OSCULATORY INTERPOLATION REVISITED

H. Lock Oh and Fritz J. Scheuren, Internal Revenue Service

The smoothing of empirical sampling distributions from grouped data is a very old topic in economics and statistics. The modelling of income and wealth distributions has had a particularly long history.

Economists have typically applied global models to the sample cumulatives for income and wealth, notably of the log normal or Pareto type [1]. These global fitting procedures are attractive because they can be used in a behavioral context once the parameters are estimated. The problem with such procedures, however, is that, despite their behavioral motivation, they simply don't fit U.S. empirical data particularly well over the whole range of many income or wealth distributions.

Local fitting procedures, unlike global ones, can be made to calibrate the U.S. data exactly and in a smooth way. Osculatory interpolation is one such procedure which, as we will see, has many useful properties.

The present paper describes recent applications and extensions of osculatory interpolation methods at the Internal Revenue Service. The material is divided into five sections. First, we provide a little background concerning our interest in and use of the osculatory interpolation approach. This is followed by a formal statement of some of the problems posed by using grouped income data, as well as brief descriptions of three variations of the methodology for estimating percentiles [2]. The next section provides further extensions of the interpolation functions and the details of a new approach particularly useful for estimating cumulative totals. This is followed in the fourth section by some results. The final section makes a few concluding comments and discusses future plans.

BACKGROUND ON APPROACH

Our initial interest in the osculatory interpolation of grouped data arose about 20 years ago at the Office of Economic Opportunity (OEO), when we tried to improve on the methods then being employed by the U.S. Bureau of the Census in estimating income percentiles from tabulated data in the Current Population Survey [3].

We spent a lot of time with global fitting procedures, especially 3-parameter log normal fits of the bottom tail of the income distribution. An algorithm was created for iteratively fitting the 3-parameter log normal, using an information theoretic approach [4], but the fits weren't usable; in fact, the residuals had problems both within years and over time [5].

Later on, in the middle 70's, we were working together at the Social Security Administration on a series of problems involving mortality estimation. More specifically, we were looking at what are called estate tax multiplier wealth estimates [6]. In order to do one part of this research, we had to develop life tables for social security earners, so we started to study the tools used by demographers and came across a lot of literature on local smoothing functions. One function that we particularly liked was known as the Karup-King Osculatory Interpolation Method. This method is a form of piecewise curve-fitting that joins the pieces so that they come together smoothly (in the sense that the derivatives from the right and left are equal) [7-10].

After completing the life tables, we thought these new methods (which were programmed) might be generalized and applied to income data. This was done and the results were reported in a paper given at these ASA meetings in 1977 [2]. We were still looking at percentile estimation from which some fairly good results were obtained. They also satisfied certain bounds that had been set by Gastwirth and Glauberman at about that time, and they outperformed any of the known competitors of that era [11-12].

After that, nothing really happened on this issue for quite a while. Then, a few years ago, the "Supply Siders" conjectured that cutting tax rates would increase the amount of taxes paid by the upper income groups. By that time, we had moved to the Internal Revenue Service and so we got the job of developing a good time series on the proportion of taxes paid by the top one percent, top five percent, etc., of all taxfilers. This was a different problem from those tackled in 1976 and 1977. We needed really good estimates of totals and not just percentiles.

Since we had all the microdata, we could have simply gotten out the old files, sorted them and retabulated the already published data <u>or</u> we could have tried to extend the 1977 results to this new problem. In the end, we did <u>both</u>, a little of the first and a lot of the second. The next two sections describe the interpolation approaches we considered.

PROBLEM STATEMENT

A typical grouped income data problem consists of i=1, 2, ..., I classes each having estimated proportions

$$\hat{P}_i = \frac{\hat{N}_i}{\hat{N}}, \quad \text{where } N = \sum_{i=1}^{I} \hat{N}_i ,$$

with the N_i being the weighted number of cases in the i^{th} interval; and mean incomes $\bar{x}_i$ of all incomes falling in the income size class $[x_{i-1}, x_i)$.

