Interpolation Along a Curve

by Joseph Boor

ABSTRACT

Actuaries quite often have to interpolate data to obtain quantities such as loss development factors (LDFs) for maturities in between the maturities included in a loss development triangle, or increased limits factors for limits between the data points used in the increased limits analysis. This paper presents an approach that includes the advantages of using fitted curves for non-linear data, and that avoids the errors arising from mismatches between patterns in the data and patterns inherent to the curve family used for interpolation.

KEYWORDS

Interpolation, loss development, increased limits, curve fitting

1. Introduction

There are several instances in which actuaries are either forced to use fitted curves to interpolate data or find them preferable to other forms of analysis for interpolation. For example, when analyzing loss development for medium tail lines of business, it is not unusual to need to convert December-to-ultimate loss development factors (LDFs) to factors suitable for use with June data. It is not unusual to perform an increased limits analysis at a few choice points and then interpolate the remaining points.

There are concerns with the use of interpolation on insurance claims data. In many of the scientific disciplines, there are overriding laws that create specific forms for mathematical functions that underlie data points. So, given such a circumstance, all one need do is to determine which member of a family of functions is the specific one that underlies the data, then determine the intermediate values directly from the fitted curve.

However, in casualty actuarial science, there is generally no such functional family or form that perfectly describes either the pattern of loss development or the relative frequency of losses by size. Actuaries often attempt to use such models anyway. The Weibull curve is often used to model the loss emergence underlying loss development factors (see Heyer 2001 for an example). Many curve families, including the Pareto, lognormal, and gamma (among others) have been used to model loss severity distributions (Hogg and Klugman 1984). Those models are widely known to capture the general character of the underlying development or severity phenomenon. But they are also known to be very imperfect estimators of the values at any particular point. One need only compare the development triangle of a large insurance company with fully credible data to any Weibull curve to see that the fitted curve fits the data imperfectly. Similarly, typical fully credible short tail data by size rarely follows a perfect mathematical severity curve. Therefore, in the less-than-perfect situations of loss development triangles with intermediate credibility, and of long-tail loss severity data with limited credibility in the upper tail, one would not expect perfect compliance with any common mathematical curve.

The problem that arises in current usage involves a trade-off of errors. Using the fitted curve values directly as interpolated values creates a distortion because the curve is an imperfect fit. Specifically, the curve will not match the known values of the phenomenon being analyzed. In other words, when the goal is to interpolate from loss development factors or increased limits factors carefully and reliably calculated at a limited set of points, the fitted curve typically does not reproduce those carefully and reliably determined data points. On the other hand, one could reproduce those data points exactly by simply using linear interpolation between the known values. One might even use the exponential interpolation for loss payouts espoused in Berquist and Sherman (1977). But, since either interpolation process just involves two data points, it will likely miss the general overall shape of the curve that the fitted curve captures. This paper offers an alternative that incorporates what is best in both approaches, as an expansion of an analysis presented conceptually in a paper by Boor (2006).

2. The formula for interpolation along the curve

This alternative method begins by fitting a curve from the curve family that is expected to be close to the pattern underlying the data. Then each segment (between two adjacent actual data points) of the fitted curve is adjusted so that the curve exactly matches the two actual data points. In other words, if we have actual data points $d(t_0)$, $d(t_1)$, $d(t_2)$, . . . , $d(t_m)$ and a curve fitted to those points of $g(t)$, and we desire an estimate at t^*, $t_a < t^* < t_{a+1}$, $a \in$ 0, 1, 2, . . . , $m - 1$, we take

$$\hat{d}(t^*) = d(t_a) + \frac{g(t^*) - g(t_a)}{g(t_{a+1}) - g(t_a)}[d(t_{a+1}) - d(t_a)]. \quad (1)$$

So, the curve is used as a guidepost for the changes between the observed points, and the actual observed values are maintained.

Further, if the function approaches zero at infinity, values beyond the last observed data point may be obtained by extrapolation. If t_m is the location of the last observation and $t^* > t_m$, then we get

$$\hat{d}(t^*) = g(t^*) \times \frac{d(t_m)}{g(t_m)},\ t^* > t_m, \tag{2}$$

which fits the observed values $d(t_m)$ and $d(\infty) = 0$ exactly. So, a very nice property of this approach is that it allows for extrapolation[1] as well as interpolation. In fact, an astute observer can see that formula (2) is just a special case of equation (1) where $d(\infty) = g(\infty) = 0$. In fact, as long as $d(\infty)$ and $g(\infty)$ are known, finite and close enough to use g as a basis for approximation with large numbers, formula (2) may be used to extrapolate towards infinity when $d(\infty) \neq 0$.

3. Interpolating loss development factors along a Weibull curve

The example below shows how this approach may be used in practice. The starting point is a series of loss development factors. It should be clear that asymptotically the loss development factors will converge to unity, since all the claims must eventually be closed. And it is recognized within the actuarial profession that in most circumstances the loss development factors decline monotonically toward unity as the time since inception gets larger. But the loss development factor at time zero is infinity, not zero (since typically no claims are even reported at the beginning of the process). Further, as time wears on, the loss development factors approach unity, not zero. So, the loss development factors themselves are not an ideal candidate for this type of interpolation/extrapolation.

But the loss development factors may be readily adapted to the approach in this paper. First, one should divide unity by each loss development factor. That produces the percentage of ultimate loss that is expected to be reported at 12, 24, etc. months into the process. Then, the percentage of loss dollars that are "IBNR" at a given time value may be computed as

$$\%IBNR = 1 - \frac{1}{\text{loss development factor}}. \tag{3}$$

So the "%IBNR" values across time begin at unity and decline monotonically to zero. Therefore, IBNR percentages are a better candidate for the approach in this paper.

