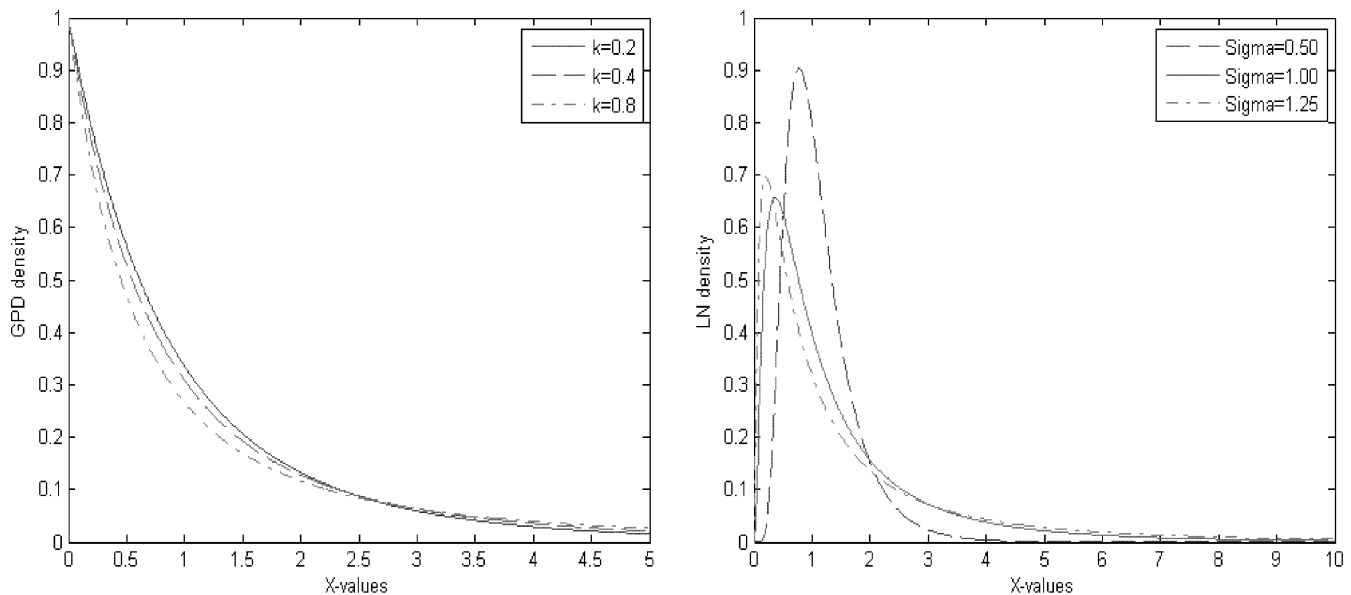


Figure 4

**Density Functions of Generalized Pareto and Lognormal Distributions Used in Simulations**



We consider four *cdf* transformations, namely, the lognormal (LN), generalized Pareto, Champernowne, and modified Champernowne. The choice of these transformations is motivated by the fact that the lognormal and generalized Pareto are commonly used to model insurance loss data (Embrechts, Klüppelberg, and Mikosch 1997; Klugman and Rioux 2006), while the Champernowne distribution approaches a form of the Pareto distribution for the extreme values (Fisk 1961) and approximates the lognormal distribution for values near zero in some cases.

For this simulation study, data are generated from GPD and LN distributions with selected parameter values that represent different shapes of the underlying distributions. For the GPD with *pdf* given by

$$f(x|k, \theta, \sigma) = \left(\frac{1}{\sigma}\right) \left(1 + k \frac{(x - \theta)}{\sigma}\right)^{-1-1/k} \tag{4.1}$$

for  $x > \theta$  when  $k > 0$ , or for  $\theta < x < -\sigma/k$  when  $k < 0$ , we consider the following sets of parameter values:  $(k = 0.20, \sigma = 1.0, \theta = 0)$ ,  $(k = 0.40, \sigma = 1.0, \theta = 0)$ , and  $(k = 0.8, \sigma = 1.0, \theta = 0)$ . For LN with the *pdf*

$$f(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0,$$

the selected parameter values are  $(\mu = 0, \sigma = 0.5)$ ,  $(\mu = 0, \sigma = 1.0)$ , and  $(\mu = 0, \sigma = 1.25)$ . Figure 4 plots the density functions for these theoretical distributions. Three sample sizes,  $n = 100, 250,$  and  $500$ , are drawn, and for each case 1,000 replications are carried out.