The estimates required in this setting often are--

- given an income cutoff $x \varepsilon [x_{i-1}, x_i)$, find the proportion, p, or the total aggregate income, s, of the population having income less than or equal to x, or
- given a proportion of the population $p \varepsilon [p_{i-1}, p_i)$, find the income cutoff $x \varepsilon [x_{i-1}, x_i)$, the

total aggregate income, s, or the proportion of total aggregate income (Lorenz curve) attributable to this population.

Traditionally in this formulation, the desired value is interpolated based on the pattern exhibited by a sequence of ordered pairs

$$(p_i,\ x_i) \text{ or } (p_i,\ \bar{x}_i)$$

but not both. Indeed, the approach taken in the 1977 paper was of the first type.

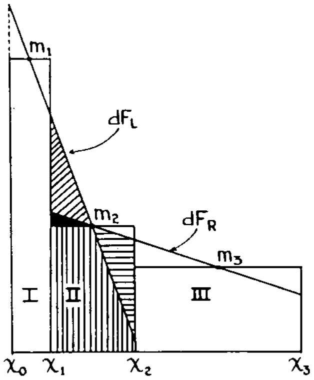
In the next section, we will see that the new procedure is an advance, in that it carries out the interpolation using sequences of ordered triplets

$$(p_i,\ x_i,\ \bar{x}_i).$$

To motivate the new technique, however, it is desirable to begin by looking at the basic Karup-King interpolation, gradually building in complexities. This is done below.

Monotonic Karup-King Osculatory Interpolation

Consider a hypothetical distribution of, say, adjusted gross income, three classes of which are shown geometrically below.



The horizontal axis provides, to scale, the four dollar income cut-offs (x_0, x_1, x_2, and x_3) which define the size classes I, II and III. The areas of the histograms which lie above the axis are drawn to be proportional to the percentages of returns in the corresponding classes. At the top of each histogram we have labelled the interval midpoints (i.e., m_1, m_2, and m_3).

We are now ready to define the Karup-King procedure for interpolating within any interval other than the initial or terminal ones. Consider size class II in the graph:

- To begin with, let us define two line segments dF_L and dF_R by connecting the points m_1, m_2, and m_3, as is done above.

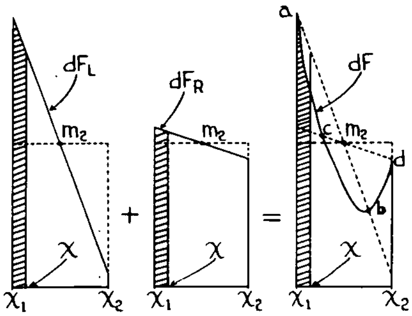
- Now, dF_L and dF_R have an interesting property; namely, that, for the middle interval, the area between each of them and the horizontal axis is the same as the rectangular area over the interval. (That is, the area under dF_L, or the cross-hatched, shaded, and vertical striped sections, equals the dimensions of interval II. Similarly, the portion under dF_R, or the shaded and the horizontal and vertical striped sections, equals the area of interval II.) Thus, since in this diagram dF_L and dF_R are always positive, we can treat them as probability density functions for the income distribution over the interval of interpolation.

- The Karup-King interpolation of the income distribution F(x) in the interval is obtained as a weighted average of the two cumulative interval distribution functions F_L and F_R. To be specific, at any point x in the interval $[x_1,\ x_2)$ it can be shown that the Karup-King distribution function is given by the expression

$$(1)\quad F\ (x) = \left(\frac{x_2 - x}{x_2 - x_1}\right) F_L\ (x) + \left(\frac{x\ - x_1}{x_2 - x_1}\right) F_R (x).$$

Since F_L (x) and F_R (x) are both quadratic functions in x, F(x) describes a cubic interpolation curve. This will always be the case in any interval other than the first or last. For the first and last intervals, where dF_L and dF_R cannot both be defined, the Karup-King interpolation curve F is simply a quadratic, since it equals either F_R (initial interval) or F_L (terminal interval) [13].