Next, the question of what class of reference functions are to be used for the interpolation. The type of function regularly used by actuaries for this particular situation is the family of Weibull distributions, with

$$\%Reported(t) = 1 - \%IBNR(t) \approx 1 - \exp(-c \times t^b), \tag{4}$$

and

$$\%IBNR(t) \approx \exp(-c \times t^b). \tag{5}$$

A strong rationale for the use of the Weibull distribution is its match to the general reporting pattern of claims dollars. The incremental pattern, or the tendency for claims dollars to be newly reported at a particular time, is the derivative of the cumulative distribution function in equation (4), or

$$\text{Incremental }\%\text{Reported}(t) = cbt^{b-1}\exp(-c \times t^b). \tag{6}$$

So, this distribution starts low at zero, picks up size as t^{b-1} increases, reaches a maximum, and then slowly tails off downward. That behavior is typical of the reporting pattern of insurance claims. So, the Weibull is a good choice for the family of interpolating distributions. Next, it is necessary to compute the coefficients b and c that define the specific Weibull distribution used in the interpolation. One may first take the natural logarithms of both sides of equation (5) to get

$$\ln(\%IBNR(t)) \approx -c \times t^b. \tag{7}$$

[1] In all but a small minority of actuarial analyses, extrapolation below zero is not needed.

Then the sign of both sides may be switched, and another natural logarithm applied to produce

$$\ln(-\ln(\%IBNR(t))) \approx \ln(c) + b\ln(t). \tag{8}$$

Then, ln(c) may be estimated as the intercept of a regression line and b as the slope of the same regression line.[2] The value of c may be determined by simply using the exponential function to invert the logarithm.

One may see the process of conversion to %IBNR, to curve fitting, and then converting back to loss development factors in Table 1. The c and b parameters are estimated by transforming the time and %IBNR values as noted in the table with the precise computations noted at the bottom.

For the sake of completeness, two technical items deserve mention that do not directly relate to the process of interpolating along the fitted curve, but do relate to the process of interpolating loss development factors. First, the number of months since the accident year began in the first column is converted to the average number of months of maturity of the losses in the data. This makes it possible to determine a loss development factor for an odd-shaped year. For example, if a company began writing policies at the beginning of the calendar year and sold them at a continuous rate, their accident year would have a triangular shape weighted towards the end of the year, with an average loss maturity of four months. So, the loss development factor corresponding to four months maturity could be used for that company.

Second, an adjustment is included to convert the loss development factors for stub periods (periods of loss data of less than a year in duration) to development factors suitable for a full year. This adjustment is necessary because the basis for all the fitted factors are loss periods that do not include losses that have not occurred as yet. Field experience with this methodology suggests that the adjustment is indeed necessary.

Figure 1 illustrates the relationship between the fitted IBNR, the values interpolated along the curve, and the data points that are being interpolated. The thin line represents the fitted curve, the thick line represents the values interpolated along the curve, and the diamonds are the actual values of the data that are to be interpolated. As one may see, the fitted curve provides the general shape, but the interpolation hits the actual values exactly.

4. Interpolating increased limits factors along a Pareto curve

Next, an example involving increased limits factors will be performed. In this case, the loss severity curve will be assumed to follow a Pareto distribution.[3] A little background is needed, though, to introduce the Pareto-based function for the increased limits factors. Mathematically, if the severity distribution is a Pareto distribution with parameter α and truncation point T, then, using the notation in Boor (2012) of $E[X_c(L)]$ for the mean value of losses capped at L, and $s_X(x)$ and $F_X(x)$ for the loss severity function and its cumulative distribution function,

$$E[X_c(L)] = \int_0^L x s_x(x)\,dx + L(1 - F_X(L)). \tag{9}$$

Substituting in the specific functions associated with the Pareto distribution yields

$$\begin{aligned} E[X_c(L)] &= \int_T^L x\alpha \frac{T^\alpha}{x^{\alpha+1}}\,dx + L\frac{T^\alpha}{L^\alpha} \\ &= \frac{T}{\alpha-1}\left[\alpha - \frac{T^{\alpha-1}}{L^{\alpha-1}}\right]. \end{aligned} \tag{10}$$

Then, if $ILF(L, B)$ represents the increased limits factor relating costs associated with a limit of L to costs associated with a basic limit B, it has the value

$$ILF(L,B) = \frac{\alpha - \left(\frac{T}{L}\right)^{\alpha-1}}{\alpha - \left(\frac{T}{B}\right)^{\alpha-1}}. \tag{11}$$

[2] This regression method for determining the Weibull coefficients is not espoused to be new or innovative. It is not espoused to be developed by this author. Actuaries working in this area are aware that this method has been in use for at least 25 years.

[3] Review of fits to the sample data using the one-parameter Pareto distribution suggests that it often provides a poor fit to the data points. Although the curve is adjusted to exactly fit the points, the one parameter was rejected on aesthetic grounds.