For each generated sample, four transformed kernel densities are obtained, as outlined and illustrated in Section 2, using the Champernowne, modified Champernowne, lognormal, and generalized Pareto *cdf* transformations. In assessing the goodness-of-fit of these estimated densities, we employ three criteria following Ait-Sahalia (1996) and Clements, Hurn, and Lindsay (2003):

$$\text{Global distance measure} = \frac{1}{n} \sum_{i=1}^n [\hat{f}(x_i) - f(x_i)]^2,$$

$$L^1 \text{ norm} = \int_0^\infty |\hat{f}(x) - f(x)| dx,$$

$$L^2 \text{ norm} = \left[ \int_0^\infty |\hat{f}(x) - f(x)|^2 dx \right]^{1/2},$$

where  $f(x)$  is the true distribution from which the data are generated. Note that Ait-Sahalia's global distance measure considers discrepancies between the estimated and underlying densities at the empirical data points, whereas the  $L^1$  and  $L^2$  norms consider discrepancies over the support of the underlying distribution. As pointed out by Clements, Hurn, and Lindsay (2003) the  $L^1$  norm is more sensitive to errors in density estimation at the tail of the distribution than the  $L^2$  norm. By considering errors of estimation only at actual data points, the Ait-Sahalia measure takes into account the likelihood of the error occurring.

Table 2 presents the percentages of time the transformed kernel density outperforms the classical (untransformed) kernel estimation according to the three criteria. The results show that that the *cdf*-transformed kernel generally does better than the classical kernel. The advantage of transformation seems to depend on the heaviness of the tail of the underlying distribution. For example, when data

Table 2

**Proportion of Time That the Transformed Kernel Outperforms Classical Kernel Density Estimation Using the Global Distance Criterion,  $L^1$  and  $L^2$  Norms<sup>a</sup>**

cdf Transformation	n	LN( $\mu, \sigma$ )		
		(0.0, 0.50)	(0.0, 1.00)	(0.0, 1.25)
Champernowne	100	46.7(48.8, 49.3)	65.2(63.1, 68.8)	84.8(87.7, 87.4)
	250	45.8(50.0, 48.6)	71.4(69.6, 77.6)	92.3(95.2, 95.0)
	500	48.0(52.9, 51.5)	79.5(76.0, 86.0)	96.2(98.4, 98.5)
Modified Champernowne	100	44.6(46.2, 46.3)	63.2(60.1, 65.8)	61.4(62.0, 62.4)
	250	45.5(49.5, 48.5)	72.9(70.4, 78.8)	78.7(80.2, 80.7)
	500	48.2(52.5, 52.0)	80.8(77.5, 86.9)	91.0(92.2, 93.1)
GPD	100	80.4(92.2, 92.6)	90.5(92.5, 89.2)	94.6(96.2, 89.8)
	250	80.6(94.1, 93.8)	93.7(95.7, 97.5)	97.7(98.5, 94.5)
	500	82.7(94.7, 95.0)	97.7(98.7, 99.9)	99.4(99.6, 98.5)
LN	100	59.1(62.9, 62.0)	74.6(77.3, 76.3)	87.7(91.2, 88.8)
	250	59.2(63.7, 61.6)	79.9(79.1, 80.7)	91.4(94.6, 92.6)
	500	62.6(66.0, 63.9)	85.0(83.5, 85.8)	95.0(96.8, 95.7)
		GPD( $k, \sigma, \theta$ )		
		(0.2, 1.0, 0.0)	(0.4, 1.0, 0.0)	(0.8, 1.0, 0.0)
Champernowne	100	40.0(57.0, 47.1)	69.1(78.3, 71.3)	92.2(95.2, 93.3)
	250	29.4(60.5, 44.4)	72.9(86.9, 77.7)	98.4(99.2, 98.1)
	500	13.7(58.0, 30.8)	73.0(92.4, 79.2)	99.9(100.0, 100.0)
Modified Champernowne	100	68.5(61.9, 66.8)	47.7(43.7, 46.0)	38.4(35.8, 37.4)
	250	68.3(57.3, 65.7)	42.1(35.0, 40.7)	35.0(32.5, 34.6)
	500	64.2(50.1, 62.2)	37.1(26.2, 35.4)	33.2(30.0, 33.0)
GPD	100	92.4(94.7, 93.3)	93.5(95.3, 94.1)	95.7(97.2, 95.9)
	250	97.2(98.2, 97.6)	98.2(99.1, 98.3)	99.0(99.4, 99.4)
	500	99.1(99.6, 99.6)	99.7(100, 99.9)	100.0(100.0, 100.0)
LN	100	30.2(62.3, 55.2)	58.5(76.3, 70.3)	86.4(91.8, 88.5)
	250	15.7(64.1, 53.2)	48.5(82.6, 74.1)	91.8(96.6, 94.2)
	500	2.6(62.2, 45.8)	28.7(83.5, 72.0)	95.0(98.5, 96.8)

<sup>a</sup> Values within parentheses refer to  $L^1$  and  $L^2$  norms, respectively.