Schematically we have shown that the area under the Karup-King density function dF is related to the (appropriately weighted) areas under dF_L and dF_R as:



Several basic observations on this second chart may be worth making:

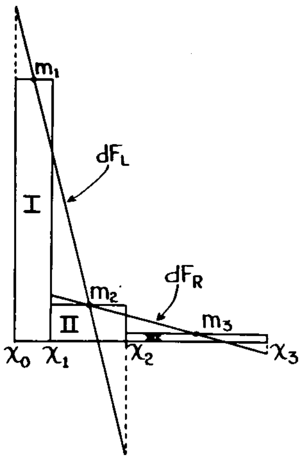
- All of the areas under dF_L and dF_R and dF are the same. This means that the original given series of data points are reproduced exactly in the interpolation.

- The shaded area under dF is less than the shaded area under line dF_L, but greater than the shaded area under dF_R. This illustrates another fact about the Karup-King interpolation curve F; namely, it always lies between F_L and F_R.

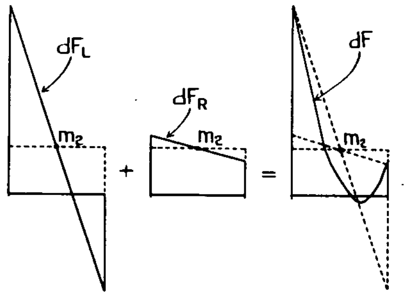
- dF intersects both dF_L and dF_R at two points each. It crosses dF_L at the beginning, or left-most point, of the interval (a), and at a point 2/3 of the way into the income class (b). Similarly, dF and dF_R coincide at a point 1/3 of the way into the class (c) and at the right-most point in the interval (d). These additional points of juncture round out the curve, giving a "smoothness" to the overall income distribution at the intersection of any two interpolation curves for adjoining intervals.

Nonmonotonic Karup-King Osculatory Interpolation

To illustrate a situation where the Karup-King procedure will not yield a monotonic curve F, let us examine a variation on the first graph.



Everything is defined the same way as before, except that dF_L can take on negative values and, therefore, is no longer a density function over the interval $[x_1, x_2)$. Furthermore, since F(x) decreases in the region where dF lies under the horizontal axis, the resulting Karup-King curve is not monotonic.



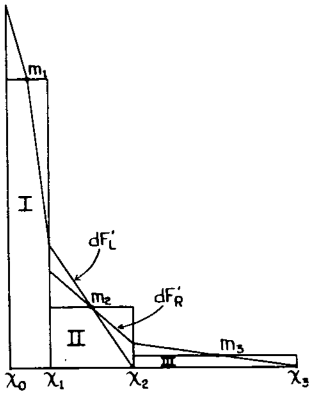
Modified Karup-King Osculatory Interpolation

Our "solution" in 1977, to cases where the Karup-King yielded a nonmonotonic distribution function, was to proceed as follows:

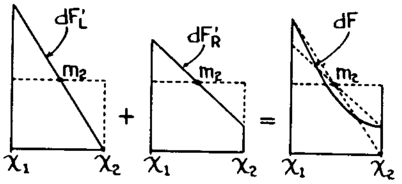
- If dF_L was negative at any point in the size classes I and II, then a modified density dF_L' was defined, which, instead of being a single straight line, consisted of two straight lines--one for each interval, such that
  (a) one line passed through m_1 and one line passed through m_2 ;
  (b) both lines lay above the horizontal axis in the interval in which they were defined; and
  (c) the two lines intersected at the juncture of the intervals in such a way that the absolute difference of the slopes was a minimum.
- If dF_R was negative at any point in classes II and III, then a modified density dF_R' was defined in a manner similar to that for dF_L'.

Only when dF_L and dF_R both lay above the horizontal axis in the intervals over which they were defined did we use the Karup-King procedure without modification.

In the particular case we examined, the modified technique yielded dF_L' and dF_R' as shown below.



A price has been paid for adopting the modified procedure. The cumulative distribution function F(x) will no longer be differentiable at the points of juncture, as before. It will, however, be continuous and strictly monotonically increasing; obviously, too, the resulting dF will be nonnegative.