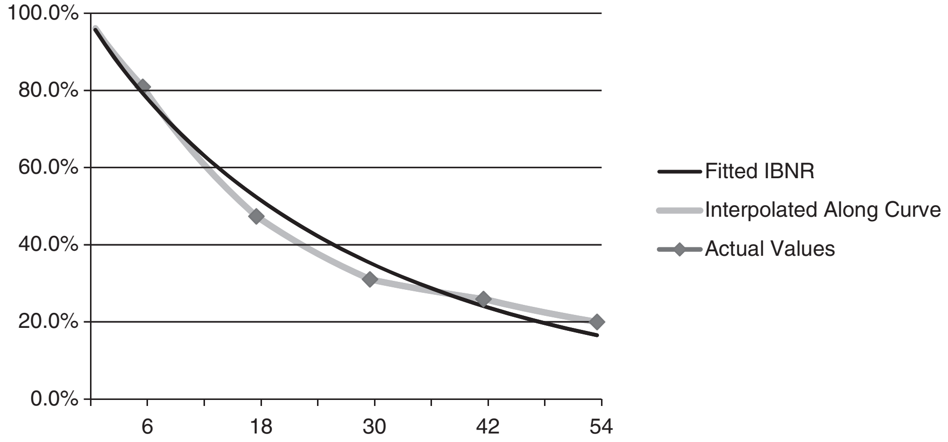
Table 1. Example: Interpolation of loss development factors using interpolation of %IBNR along complement of Weibull cumulative distribution

Months Since AY Began	Avg. Loss Maturity in Mos.	Actual L.D.F.	Actual % Reported	Actual % IBNR	ln (IBNR)	ln (–ln (IBNR))	ln (Mos.)	Fitted Weibull IBNR	Interpolated Along Curve	Implied L.D.F. for Maturity	Implied Full AY LDF
1	0.5							97.7%	97.9%	47.806	573.673
2	1							95.7%	96.1%	25.349	152.095
3	1.5							93.8%	94.3%	17.559	70.234
4	2							92.0%	92.6%	13.567	40.701
5	2.5							90.2%	91.0%	11.129	26.710
6	3							88.5%	89.5%	9.481	18.963
7	3.5							86.8%	87.9%	8.291	14.213
8	4							85.2%	86.5%	7.389	11.084
9	4.5							83.7%	85.0%	6.682	8.910
10	5							82.2%	83.6%	6.112	7.335
11	5.5							80.7%	82.3%	5.643	6.156
12	6	5.25	19.0%	81.0%	–0.211	–1.554	1.792	79.2%	81.0%	5.250	5.250
13	7							76.4%	77.5%	4.437	4.437
14	8							73.8%	74.1%	3.864	3.864
15	9							71.2%	70.9%	3.440	3.440
16	10							68.8%	67.9%	3.113	3.113
17	11							66.4%	64.9%	2.852	2.852
18	12							64.2%	62.1%	2.640	2.640
19	13							62.0%	59.4%	2.464	2.464
20	14							59.9%	56.8%	2.316	2.316
21	15							57.9%	54.3%	2.189	2.189
22	16							56.0%	51.9%	2.080	2.080
23	17							54.1%	49.6%	1.984	1.984
24	18	1.9	52.6%	47.4%	–0.747	–0.291	2.890	52.4%	47.4%	1.900	1.900
25	19							50.6%	45.7%	1.842	1.842
26	20							49.0%	44.1%	1.790	1.790
27	21							47.4%	42.6%	1.742	1.742
28	22							45.8%	41.1%	1.699	1.699
29	23							44.4%	39.7%	1.658	1.658
30	24							42.9%	38.3%	1.622	1.622
31	25							41.5%	37.0%	1.587	1.587
32	26							40.2%	35.7%	1.556	1.556
33	27							38.9%	34.5%	1.526	1.526
34	28							37.7%	33.3%	1.499	1.499
35	29							36.5%	32.1%	1.474	1.474
36	30	1.45	69.0%	31.0%	–1.170	0.157	3.401	35.3%	31.0%	1.450	1.450
37	31							34.2%	30.5%	1.439	1.439
38	32							33.1%	30.0%	1.429	1.429
39	33							32.1%	29.6%	1.420	1.420
40	34							31.1%	29.1%	1.410	1.410

Table 1. (*Continued*) Example: Interpolation of loss development factors using interpolation of %IBNR along complement of Weibull cumulative distribution

Months Since AY Began	Avg. Loss Maturity in Mos.	Actual L.D.F.	Actual % Reported	Actual % IBNR	ln (IBNR)	ln (–ln (IBNR))	ln (Mos.)	Fitted Weibull IBNR	Interpolated Along Curve	Implied L.D.F. for Maturity	Implied Full AY LDF
41	35							30.1%	28.7%	1.402	1.402
42	36							29.1%	28.2%	1.393	1.393
43	37							28.2%	27.8%	1.385	1.385
44	38							27.3%	27.4%	1.378	1.378
45	39							26.5%	27.0%	1.370	1.370
46	40							25.7%	26.6%	1.363	1.363
47	41							24.9%	26.3%	1.356	1.356
48	42	1.35	74.1%	25.9%	–1.350	0.300	3.738	24.1%	25.9%	1.350	1.350
49	43							23.3%	25.3%	1.339	1.339
50	44							22.6%	24.8%	1.329	1.329
51	45							21.9%	24.2%	1.320	1.320
52	46							21.2%	23.7%	1.310	1.310
53	47							20.6%	23.2%	1.302	1.302
54	48							20.0%	22.7%	1.293	1.293
55	49							19.3%	22.2%	1.285	1.285
56	50							18.8%	21.7%	1.278	1.278
57	51							18.2%	21.3%	1.270	1.270
58	52							17.6%	20.8%	1.263	1.263
59	53							17.1%	20.4%	1.256	1.256
60	54	1.25	80.0%	20.0%	–1.609	0.476	3.989	16.6%	20.0%	1.250	1.250
					Regression for Weibull parameters						
				Natural Log of *c*	–3.1240	(Intercept of –ln (–ln (IBNR)) At ln (Time)=0)					
				c	0.0440	(exp (ln (*c*)) per ln (*c*) above)					
				Power "*b*"	0.9303	(Slope of –ln (–ln (IBNR)) Across ln (Time))					

Figure 1. Example: Graph of fitted IBNR points, points interpolated along curve, and actual IBNR data points



That expression is initially undesirable, because it does not lend itself to a simple regression-type calculation of the parameters. However, most spreadsheet software in use today contains goal seek and other optimization functionalities. Those may be used in lieu of the regression approach used in fitting loss development factors.