Table 3  
**Error Rates for Transformed and Classical Kernel Density Estimation Using the Global Distance Criterion,  $L^1$  and  $L^2$  Norms<sup>a</sup>**

cdf Transformation	n	LN ( $\mu, \sigma$ )		
		(0.0, 0.50)	(0.0, 1.00)	(0.0, 1.25)
Champernowne	100	0.0108(0.0754, 0.0925) [0.0091(0.0753, 0.0906)]	0.0053(0.0474, 0.0655) [0.0064(0.0547, 0.0783)]	0.0056(0.0452, 0.0685) [0.0100(0.0651, 0.0966)]
	250	0.0055(0.0544, 0.0679) [0.0048(0.0547, 0.0671)]	0.0030(0.0350, 0.0500) [0.0041(0.0425, 0.0637)]	0.0033(0.0339, 0.0534) [0.0068(0.0522, 0.0810)]
	500	0.0033(0.0425, 0.0535) [0.0031(0.0434, 0.0535)]	0.0019(0.0278, 0.0409) [0.0030(0.0357, 0.0555)]	0.0022(0.0273, 0.0446) [0.0052(0.0441, 0.0718)]
Modified Champernowne	100	0.0114(0.0783, 0.0953) [0.0091(0.0749, 0.0912)]	0.0057(0.0511, 0.0711) [0.0064(0.0554, 0.0798)]	0.0180(0.0808, 0.1129) [0.0100(0.0675, 0.1009)]
	250	0.0055(0.0549, 0.0682) [0.0048(0.0546, 0.0670)]	0.0030(0.0354, 0.0509) [0.0041(0.0426, 0.0640)]	0.0086(0.0518, 0.0759) [0.0068(0.0536, 0.0837)]
	500	0.0033(0.0424, 0.0533) [0.0031(0.0433, 0.0535)]	0.0019(0.0278, 0.0408) [0.0030(0.0357, 0.0553)]	0.0040(0.0340, 0.0533) [0.0052(0.0448, 0.0730)]
GPD	100	0.0072(0.0604, 0.0760) [0.0091(0.0718, 0.0889)]	0.0036(0.0495, 0.0801) [0.0064(0.0632, 0.0956)]	0.0038(0.0437, 0.0808) [0.0100(0.0691, 0.1055)]
	250	0.0036(0.0415, 0.0535) [0.0048(0.0527, 0.0661)]	0.0020(0.0387, 0.0654) [0.0041(0.0519, 0.0830)]	0.0022(0.0343, 0.0683) [0.0068(0.0569, 0.0916)]
	500	0.0022(0.0316, 0.0413) [0.0031(0.0420, 0.0531)]	0.0014(0.0324, 0.0561) [0.0030(0.0451, 0.0752)]	0.0015(0.0289, 0.0607) [0.0052(0.0492, 0.0833)]
LN	100	0.0075(0.0614, 0.0766) [0.0091(0.0754, 0.0918)]	0.0039(0.0395, 0.0559) [0.0064(0.0554, 0.0778)]	0.0042(0.0381, 0.0588) [0.0100(0.0666, 0.0979)]
	250	0.0040(0.0454, 0.0574) [0.0048(0.0548, 0.0672)]	0.0022(0.0296, 0.0430) [0.0041(0.0429, 0.0626)]	0.0024(0.0288, 0.0458) [0.0068(0.0532, 0.0814)]
	500	0.0024(0.0361, 0.0459) [0.0031(0.0434, 0.0536)]	0.0014(0.0238, 0.0352) [0.0030(0.0359, 0.0542)]	0.0016(0.0233, 0.0382) [0.0052(0.0450, 0.0716)]
		GPD(k, $\sigma, \theta$ )		
		(0.2, 1.0, 0.0)	(0.4, 1.0, 0.0)	(0.8, 1.0, 0.0)
Champernowne	100	0.0228(0.0775, 0.1290) [0.0195(0.0819, 0.1247)]	0.0178(0.0670, 0.1170) [0.0223(0.0878, 0.1368)]	0.0110(0.0518, 0.0965) [0.0281(0.0990, 0.1614)]
	250	0.0181(0.0619, 0.1120) [0.0148(0.0664, 0.1072)]	0.0136(0.0530, 0.1012) [0.0173(0.0728, 0.1197)]	0.0070(0.0392, 0.0797) [0.0227(0.0852, 0.1455)]
	500	0.0160(0.0541, 0.1035) [0.0119(0.0565, 0.0950)]	0.0119(0.0463, 0.0935) [0.0141(0.0626, 0.1073)]	0.0056(0.0336, 0.0727) [0.0189(0.0753, 0.1331)]
Modified Champernowne	100	0.0187(0.0900, 0.1277) [0.0195(0.0874, 0.1353)]	0.0413(0.1314, 0.1833) [0.0223(0.0916, 0.1447)]	0.0841(0.1836, 0.2563) [0.0281(0.0997, 0.1633)]
	250	0.0130(0.0739, 0.1101) [0.0148(0.0724, 0.1197)]	0.0318(0.1181, 0.1648) [0.0173(0.0773, 0.1294)]	0.0815(0.1811, 0.2510) [0.0227(0.0860, 0.1477)]
	500	0.0110(0.0680, 0.1025) [0.0119(0.0625, 0.1083)]	0.0266(0.1110, 0.1534) [0.0141(0.0673, 0.1177)]	0.0720(0.1699, 0.2351) [0.0189(0.0763, 0.1350)]
GPD	100	0.0082(0.0542, 0.0867) [0.0195(0.0884, 0.1373)]	0.0079(0.0928, 0.1470) [0.0223(0.0928, 0.1470)]	0.0075(0.0462, 0.0822) [0.0281(0.1008, 0.1653)]
	250	0.0052(0.0410, 0.0702) [0.0148(0.0726, 0.1202)]	0.0050(0.0385, 0.0687) [0.0173(0.0776, 0.1302)]	0.0046(0.0347, 0.0664) [0.0227(0.0869, 0.1493)]
	500	0.0039(0.0344, 0.0620) [0.0119(0.0625, 0.1085)]	0.0038(0.0324, 0.0609) [0.0141(0.0674, 0.1181)]	0.0035(0.0292, 0.0589) [0.0189(0.0770, 0.1370)]
LN	100	0.0279(0.0812, 0.1377) [0.0195(0.0895, 0.1340)]	0.0230(0.0709, 0.1262) [0.0223(0.0949, 0.1455)]	0.0152(0.0555, 0.1073) [0.0281(0.1047, 0.1685)]
	250	0.0225(0.0649, 0.1177) [0.0148(0.0732, 0.1156)]	0.0185(0.0565, 0.1078) [0.0173(0.0791, 0.1276)]	0.0114(0.0430, 0.0899) [0.0227(0.0904, 0.1520)]
	500	0.0204(0.0571, 0.1090) [0.0119(0.0622, 0.1022)]	0.0168(0.0497, 0.0999) [0.0141(0.0682, 0.1141)]	0.0101(0.0375, 0.0829) [0.0189(0.0799, 0.1388)]