EXTENSIONS AND NEW APPROACHES

The approaches we have been considering are all of the form

$$(2)\quad F(x) = (1-\alpha)F_L(x) + \alpha F_R(x)$$

$$\text{with } \alpha = \frac{x - x_{i-1}}{x_i - x_{i-1}},$$

where F_L and F_R are quadratic and adjusted to be monotonic. Moving from an interpolation function based on the ordered sequence (p_i, x_i) to one based on the sequences $(p_i, x_i, \bar{x}_i)$ means essentially that we need to impose additional constraints related to x_i. In particular, we require the interpolation function to reproduce the $(p_i, x_i, \bar{x}_i)$

for each interval $[x_{i-1}, x_i)$ in such a way that the curve generated is smoothly connected with similarly constructed curves over adjoining intervals.

Two ways of extending the 1977 work were considered. First, we looked at (positive) polynomial functions for F_L and F_R that were simply a degree higher. Second, it also turns out to be possible to use a generalized version of the Pareto to obtain F_L and F_R. The details motivating these approaches are worked out below; some results based on our data here at IRS follow in the next section.

Polynomial Fitting

For our initial choice of curves, we considered the class of (positive) polynomials for $F_L(x)$. The Lagrange interpolating polynomial is the polynomial of degree n-1 which agrees with a given function at n distinct points (or constraints). Hence, we could use a cubic function:

$$F_L(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3,\ x \varepsilon [x_{i-2}, x\ ).$$

The coefficients b_0, b_1, b_2, and b_3 are determined by solving the following simultaneous equations reflecting the four constraints:

$$p_{i-2} = b_0 + b_1\, x_{i-2} + b_2 x_{i-2}^2 + b_3 x_{i-2}^3$$

$$p_{i-1} = b_0 + b_1\, x_{i-1} + b_2 x_{i-1}^2 + b_3 x_{i-1}^3$$

$$p_i = b_0 + b_1\, x_i + b_2 x_i^2 + b_3 x_i^3$$

$$p_i \bar{x}_i = \int_{x_{i-1}}^{x_i} t d\, F_L(t)$$

$$= \tfrac{1}{2} b_1 (x_i^2 - x_{i-1}^2) + \tfrac{2}{3}\, b_2 (x_i^3 - x_{i-1}^3) +$$

$$\tfrac{3}{4}\, b_3 (x_i^4 - x_{i-1}^4).$$

Similarly, we could construct F_R and, indeed, finally obtain an interpolation formula for F(x) by combining F_L and F_R using (2) to yield

$$(3)\ F(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4,$$
$$x \varepsilon (x_{i-1}, x_i).$$

Interpolation methods employing the class of polynomials given in expression 3 above can, again, run into problems of monotonicity. Let

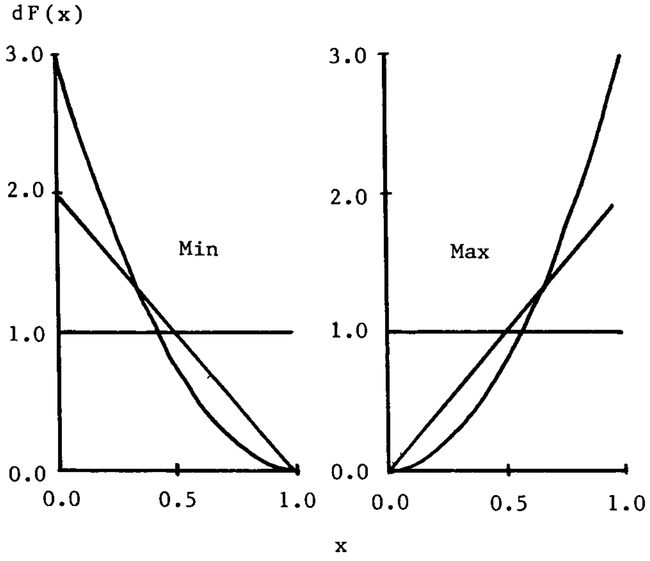
$$\bar{x}_i = x_{i-1} + k(x_i - x_{i-1}).$$

In order for $dF(x) \geq 0$ in the interval $[x_{i-1}, x_i)$, then k must lie within certain limits, i.e.,

$$k_{min} \leq k \leq k_{max},$$

which depend on the interpolation function being fit. For example, it can be shown that for $dF(x) \geq 0$ for all $x \varepsilon [x_{i-1}, x_i)$, the minimum k under uniform, linear, and quadratic density functions is 1/2, 1/3, 1/4, respectively, and the maximum k is 1/2, 2/3, 3/4, respectively. (See Figure A.)