It may help to provide an example using Pareto interpolation along a curve which requires goal seek for curve fitting. One might begin with the sample increased limits factors for \$25,000, \$100,000, \$500,000 and \$2 million in Table 2. Using equation (11), one could compute the increased limits factors that correspond to any assigned values of α and T. As is shown in Table 2, one might compute the squared differences between the known increased limits factors used as input in the process and those computed using the assigned values of α and T. Then, the sum of those squared differences may be computed, as is shown in a subsequent column in the table. The last step is simply to execute an instruction that directs the spreadsheet goal seek routine to find the values of α and T that minimize the sum of squared differences. After that, the α and T parameter values for the fitted curve are determined and may be carried to the next step.

Having the values of α and T determined, the next step is to show the fitted curve values at all the desired limit values. After that, one should rotate and scale the fitted curves in each interval using equation (1) so that they match the endpoints exactly. The result yields the desired estimates of the intermediate increased limits factors, as is illustrated in Table 3.

Table 2. Example: Determination of Pareto parameters from increased limits data

	"α" parameter to use = 1.103 Truncation point "T" to use = 15,000			
Original Limit ("*L*")	Actual ILF	Fitted Pareto Capped Expectation	Implied Fitted ILF	Squared Diff. Actual vs. Fitted
\$25,000	0.500	22,465	0.55	0.0025
\$100,000	1.000	40,857	1.00	—
\$500,000	1.500	59,171	1.45	0.0027
\$2,000,000	1.750	72,693	1.78	0.0009
Sum of Squared Differences = Minimization Target = 0.0060				

5. Mathematical rationale and summary

Again, note that a key assumption in this approach is that the data points $d(t_1)$, $d(t_2)$, . . . are fairly good approximations of the true underlying values, if not the exact underlying values. Further, it also assumes that whatever curve is fit to the data points is a reasonable approximation of the underlying pattern. Presumably, though, the fitted curve is not the true, more complex curve of the underlying phenomenon. Specifically, as long as the data points $d(t_i)$ are known to be much higher quality approximations to the true underlying phenomenon than the fitted curve values $g(t_1)$, $g(t_2)$, . . . , it is logically preferable to modify the curve to match the $d(t_i)$'s exactly.

Another key question to ask involves how the mathematics justify this approach, especially when the Taylor series-based interpolation processes such as linear and polynomial interpolation are so prominent in mathematics. It is important to address why interpolation along the curve should be an improvement, in actuarial contexts, over that class of methods. First, purely linear approximations deserve attention. Obviously, the efficacy of these methods depends on the degree to which the character of the fitted curve captures the character of the underlying data pattern. But it would appear, to the extent that the general characteristics of the fitted curve used in the interpolation match the data pattern, that in some sense the method in this paper captures aspects of the second (and possibly higher) derivatives. So, while it is unlikely that an exact second derivative match will be generated by this approximation, the accuracy resulting from this enhancement should usually be superior to that of linear interpolation. A secondary consideration results from comparing this methodology to, say, cubic splines. While the cubic splines approach has more theoretical mathematical support, the method in

Table 3. Example: Interpolation along the curve to estimate intermediate increased limits factors—given Pareto parameters for approximation

(A) Limit (*L*)	(B) Original ILF	(C) Fitted Pareto ILF α, *T* as Below	(D) Original ILF Change in Interval	(E) Fitted ILF Change in Interval	(F) Fitted Value Less Value at Bottom of Interval	(G) =Last (B) +(F) × (D)/(E) Interpolated Value Along Curve
$25,000	0.500	0.550	0.500	0.450	0.000	0.500
$50,000	—	0.783	0.500	0.450	0.233	0.759
$75,000	—	0.912	0.500	0.450	0.362	0.902
$100,000	1.000	1.000	0.500	0.448	0.000	1.000
$150,000	—	1.120	0.500	0.448	0.120	1.134
$200,000	—	1.202	0.500	0.448	0.202	1.226
$250,000	—	1.264	0.500	0.448	0.264	1.295
$350,000	—	1.355	0.500	0.448	0.355	1.396
$500,000	1.500	1.448	0.250	0.331	0.000	1.500
$750,000	—	1.550	0.250	0.331	0.102	1.577
$1,000,000	—	1.620	0.250	0.331	0.171	1.629
$1,500,000	—	1.714	0.250	0.331	0.266	1.701
$2,000,000	1.750	1.779	0.250	0.331	0.331	1.750

"α" parameter used = 1.103

Truncation point "*T*" used = 15,000

this paper will work better in an actuarial context for four reasons:

- As long as the value at infinity is finite, extrapolation along the fitted curve may be used, whereas cubic splines are not designed for extrapolation;
- Business audiences are generally not receptive to methods that could feature unexplained reversals (e.g., a fitted curve going down, then up, then down again when locally it should decrease monotonically) such as can sometimes happen with higher order approximations;
- The mathematics of this approach (at least as long as the curve fit is susceptible to easy explanation) may be presented in a relatively simple spreadsheet, whereas cubic splines require a more complex spreadsheet; and
- Testing confirms (see section 6) that interpolating along the curve generally produces more accurate values than pure curve fitting.