<sup>a</sup> Values within parentheses refer to  $L^1$  and  $L^2$  norms, respectively, and values in square brackets refer to classical kernel density estimation.

are generated by the relatively short-tailed  $LN(0.0, 0.5)$  for  $n = 250$ , the  $LN$   $cdf$ -transformed kernel outperforms the untransformed kernel in 59.2(63.5, 61.6) percent of the cases, according to the three criteria. However, if data had come from a heavier-tailed  $LN(0.0, 1.25)$ , the corresponding percentages are 91.4(94.6, 92.6).

On comparing the different  $cdf$  transformations, it seems that transforming the data with the  $cdf$  of the underlying data-generating process yields good density estimates, even for short-tailed data. This can be seen from the relatively high percentages reported in Table 1 for the cases of  $LN$  and  $GPD$  transformations when data came from the  $LN$  and  $GPD$  distributions, respectively.

Among the four  $cdf$  transformations, the  $GPD$  performs best. For all the sample sizes, parameter values, and both underlying data-generating distributions considered,  $GPD$   $cdf$  transformation yields the highest percentages in Table 2. Even when the data-generating distribution is  $LN$ , it gives higher percentages than the  $LN$   $cdf$  transformation. This shows the robustness of the  $GPD$  transformation to the data-generating process.

The relative importance that each of the three measures places on the tail of the distribution is clearly illustrated by the case where data are generated by the long-tailed  $GPD(0.4, 1.0, 0.0)$ , and the  $LN$   $cdf$  transformation is employed in the transformed kernel density estimation. For  $n = 500$ , while the global distance measure shows that the transformed kernel outperforms the untransformed in only 28.7% of the time,  $L^1$  and  $L^2$  norms recorded 83.5% and 72.0%, respectively. These results suggest that relatively large discrepancies between the untransformed kernel and the underlying data-generating distribution occur at the tail of the distribution, because the global distance measure gives less weight to the tail when compared to  $L^1$  and  $L^2$ .