Figure A.--Minimum and Maximum k Under Uniform, Linear and Quadratic Densities



In our AGI data for 1984 (see Figure B), k lies in a fairly narrow range before declining sharply from 0.5 after the 95th percentile or so; this seems to indicate that the new method may yield negative values for the density in some part of the interpolation interval, resulting--unless adjusted--in a decreasing cumulative distribution function.

Figure B.--1985 Individual Returns

AGI Size Class	Cumulative Percent	Percent in Interval	Class Mean (in $)	k Value
Under $1,000	2.2	2.3	574	.57
$1,000 under $2,000	5.5	3.3	1,500	.50
$2,000 under $3,000	9.0	3.4	2,491	.49
$3,000 under $4,000	12.3	3.3	3,500	.50
$4,000 under $5,000	15.6	3.3	4,503	.50
$5,000 under $6,000	18.9	3.3	5,493	.49
$6,000 under $7,000	22.1	3.3	6,491	.49
$7,000 under $8,000	25.3	3.2	7,502	.50
$8,000 under $9,000	28.8	3.4	8,508	.51
⋮	⋮	⋮	⋮	⋮
$40,000 under $50,000	92.0	6.6	44,455	.45
$50,000 under $75,000	97.5	5.6	59,288	.37
$75,000 under $100,000	98.8	1.3	85,028	.40
$100,000 under $200,000	99.7	0.9	131,082	.31
$200,000 under $500,000	99.9	0.2	289,751	.30
$500,000 under $1,000,000 ..	100.0	0.0	669,994	.34

Pareto Fitting

As an alternative to the (positive) polynomial, we investigated the Pareto distribution

$$F(x) = \frac{m}{n} \frac{1}{x^n},\ \text{with}\ 1 < n < m < x$$

which has often been used in fitting the upper tail of income and wealth distributions.

The Pareto seems intuitively to be a suitable density for cases where the polynomial function fails because of small k. In fact, it can be shown that there is always a positive Pareto dF(x) no matter how small k becomes. To illustrate this, Figure C shows values of the Pareto dispersion parameter n by selected values of k, given the interpolation interval as a multiple of the lower class limit; Figure D presents values of k for each selected n value.

Figure C.--Dispersion Parameter n of Pareto for Selected Values of k and Interpolation Interval

k	Interpolation Interval as Multiple of Lower Class Limit				
	0.5	1	1.5	2	5
0.25	7.49	3.86	2.63	2.00	0.82
0.20	10.16	5.32	3.68	2.85	1.29
0.10	20.96	10.94	7.59	5.91	2.82
0.05	41.00	21.00	14.33	11.00	4.99

Figure D.--Interval Proportion, k, for Selected Values of Dispersion Parameter n of Pareto and Interpolation Interval

n	Interpolation Interval as Multiple of Lower Class Limit				
	0.5	1	1.5	2	5
1.25	0.425	0.373	0.334	0.304	0.204
1.50	0.416	0.359	0.317	0.285	0.181
1.75	0.408	0.346	0.301	0.267	0.161
2.00	0.400	0.333	0.286	0.250	0.143
2.50	0.384	0.309	0.257	0.219	0.114
3.50	0.353	0.264	0.208	0.169	0.077

Since our method requires four constraints, the Pareto cannot be used as such, but a polynomial with negative exponents will behave almost just like a Pareto. We selected n=1.5, 2.5, 3.5 from the chart and solved the Lagrange equations, as above, this time obtaining for F_L the form

$$(4)\quad F_L(x) = b_0 + b_1x^{-1.5} + b_2x^{-2.5} + b_3x^{-3.5}$$

Again, using the procedure outlined earlier for the (positive) polynomial fit, a similar expression to expression 4 can be obtained for F_R; and, finally, combining the Pareto versions of F_L and F_R, we can derive a Pareto version of expression 2.