So, this approach represents a significant enhancement over curve fitting, linear interpolation and cubic splines in a practical actuarial context.

6. Testing

The previous sections provide a strong argument that interpolation along a curve provides an improved prediction of loss development factors and increased limits factors at intermediate input values. However, it is relevant to provide some actual field testing of the efficacy of this approach. To that end, three major data sets were collected. First, industry aggregate development triangles for various Schedule P lines of business as of 12/31/2003 were provided by (and consequential data used here by permission of) the National Association of Insurance Commissioners. Secondly, 10 Schedule P triangles, each from a company with a limited volume of development data, were selected. Lastly, the National Council on Compensation Insurance provided two sets of seven copyrighted excess loss ratios, seven for 2007 and seven for 2008 (and gave permission for their use). Using those datasets, loss development factor and increased limits factor datasets were determined. Then, one may evaluate the accuracy of various approaches to interpolation by eliminating, for example, all the even-numbered values and interpolating them using the odd-numbered values. Since their true values are known, the interpolation error could then be computed precisely.

6.1. Test using the NAIC aggregate reserving data

One advantage of the NAIC Aggregate Reserving data is that any process variance volatility in the link ratios due to inadequate premium volume is minimal to nonexistent. So, this test measures the quality of interpolation of loss development factors when the true long-term average value of the loss development factors are known. To execute this test, 2003 aggregate industry triangles using Schedule P parts 2, 3, and 4 were prepared for the homeowners, private passenger auto liability, workers compensation, commercial multi-peril, occurrence medical malpractice, claims/made medical malpractice, occurrence other liability, claims/made other liability, occurrence products liability, and claims/made products liability. The case incurred (Part 2–Part 4) and paid loss and defense, and cost containment were entered into different tabs in the same workbook. Link ratios were mechanically selected as the weighted average of the last three link ratios. Tail factors were selected in a fairly mechanical fashion using the Sherman/Boor tail factor algorithm (Boor 2006). In some cases, link ratios below unity were observed.[4] Up to three link ratios in each development pattern might be replaced by factors of "1.0001".[5] Consequently, all the paid and incurred loss development patterns could be made usable except that of the claims made medical malpractice incurred loss development. For each paid or case incurred development pattern, the factors at 12, 36, 60, and 84 months were extracted, and then a Weibull curve was fit to those extracted factors.[6] The results of that curve fit yielded the fitted curve estimates of the loss development factors at the intermediate maturities of 24, 48, 72, and 108. Further, the availability of the fitted curve values accommodated fitting along the curve as well. Geometric (exponential) interpolation, and linear interpolation of the loss development factors were performed. Lastly, linear and geometric interpolation[7] of the percentages paid or incurred were performed as well.

Since the goal was to score the various interpolating methods, two key datasets were created. First, the squared error between each estimate and the actual value from the curve was computed. Then, for each intermediate value, the interpolation method with the lowest squared error was determined.

As one may see, the interpolation along the curve has a much more consistent success percentage than the other other methods. Further, as noted in the last column of Table 4, it is not unusual for the Weibull interpolation to lie outside the endpoints of the range which are to be interpolated. Therefore, there is a strong reason to use the fitting along the curve as the benchmark interpolation approach.

To provide a more explicit accuracy test, the squared errors of each interpolation method for each target dataset were computed. Ratios of the squared error generated by the various approximations to each intermediate point to the benchmark interpolation along the curve were computed. The results were capped from below by 5% and above by 2000%,[8] and geometric averages computed across the intermediate values. The results are shown in Table 5.

As one may see, fitting along the curve is more accurate by an order of magnitude compared to the various non–curve fit methodologies. The unadjusted Weibull curve fit contains 336% as much error (180% as much standard error) as the fit along the Weibull curve described in this paper. So, the fit along the curve is demonstrably superior.

[4] Note that that link ratios below unity are contrary to the assumptions of the Weibull curve.

[5] It should be noted that use of of these very flat development factors tend to generate results that favor linear interpolation. So, should linear interpolation arise as the best estimator in conjunction with those factors, it should be viewed skeptically.

[6] Technically, the Weibull curves "$f(t)$" were fit to 1.0/LDF(t).

[7] Specifically, this would involve interpolating the percentage paid or incurred, "1/LDF(t)" to the time t_0, then dividing that percentage paid or incurred into unity to get the interpolated loss development factor.

[8] Note that 2000% is the multiplicative inverse of 5%.