In Table 3 we tabulate the error rates for each of the three measures. The error rate is defined as the average discrepancies between the estimated and underlying population densities taken over 1,000 replications. The results indicate that, in general, transformed kernel performs better than the classical kernel, particularly for long-tailed data, which agrees with our earlier observation in Table 2.

When we compare the different transformations, one conclusion that stands out is that using the  $cdf$  of the data-generating process yields good results. When data are generated from  $LN$  or  $GPD$  distributions, transformed kernel densities based on  $LN$  or  $GPD$   $cdf$  often give the smallest error rates.

Finally, transformation does not always outperform nontransformation. In the cases when data are generated by  $LN(0.0, 0.50)$ ,  $GPD(0.2, 1.0, 0.0)$ , and  $GPD(0.4, 1.0, 0.0)$ , there are occasions when the classical kernel density has smaller error rates. Note that all these cases are characterized by data generated by relatively short-tailed distributions.

## 5. ANALYSIS OF SOA MEDICAL CLAIMS DATA

In this section we report our attempt to model SOA medical claims data using two parametric distributions,  $GLD$  and  $GPD$ , and the semiparametric transformed kernel density, as outlined and illustrated in Section 2. The data consist of all claim amounts exceeding \$25,000 over the period 1991–92 and are contained in the Group Medical Insurance Large Claims Database available at [www.soa.org](http://www.soa.org). Total claims are split into hospital charges and other expenses. For our analysis we consider only the total claim amount. The data set contains 75,789 observations with a mean of \$58,410. The bulk of the observations lie below \$61,320, but there are a significant number of very high claims, the largest being \$4,518,000. The data are therefore strongly skewed to the right with a skewness coefficient of 13.21.

As the data-generating process is unknown, a flexible distribution like the  $GLD$  affords much laxity to the data to identify the underlying distribution. Similarly the kernel density is a data-based approach to estimating the underlying distribution. Since our simulations show that the  $GPD$   $cdf$ -transformed kernel yields good results and is robust to the data-generating process, we employ it to model the SOA claims data. Further, in modeling exceedances above a certain threshold, we employ the generalized Pareto distribution, following a key result in EVT due to Pickands (1975) and Balkema and de Haan

(1974). We use a threshold of \$200,000 as suggested earlier by Cebrián, Denuit, and Lambert (2003) in their analysis of the same data.

The GPD distribution is estimated by maximum likelihood, while GLD is estimated by the method of percentiles. Only 1991 data are used for the estimation; the 1992 data are used as a holdout sample to assess the out-of-sample performance of the estimated models. Figure 5 shows the QQ plots for the transformed kernel density on the 1991 and 1992 data, and the exceedances above the \$200,000 threshold. The plots show a generally good-fitting transformed kernel density to the data. The only aberration is for exceedances in the 1992 data where the fitted kernel appears to exhibit a slightly thicker tail than the empirical distribution. To conserve space, we have not presented the QQ plots for the estimated GPD and GLD, the results of which are similar to those for the transformed kernel.

Table 4 compares the deciles of the estimated distributions with the empirical deciles. For both the 1991 and 1992 complete data, the empirical deciles are very close to the corresponding values of the fitted transformed kernel and GLD distributions. The largest discrepancy of 403 occurs at the 60th percentile of the transformed kernel and the 1992 data, which amounts to less than 1% of the claim size of \$46,179. The average absolute discrepancies between the empirical deciles and the estimated deciles for 1991 data are 109.78 and 192.44, for the transformed kernel and GLD, respectively. These discrepancies amount to less than  $\frac{1}{2}\%$  of the average empirical deciles. Similarly the results for 1992