RESULTS AND IMPLICATIONS

Tables 1 and 2 provide some results on the three interpolation methods that we have discussed--the initial 1977 approach (Method I), the (positive) polynomial generalization (Method II) and the Pareto or negative polynomial generalization (Method III).

Percentile Estimates (Table 1)

For the dollar cutoffs or percentile estimates Method I, as expected, behaves quite well over the bulk of the income distribution, becoming, however, unreliable in the extreme upper tails. Method II behaves generally better than Method I, although it too becomes unreliable in the extreme upper tail of the adjusted gross income (AGI) distribution. Method III does very well, as expected, in the upper tail and fits about as well as Method II in the lower part of the distribution.

Aggregate Estimates (Table 2)

For the aggregate estimates within each AGI income interval, Method I does fairly poorly. It consistently, in these data, overestimates the aggregates nearly everywhere. Method II does extremely well except in the upper tail, where, as with the percentile data, its performance continues to be unreliable. Method III achieves really phenomenal results with these data, justifying the heavy reliance we have placed on it in looking at the Supply Sider's questions. (Our data, by the way, do show the predicted effects, although not necessarily in the magnitude expected. (See Figure E below.)

Figure E.--Federal Income Tax Payments

Tax Year	Taxes Paid		Percent of Total Paid	
	Rich	Poor	Rich	Poor
1981	$51.0	$21.0	18.05%	7.45%
1982	53.6	20.3	19.41	7.35
1983	54.1	19.5	19.93	7.17
1984	62.7	21.9	21.10	7.35
1985	72.1	23.1	22.13	7.10

Income taxes after credits; in billions of dollars. 1985 data are preliminary.

Top 1% of taxpayers, measured by adjusted gross income.

Lowest 50% of taxpayers, measured by adjusted gross income.

Source: Internal Revenue Service
(Reprinted here from [14].)

CONCLUSIONS AND AREAS FOR FUTURE STUDY

Four general remarks might be made about the direction that this work has taken so far. (In particular, some comments are in order about where we have been and where we should be going.)

At this point we are reasonably happy with the improvements that have been made over the 1977 results. The new polynomial fits, especially those based on the Pareto, calibrate the data very nicely. We have more work to do on the problem, however. For one thing, there is a growing literature in this area, including a paper given at these meetings [15-18]. Undoubtedly the results of others may be worth programming and testing to see if there are further improvements worth making.

TABLE 1.--INTERPOLATED VALUE OF ADJUSTED GROSS INCOME (AGI) SIZE FOR SOI-85 INDIVIDUAL TAX RETURNS BY AGI SIZE CLASS: DELETING ONE SIZE CLASS (AS PERCENT OF PUBLISHED VALUES)