Table 4. Winning percentage of intermediate LDF value estimates from various interpolation methods vs. interpolation along the curve using NAIC aggregate data—percentage of the tests in which each method was superior to interpolation along the curve

Curve Fit to:	Number of Curves Fit	Winning % of Interp. Along the Curve	Geometric Interpolation	Linear Interpolation	Linear % Pd or Incrd Interpolation	Geometric % Pd or Incrd Interpolation	Unadjusted Weibull	Number of Times Weibull Outside Range
Even Maturity Paid LDFs	11	73%	6%	6%	11%	0%	18%	1
Odd Maturity Paid LDFs	11	77%	5%	5%	9%	2%	14%	5
Even Maturity Incrd LDFs	10	58%	13%	10%	15%	3%	30%	1
Odd Maturity Incrd LDFs	10	68%	18%	15%	25%	5%	18%	3
Straight Average		69%	10%	9%	15%	3%	20%	

Table 5. Geometric average ratio of error of intermediate LDF value estimates from various interpolation methods vs. interpolation along the curve per NAIC aggregate data—geometric average of ratio of squared error of various methods to squared error of interpolation along Weibull curve (individual ratios in average capped at 2000% above and 5% below)

Curve Fit to:	Number of Curves Fit	Geometric Interpolation	Linear Interpolation	Linear % Pd or Incrd Interpolation	Geometric % Pd or Incrd Interpolation	Unadjusted Weibull
Even Maturity Paid LDFs	11	1034%	1183%	568%	1781%	316%
Odd Maturity Paid LDFs	11	1277%	2278%	548%	2931%	527%
Even Maturity Incrd LDFs	10	801%	864%	655%	1401%	190%
Odd Maturity Incrd LDFs	10	694%	943%	366%	1904%	386%
Straight Average		935%	1235%	524%	1947%	336%

6.2. Test using the small company reserving data

As a contrast to the high data volume displayed in the NAIC aggregate data, 10 small company triangle sets were identified. All the paid loss and DCC triangles were usable, but only seven of the then incurred loss triangles could be used. Errors were computed as before, and the interpolation methods with the lowest errors are shown in Table 6.

As one may see, interpolation along the curve does not have as much benefit for small company data. Further, Table 7 shows the average error relative to the interpolation along the curve. Interpolation along the curve still produces benefits here, but it does not provide the same error reduction that it does with industry aggregate data. This suggests that interpolation along the curve tends to work best with larger, more reliable volumes of data.

6.3. Test using NCCI increased limits factors

As noted earlier, interpolation along the curve may be used to interpolate increased limits factors as well as loss development factors. Therefore, it makes sense to test interpolation along a curve in the context of estimating increased limits factors. To that end, the National Council on Compensation Insurance supplied tables of excess loss factors for Florida from 2007 and 2008, along with permission to use them. On review of the data, it was apparent that the 2008 factors were so close to the 2007 factors that they could not truly represent an independent test of the accuracy of the various interpolation methods. The NCCI data is available at a wide range of attachments and for seven hazard groups. To simulate the classic ILF interpolation problem, the excess factors were converted to ILFs centered at $250,000. For each of the seven hazard groups, ILFs were extracted for $25,000,

Table 6. Winning percentage of intermediate LDF value estimates from various interpolation methods vs. interpolation along the curve from small company data—percentage of the tests in which each method was superior to interpolation along the curve

Curve Fit to:	Number of Curves Fit	Winning % of Interp. Along the Curve	Geometric Interpolation	Linear Interpolation	Linear % Pd or Incrd Interpolation	Geometric % Pd or Incrd Interpolation	Unadjusted Weibull	Number of Times Weibull Outside Range
Even Maturity Paid LDFs	10	38%	27%	23%	33%	7%	43%	3
Odd Maturity Paid LDFs	10	50%	20%	20%	25%	8%	30%	4
Even Maturity Incrd LDFs	7	43%	38%	33%	29%	19%	36%	3
Odd Maturity Incrd LDFs	8	28%	28%	28%	34%	9%	38%	5
Straight Average		40%	28%	26%	30%	11%	36%	

Table 7. Geometric average ratio of error of intermediate LDF value estimates from various interpolation methods vs. interpolation along the curve per small company data—geometric average of ratio of squared error of various methods to squared error of interpolation along Weibull curve (individual ratios in average capped at 2000% above and 5% below)

Curve Fit to:	Number of Curves Fit	Geometric Interpolation	Linear Interpolation	Linear % Pd or Incrd Interpolation	Geometric % Pd or Incrd Interpolation	Unadjusted Weibull
Even Maturity Paid LDFs	10	365%	419%	232%	943%	147%
Odd Maturity Paid LDFs	10	473%	658%	302%	1362%	265%
Even Maturity Incrd LDFs	7	238%	255%	336%	423%	110%
Odd Maturity Incrd LDFs	8	341%	406%	177%	1453%	201%
Straight Average		355%	429%	253%	985%	176%

$50,000, $75,000, $100,000, $150,000, $350,000, $500,000, $750,000, $1,000,000, $2,000,000, $3,000,000, $5,000,000, $7,000,000, and $10,000,000 in addition to the unity ILF at $250,000. As with the loss development factor interpolation testing, the increased limits factors were split into the odd numbered inputs and even numbered inputs. Using standard spreadsheet goal seek software, the Pareto-induced ILF that had a basic limit of unity at $250,000 and minimized the squared differences against the selected ILFs was computed for each hazard group, one using the even numbered ILFs, and another the odd ILFs. Then, as with the development factors, the accuracy of the various methods at filling in the (known) intermediate values was computed.

For reference, the percentage of the time that each method was the best estimate is shown in Table 8.

Note that interpolation along the curve is clearly preferable, in spite of the fact that the nature of the curve fit to the data excluding the basic limit forces the fitted curve to be a perfect approximation (i.e., unity).

Table 8. Winning percentage of intermediate ILF value estimates from various interpolation methods vs. interpolation along the curve

Fitted Curve	Interp. Along the Curve	Linear Interpolation	Geometric Interpolation
12%	81%	7%	0

Table 9. Geometric average ratio of squared errors relative to ILF interpolation along the curve

	Fitted Curve	Interp. Along the Curve	Linear Interpolation	Geometric Interpolation
Ratio of Sq. Errors	703%	100%	592%	740%

The geometric averages[9] of the relative squared errors of the various interpolation methods are shown in Table 9.