Figure 5

**QQ Plot of Fitted Kernel Density for SOA 1991 and 1992 Claims Data with and without Threshold**

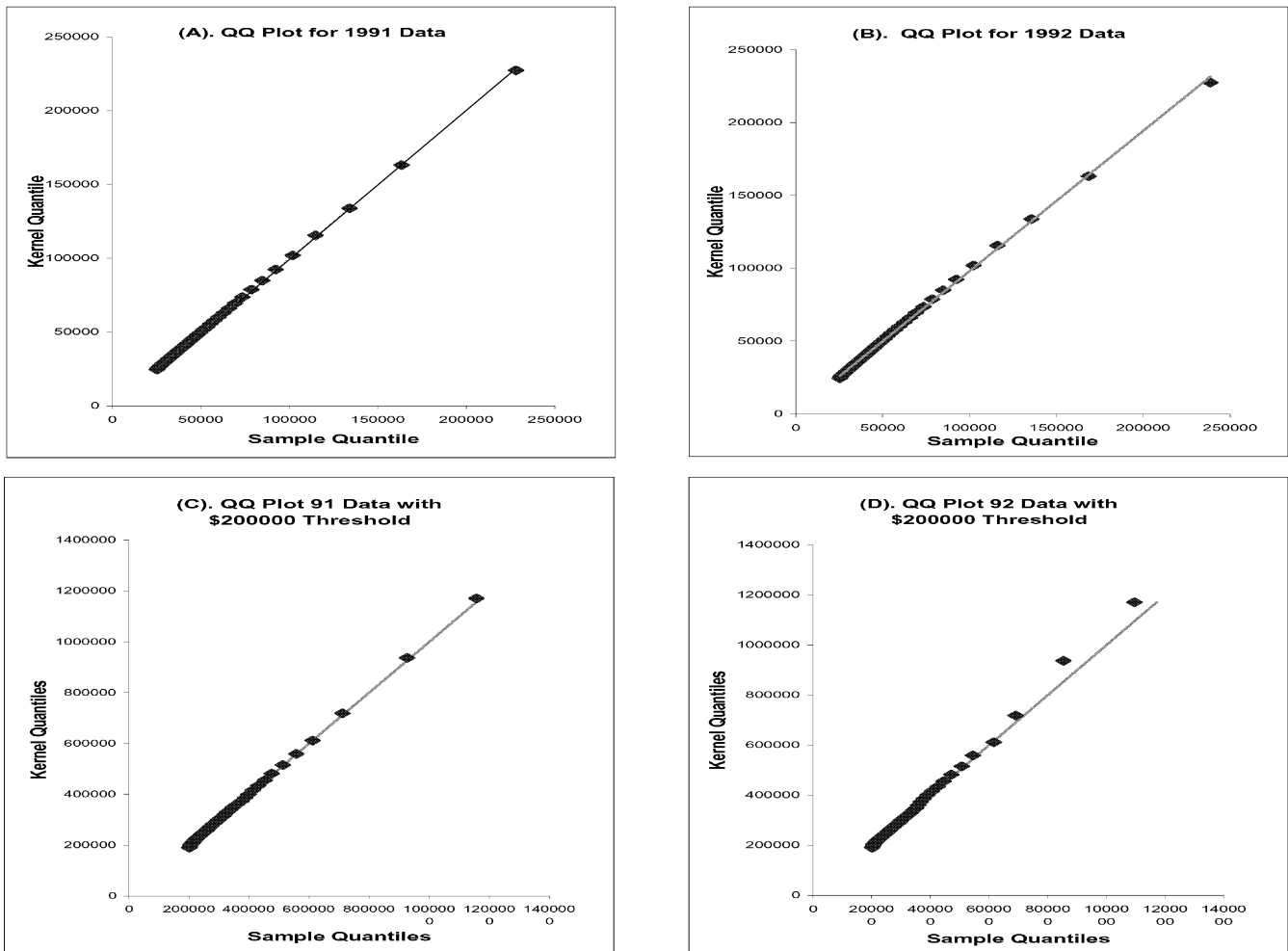


Table 4

**Comparison of Selected Quantiles with Empirical Quantiles of the SOA Data, Transformed Kernel Density, Generalized Pareto, and Generalized Lambda Fits**

	No Threshold								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
	Empirical 91	26,874	29,181	32,029	35,625	40,224	46,470	55,435	69,332
Empirical 92	26,892	29,175	31,939	35,502	40,107	46,179	54,958	69,299	102,301
Kernel	26,867	29,285	32,120	35,689	40,320	46,582	55,478	69,524	102,124
GLD	26,956	29,430	32,284	35,752	40,222	46,237	55,064	69,679	102,062
	Threshold 200,000								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Empirical 91	210,786	222,197	235,941	253,313	272,459	298,157	337,366	398,172	512,458
Empirical 92	210,328	222,590	237,894	254,707	275,364	301,139	341,418	389,864	507,470
Kernel	207,795	222,616	237,134	253,774	274,067	300,594	339,124	400,326	514,332
GPD	210,258	221,750	235,455	252,158	272,736	299,742	337,273	396,534	518,241
GLD	210,450	222,305	236,276	252,994	273,169	299,102	334,520	390,032	506,684

data are equally encouraging: average absolute discrepancies are also less than  $\frac{1}{2}\%$  of the average empirical deciles.