SIZE OF AGI	NUMBER OF RETURNS		PRO-PORTION CLASS MEAN AGI	DOLLAR CUTOFF				AGI AMOUNT (IN MILLION DOLLARS)			
	(IN PERCENT)	CUMULATIVE (IN PERCENT)		PUB-LISHED ACTUAL	INTERPOLATED			PUB-LISHED ACTUAL	INTERPOLATED		
					I	II	III		I	II	III
2,000 - 3,000......	3.433	8.973	.49121	100.00	110.44	99.93	100.10	100.00	102.04	100.19	100.05
3,000 - 4,000......	3.302	12.275	.49559	100.00	106.94	100.01	100.03	100.00	114.15	100.03	100.10
4,000 - 5,000......	3.342	15.616	.50276	100.00	105.89	99.98	100.00	100.00	118.66	99.90	99.88
5,000 - 6,000......	3.280	18.896	.49264	100.00	104.70	99.81	99.82	100.00	121.25	99.99	99.97
6,000 - 7,000......	3.250	22.147	.49088	100.00	104.35	99.96	99.98	100.00	122.69	100.06	100.02
7,000 - 8,000......	3.182	25.328	.50207	100.00	103.03	99.97	99.96	100.00	124.02	100.05	100.10
8,000 - 9,000......	3.441	28.769	.50767	100.00	102.93	100.04	100.03	100.00	123.97	99.52	99.91
9,000 - 10,000.....	3.236	32.005	.48625	100.00	102.79	99.90	99.91	100.00	124.51	100.00	99.99
10,000 - 11,000....	2.960	34.966	.48602	100.00	102.35	99.86	99.86	100.00	125.23	100.00	100.00
11,000 - 12,000....	2.847	37.813	.48713	100.00	102.47	99.92	99.93	100.00	125.43	100.02	100.01
12,000 - 13,000....	2.701	40.514	.49208	100.00	102.19	99.95	99.95	100.00	126.17	100.02	100.02
13,000 - 14,000....	2.690	43.204	.49799	100.00	101.92	99.91	99.91	100.00	126.18	99.95	99.95
14,000 - 15,000....	2.664	45.867	.48360	100.00	101.65	99.97	99.96	100.00	126.26	100.08	100.07
15,000 - 16,000....	2.545	48.413	.50230	100.00	101.92	100.10	100.10	100.00	125.93	100.00	99.99
16,000 - 17,000....	2.320	50.732	.50289	100.00	101.44	100.02	100.02	100.00	127.07	100.00	100.00
17,000 - 18,000....	2.329	53.061	.50325	100.00	101.65	100.01	100.01	100.00	126.46	99.96	99.95
18,000 - 19,000....	2.185	55.247	.48783	100.00	101.28	99.93	99.92	100.00	127.11	100.01	100.01
19,000 - 20,000....	2.151	57.397	.49050	100.00	101.42	99.99	99.99	100.00	126.62	100.01	100.01
20,000 - 21,000....	1.988	59.385	.49348	100.00	101.20	99.91	99.91	100.00	126.99	99.96	99.95
21,000 - 22,000....	1.913	61.298	.47575	100.00	101.08	99.97	99.97	100.00	126.96	100.08	100.08
22,000 - 23,000....	1.788	63.086	.50649	100.00	101.38	100.06	100.06	100.00	126.44	99.96	99.95
23,000 - 24,000....	1.600	64.686	.49269	100.00	101.11	99.95	99.95	100.00	127.92	100.00	100.00
24,000 - 25,000....	1.632	66.318	.49604	100.00	101.08	100.02	100.02	100.00	127.65	100.03	100.03
25,000 - 26,000....	1.602	67.919	.50873	100.00	101.13	100.03	100.03	100.00	127.29	99.57	99.97
26,000 - 27,000....	1.523	69.443	.49250	100.00	101.08	100.00	100.00	100.00	127.90	100.02	100.02
27,000 - 28,000....	1.507	70.949	.50559	100.00	100.71	99.97	99.97	100.00	127.91	99.97	99.97
28,000 - 29,000....	1.505	72.454	.48230	100.00	101.26	100.02	100.02	100.00	125.28	100.03	100.02
29,000 - 30,000....	1.275	73.729	.49770	100.00	100.41	99.92	99.93	100.00	126.47	99.96	99.97
30,000 - 32,000....	2.760	76.489	.50213	100.00	101.90	100.07	100.06	100.00	125.63	99.95	99.98
32,000 - 34,000....	2.446	78.935	.49140	100.00	101.25	99.96	99.96	100.00	126.75	100.01	100.01
34,000 - 36,000....	2.361	81.296	.49466	100.00	101.45	99.98	99.98	100.00	125.80	99.96	99.95
36,000 - 38,000....	2.087	83.383	.47785	100.00	100.96	99.96	99.96	100.00	124.56	100.05	100.05
38,000 - 40,000....	1.909	85.293	.45733	100.00	100.62	100.04	100.04	100.00	123.26	99.99	99.99
40,000 - 45,000....	3.911	89.204	.47720	100.00	100.02	100.12	100.12	100.00	116.51	100.02	100.01
45,000 - 50,000....	2.749	91.952	.47954	100.00	101.25	99.82	100.07	100.00	112.36	99.84	99.94
50,000 - 75,000....	5.594	97.546	.37152	100.00	106.30	98.70	100.29	100.00	106.48	99.96	100.02
75,000 - 100,000...	1.256	98.802	.40110	100.00	119.86	86.10	99.24	100.00	107.20	94.25	99.75
100,000 - 200,000..	.904	99.705	.31082	100.00	160.63	63.09	98.60	100.00	106.52	85.67	99.41

Note: The interpolated value was obtained by using the published cumulative AGI size distribution excluding the size class that is to be estimated.