[9] Capped at 5% from below and 2000% from above.

Table 10. Winning percentage of intermediate LDF value estimates from cubic splines vs. interpolation along the curve using NAIC aggregate data—percentage of the tests in which cubic splines was superior to interpolation along the curve

Curve Fit to:	Number of Curves Fit	Winning % of Interp. Along the Curve	Winning % of Cubic Splines	Number of Cubic Splines Consistency Errors
Even Maturity Paid LDFs	11	64%	36%	3
Odd Maturity Paid LDFs	11	55%	45%	5
Even Maturity Incrd LDFs	10	75%	25%	1
Odd Maturity Incrd LDFs	10	73%	28%	3
Straight Average		66%	34%	

As one may see, interpolation along the curve is a superior algorithm, at least for the increased limits problem analyzed above.

6.4. Comparison to cubic splines

The previous testing presents strong reasons to prefer interpolation along the curve to other common actuarial interpolation methods. However, one might argue that numerical analysis, with its focus on interpolating polynomials, could offer a superior methodology. One could then argue that the adjustments to the fitted curves inherent in interpolation along the curve are similar to numerical analysis concepts such as linear interpolation. However, it makes sense to compare interpolation to a representative interpolation method from numerical analysis. Therefore, interpolation along the curves is compared to interpolation via cubic splines within this section. An appendix is included for readers who are not familiar with the mathematics of cubic splines.

The three data sets used previously were reused in this section. The fitting to industry loss development data produced the results in Table 10.

As one may see, interpolation along the curve is preferable twice as often as cubic spline interpolation. Further, in several instances cubic splines interpolation suggests negative development when positive development is what is actually expected. In one case it even indicates that inception-date loss at one development stage is negative rather than positive. The number of those consistency errors is shown in the table. As further support for the preference[10] for interpolation along the curve, the ratios (with capping, as before) of the squared error in cubic splines to that of interpolation along the curve are shown in Table 11.

Table 11. Geometric average ratio of error of intermediate LDF value estimates from cubic splines vs. interpolation along the curve per NAIC aggregate data—geometric average of ratio of squared error of cubic splines to squared error of interpolation along Weibull curve (individual ratios in average capped at 2000% above and 5% below)

Curve Fit to:	Number of Curves Fit	Error Ratio
Even Maturity Paid LDFs	11	188%
Odd Maturity Paid LDFs	11	133%
Even Maturity Incrd LDFs	10	271%
Odd Maturity Incrd LDFs	10	152%
Straight Average		178%

It is also relevant to evaluate the relative performance against the small company data. The winning percentage of cubic splines is shown in Table 12. The average error ratio (subject to the previous protocols) is shown in Table 13. As one may see, cubic splines performed slightly worse (relative to interpolation along the curve) on small company data than on the NAIC aggregate data.

To complete the review, cubic splines were compared to interpolation along the curve using the same

[10] It was noted that the best performance of cubic splines was in the relatively “flat” distributions associated with long-tail business. However, that is also the broad area where the negative inception-to-date loss estimate was observed. Also, cubic splines can be used when losses develop downward, whereas interpolation along a Weibull curve cannot.

Table 12. Winning percentage of intermediate LDF value estimates from cubic splines vs. interpolation along the curve using small company data—percentage of the tests in which cubic splines was superior to interpolation along the curve

Curve Fit to:	Number of Curves Fit	Winning % of Interp. Along the Curve	Winning % of Cubic Splines	Number of Cubic Splines Consistency Errors
Even Maturity Paid LDFs	10	68%	33%	4
Odd Maturity Paid LDFs	10	63%	38%	5
Even Maturity Incrd LDFs	7	93%	7%	1
Odd Maturity Incrd LDFs	8	75%	25%	3
Straight Average		74%	26%	

Table 13. Geometric average ratio of error of intermediate LDF value estimates from cubic splines vs. interpolation along the curve per small company data—geometric average of ratio of squared error of cubic splines to squared error of interpolation along Weibull curve (individual ratios in average capped at 2000% above and 5% below)

Curve Fit to:	Number of Curves Fit	Error Ratio
Even Maturity Paid LDFs	10	274%
Odd Maturity Paid LDFs	10	169%
Even Maturity Incrd LDFs	7	201%
Odd Maturity Incrd LDFs	8	142%
Straight Average		193%

NCCI increased limits data that was used in the previous testing. The winning percentage in this case was simply a single number. Specifically, interpolation along a Pareto curve produced more accurate answers in 66 (81%) of 77 cases. Further, the overall average ratio of cubic splines interpolation error to the error of interpolation along a Pareto curve across the seven hazard groups is 254%. In summary, interpolation along the curve appears to be a much more reliable method than cubic splines for interpolating actuarial data.

6.5. Summary of testing

The testing generally supports the main thesis of this paper—that interpolation along the curve is much more accurate and reliable overall than the alternative interpolation methods employed by actuaries. Further, it also suggests that, for common actuarial applications, it should be preferred to the standard cubic splines method employed by numerical analysts.

7. Summary

As the the analysis and testing in this paper show, interpolation along the curve is an enhancement to the use of curves fitted from some family of curves. Such a curve fit would often be performed anyway. Further, by adjusting the fitted curve to exactly match the observed data points, the interpolation along the curve should be more accurate[11] than curve fitting alone. So, use of this method should enhance the quality of actuarial predictions.