For exceedances, the empirical deciles are also closely estimated by the estimated densities. Average absolute discrepancies amount to about 1% of the average empirical deciles for both 1991 and 1992 data. Provided that \$200,000 is an appropriate threshold, EVT shows that the underlying data-generating process is generalized Pareto. Both the transformed kernel and GLD approximate the GPD very well, although the GPD exhibits a slightly thicker tail at the extreme right end of the distribution.

For reinsurance purposes, actuaries are often concerned with a stop-loss premium, that is, the net cost of providing coverage beyond a stop-loss limit. If  $X$  is the claim amount and  $L$  is the stop-loss limit, then the exceedance above the stop-loss premium,  $Y$ , is given by

$$Y = \begin{cases} 0 & \text{for } X < L \\ X - L & \text{for } X \geq L \end{cases}$$

The stop-loss premium is then  $E[Y]$ . Using a stop-loss limit of \$200,000, the transformed kernel and the GLD estimated a stop-loss premium of \$3,282 and \$2,790, respectively, compared to the corresponding empirical value of \$3,644 for 1991 data.

In summary, the results show that the GPD *cdf*-transformed kernel density and GLD fit the full range of SOA medical claims data well. The encouraging results obtained here for such highly skewed data suggest that both these models may find wider applications to claims of other classes of insurance.

## 6. CONCLUDING REMARKS

In this article we attempt to shed light on modeling long-tailed data using the parametric generalized lambda distribution and a semiparametric transformed kernel density function. Our simulation results show that both these models seem promising in modeling long-tailed data. We have shown that although the transformed kernel generally outperforms the classical kernel, especially in estimating the tail of a distribution, one needs to be careful in selecting the particular *cdf* for the transformation. For highly skewed data, long-tailed *cdf* transformations seem to perform better than short-tailed ones. Among the four transformations we considered, the GPD *cdf* transformation is the most robust with respect to the underlying data generation process.

We illustrate the usefulness of the transformed kernel and GLD with the SOA medical claims data. Both models fit the empirical data well over the entire range of the data. Even for exceedances above a \$200,000 threshold where, by the extreme value theory, GPD is likely to approximate the underlying

data generation process closely, the transformed kernel and GLD provide good fit to the data. For rating purposes and for other actuarial applications, it is necessary to estimate the entire loss distribution, as opposed to just the tail of the distribution. To this end, we have shown in this article two approaches that practicing actuaries may find useful in a wide varieties of contexts.

In our simulation study in Section 4, the goodness-of-fit measures compare the fitted distributions with the underlying theoretical distribution. When the data are truncated and/or censored, these measures remain valid, as the data curtailment does not affect the underlying distribution. However, in practice, if observed data are truncated and/or censored, and the investigator is faced with the problem of choosing the most appropriate model among a set of competing models, the investigator could use the usual model selection criteria. For a good discussion of the performance of some of these criteria in choosing loss distributions, see Klugman and Rioux (2006).

## 7. ACKNOWLEDGMENT

We thank the editor and two anonymous referees for their helpful comments and suggestions that have improved this article.

This research was partially supported by the Nanyang Technological University ACRF Grant, Singapore.

## REFERENCES

- AIT-SAHALIA, Y. 1996. Testing Continuous-Time Models of the Spot Interest Rate. *Review of Financial Studies* 9(2): 385–426.
- BAIN, L. J., AND M. ENGELHARDT. 1980. Probability of Correct Selection of Weibull versus Gamma Based on Likelihood Ratio. *Communications in Statistics—Theory and Methods* A9: 375–81.
- BALKEMA, A. A., AND L. DE HAAN. 1974. Residual Life Time at Great Age. *Annals of Statistics* 2(5): 792–804.
- BOLANCÉ, C., M. GUILLÉN, T. BUCH-LARSEN, AND J. P. NIELSEN. 2005. Kernel Density Estimation for Heavy-Tailed Distributions Using the Champernowne Transformation. In *Contributed Papers: International Seminar on Nonparametric Inference*, ISNI 2005 International Seminar on Nonparametric Inference, A Coruña, Spain, 13–15 July.
- BOLANCÉ, C., M. GUILLÉN, AND J. P. NIELSEN. 2003. Kernel Density Estimation of Actuarial Loss Functions. *Insurance: Mathematics and Economics* 32: 19–36.
- BOWMAN, A. 1984. An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika* 71: 353–60.
- BOWMAN, A. W., AND A. AZZALINI. 1997. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press.
- CEBRÍAN, A. C., M. DENUIT, AND P. LAMBERT. 2003. Generalized Pareto Fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal* 7(3): 18–36.
- CHIU, S. T. 1991. Bandwidth Selection for Kernel Density Estimation. *Annals of Statistics* 19: 1883–1905.
- . 1992. An Automatic Bandwidth Selector for Kernel Density Estimation. *Biometrika* 79: 771–82.
- CLEMENTS, A., S. HURN, AND K. LINDSAY. 2003. Möbius-Like Mappings and Their Use in Kernel Density Estimation. *Journal of the American Statistical Association* 98: 993–1000.
- DUDEWICZ, E. J. 1999. Basic Statistical Methods. In *Juran's Quality Handbook*, 5th edition, edited by J. M. Juran, A. Blanton, and R. E. Godfrey, chapter 44. New York: McGraw-Hill.
- EMBRECHTS, P., C. KLÜPPELBERG, AND T. MIKOSCH. 1997. *Modelling Extremal Events for Insurance and Finance*. New York: Springer.
- FISK, P. R. 1961. The Graduation of Income Distributions. *Econometrica* 29(2): 171–85.
- JONES, M. C., J. S. MARRON, AND B. U. PARK. 1991. A Simple Root  $n$  Bandwidth Selector. *Annals of Statistics* 19: 1919–32.
- KARLAN, Z. A., AND E. J. DUDEWICZ. 2000. *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*. Boca Raton, FL: Chapman & Hall/CRC Press.
- KLUGMAN, S. A., AND J. RIOUX. 2006. Toward a Unified Approach to Fitting Loss Models. *North American Actuarial Journal* 10(1): 63–83.
- MCNEIL, A. J., AND T. SALADIN. 1997. The Peaks over Thresholds Methods for Estimating High Quantiles of Loss Distributions. In *Proceedings of the 28th International ASTIN Colloquium*, pp. 23–43, Cairns, Australia: Peeters.
- MUDHOLKAR, G. D., AND M. V. PHATAK. 1984. Quantile Function Models for Quantal Response Analysis: An Outline. In *Topics in Applied Statistics, Proc. Stat. 81 Canada*, pp. 621–27. Montreal: Concordia University Press.
- ÖZTÜRK, A., AND R. F. DALE. 1982. A Study of Fitting the Generalized Lambda Distribution to Solar Radiation Data. *Journal of Applied Meteorology* 21: 995–1004.
- PARK, B. U., AND J. S. MARRON. 1990. Comparison of Data-Driven Bandwidth Selectors. *Journal of the American Statistical Association* 85: 66–72.

- PICKANDS, J. 1975. Statistical Inference Using Extreme Order Statistics. *Annals of Statistics* 3(1): 119–31.
- PREGIBON, D. 1980. Goodness of Link Tests for Generalized Linear Models. *Applied Statistics* 29: 15–24.
- RAMBERG, J. S., AND B. W. SCHMEISER. 1972. An Approximate Method for Generating Symmetric Random Variables. *Communications of the ACM* 15: 987–90.
- . 1974. An Approximate Method for Generating Asymmetric Random Variables. *Communications of the ACM* 17: 78–82.
- RAMBERG, J. S., P. R. TADIKAMALLA, E. J. DUDEWICZ, AND E. F. MYKYTKA. 1979. A Probability Distribution and Its Uses in Fitting Data. *Technometrics* 21: 201–14.
- ROOTZÉN, H., AND N. TAJVIDI. 1996. Extreme Value Statistics and Wind Storm Losses: A Case Study. *Scandinavian Actuarial Journal*, 70–94.
- SHEATHER, S. J., AND M. C. JONES. 1991. A Reliable Data-Based Bandwidth Selection for Kernel Density Estimation. *Journal of the Royal Statistical Society B* 53: 683–90.
- SILVER, E. A. 1977. A Safety Factor Approximation Based upon Tukey's Lambda Distribution. *Operational Research Quarterly* 28(3): 743–46.
- SILVERMAN, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.
- STONE, C. J. 1984. An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *Annals of Statistics* 12: 1285–97.
- TUKEY, J. 1960. The Practical Relationship between the Common Transformations of Percentages of Counts and of Amounts. Technical Report 36, Statistical Techniques Research Group, Princeton University.
- WAND, M. P., AND E. J. JONES. 1995. *Kernel Smoothing*. Boca Raton, FL: Chapman & Hall/CRC Press.
- WAND, M. P., J. S. MARRON, AND D. RUPPERT. 1991. Transformations in Density Estimation. *Journal of the American Statistical Association* 86: 343–53.

*Discussions on this paper can be submitted until October 1, 2008. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.*