References

Berquist, J., and R. Sherman, "Loss Reserve Adequacy Testing: A Comprehensive Systematic Approach," *Proceedings of the Casualty Actuarial Society* 67, pp. 128–184, 1977.

Boor, J., "Estimating Tail Development Factors: What to Do When the Triangle Runs Out," Casualty Actuarial Society *Forum* (Winter 2006), pp. 345–390.

Boor, J., *An Analytical Approach to Estimating the Required Surplus, Benchmark Profit, and Optimal Reinsurance Retention for an Insurance Enterprise,* dissertation, Florida State University, 2012.

Heyer, D., "A Random Walk Model for Paid Loss Development," Casualty Actuarial Society *Forum,* Fall 2001, pp. 239–254.

Hogg, R., and S. Klugman, *Loss Distributions,* Hoboken, NJ: Wiley, 1984.

Kincaid, D., and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing,* Pacific Grove, CA: Brooks/Cole, 2002.

[11] This of course assumes that the actual data values used as input to the interpolation are reasonably accurate representations of the values of the phenomenon being analyzed. Should their accuracy be poor, the curve fitting might be preferable.

Appendix. A Brief Description of Cubic Splines

Interpolation using cubic splines is discussed elsewhere in this paper, but is not a common tactic for casualty actuaries. Therefore, this section presents the minimum that an actuary might need to know to put this paper in perspective. The interested reader is referred to a numerical analysis text such as Kincaid and Cheney (2002) for additional background on cubic splines.

The "natural" cubic spines interpolation between the values $y_1, y_2, \ldots, y_n$ at the "node" points $t_1, t_2, \ldots, t_n$. a set of 3*rd* degree polynomials, defined in each interval $[t_i, t_{i+1}]$ such that for each interval:

1. For the *ith* interval, a polynomial of degree three or less, $S_i(x)$, interpolates $y(x)$ for x values between t_i and t_{i+1};
2. Each $S_i(x)$ matches $y(x)$ perfectly at the endpoints, e.g., $S_i(t_i) = y_i$ and $S_{i+1}(t_{i+1}) = y_{i+1}$;
3. The combined splines have continuous first and second derivatives across the endpoints, e.g., $S'_i(t_{i+1}) = S'_{i+1}(t_{i+1})$ and $S''_i(t_{i+1}) = S''_{i+1}(t_{i+1})$; and,
4. The second derivative is zero at the outside endpoints, $S''_1(t_1) = 0 = S''_{n-1}(t_n)$.

The equations for the (unique) cubic splines that fulfill those conditions are

$$S_i(x) = \frac{C_i}{6l_i}(t_{i+1} - x)^3 + \frac{C_{i+1}}{6l_i}(x - t_i)^3$$

$$+\left(\frac{y_{i+1}}{l_i} - \frac{C_{i+1}l_i}{6}\right)(x - t_i) + \left(\frac{y_i}{l_i} - \frac{C_i l_i}{6}\right)(t_{i+1} - x)$$

(with each S_i applying between t_i and t_{i+1}).

The C_i's and l_i's of course must still be determined. Table 14 shows how the C_i's and l_i's may be computed from the t_i's and y_i's using a spreadsheet.

As the table shows, the standard column-by-column approach used in actuarial spreadsheets may be used to perform cubic splines interpolation. Unfortunately, the calculations underlying the α's, β's, γ's, and even those of the *C*'s are not intuitive. Further, the α's, β's, and γ's do not appear to represent any meaningful quantities that would lend themselves to column labels. Rather, they are simply intermediate calculations. As such, the calculations do not lend themselves to a spreadsheet format that can be readily understood by lay readers.

Table 14. Derivation of cubic spline constants, l_i's and C_i's, from the input t_i's and y_i's

Index *i*'s (Data)	Known Inputs t_i's (Data)	Known Values y_i's (Data)	Step Sizes l_i's $t_{i+1} - t_i$	Formula α_i's $6[y_{i+1} - y_i]/l_i$
1	$25,000	0.32	$50,000	0.0000336000
2	$75,000	0.60	$75,000	0.0000160000
3	$150,000	0.80	$200,000	0.0000090000
4	$350,000	1.10	$400,000	0.0000037500
5	$750,000	1.35	$1,250,000	0.0000010704
6	$2,000,000	1.57	$3,000,000	0.0000002540
7	$5,000,000	1.70	$5,000,000	0.0000000720
8	$10,000,000	1.76		

Index *i*'s (Data)	Starting β $2.0[l_2 + l_1]$	Starting γ $\alpha_2 - \alpha_1$	Formula β_i's $2.0[l_i + l_{i-1}] - l^2_{i-1}/\beta_{i-1}$	Formula γ_i's $\alpha_i - \alpha_{i-1} - l_{i-1}\gamma_{i-1}/\beta_{i-1}$	Method's Starting *C*'s (Data)	Final Values C_i's $[\gamma_i - l_i C_{i+1}]/\beta_i$
1					0	0.00000E+00
2	$250,000	(0.0000176000)	$250,000	(0.0000176000)		–6.98744E-11
3			$527,500	(0.0000017200)		–1.75189E-12
4			$1,124,171	(0.0000045979)		–3.97940E-12
5			$3,157,673	(0.0000010436)		–3.10861E-13
6			$8,005,174	(0.0000004033)		–4.95999E-14
7			$14,875,727	(0.0000000309)		–2.07502E-15
8					0	0.00000E+00