TABLE 2.--INTERPOLATED VALUE OF ADJUSTED GROSS INCOME (AGI) SIZE FOR SOI-85 INDIVIDUAL TAX RETURNS BY AGI SIZE CLASS: DELETING TWO SIZE CLASSES (AS PERCENT OF PUBLISHED VALUES)

SIZE OF AGI	NUMBER OF RETURNS		PRO-PORTION CLASS MEAN AGI	DOLLAR CUTOFF				AGI AMOUNT (IN MILLION DOLLARS)			
	(IN PERCENT)	CUMULATIVE (IN PERCENT)		PUBLISHED ACTUAL	INTERPOLATED			PUBLISHED ACTUAL	INTERPOLATED		
					I	II	III		I	II	III
2,000 - 3,000.......	3.433	8.973	.49121	100.00	111.48	100.33	100.78	100.00	89.41	100.56	100.83
3,000 - 4,000.......	3.302	12.275	.49559	100.00	109.47	99.72	100.01	100.00	121.66	99.88	99.58
4,000 - 5,000.......	3.342	15.616	.50276	100.00	106.35	99.99	100.02	100.00	107.52	99.92	99.92
5,000 - 6,000.......	3.260	18.896	.49264	100.00	106.29	99.88	99.94	100.00	123.79	100.00	99.97
6,000 - 7,000.......	3.250	22.147	.49088	100.00	104.23	100.19	100.27	100.00	115.22	100.27	100.36
7,000 - 8,000.......	3.182	25.328	.50207	100.00	104.28	99.73	99.67	100.00	122.00	100.02	100.00
8,000 - 9,000.......	3.441	28.769	.50767	100.00	103.63	100.06	99.99	100.00	120.07	99.93	99.85
9,000 - 10,000......	3.236	32.005	.48625	100.00	103.81	100.01	100.05	100.00	114.43	100.01	100.03
10,000 - 11,000.....	2.960	34.966	.48602	100.00	102.55	99.85	99.86	100.00	122.37	100.00	99.99
11,000 - 12,000.....	2.847	37.813	.48713	100.00	103.21	99.97	99.99	100.00	113.72	99.57	99.95
12,000 - 13,000.....	2.701	40.514	.49208	100.00	102.02	99.90	99.91	100.00	123.22	100.00	100.00
13,000 - 14,000.....	2.690	43.204	.49799	100.00	102.56	99.93	99.92	100.00	114.47	99.93	99.93
14,000 - 15,000.....	2.664	45.867	.48360	100.00	102.11	99.98	99.96	100.00	125.23	100.07	100.06
15,000 - 16,000.....	2.545	48.413	.50230	100.00	102.50	100.07	100.09	100.00	110.03	99.94	99.94
16,000 - 17,000.....	2.320	50.732	.50289	100.00	101.46	99.92	99.92	100.00	125.58	99.91	99.92
17,000 - 18,000.....	2.329	53.061	.50325	100.00	102.10	100.03	100.08	100.00	111.32	99.98	99.98
18,000 - 19,000.....	2.185	55.247	.48783	100.00	101.32	99.91	99.91	100.00	126.60	100.00	99.99
19,000 - 20,000.....	2.151	57.397	.49050	100.00	101.85	100.02	100.02	100.00	109.70	99.98	99.98
20,000 - 21,000.....	1.988	59.385	.49348	100.00	100.97	99.96	99.96	100.00	127.24	99.95	99.99
21,000 - 22,000.....	1.913	61.298	.47575	100.00	101.47	99.99	99.98	100.00	106.90	100.07	100.08
22,000 - 23,000.....	1.788	63.086	.50649	100.00	101.40	100.09	100.09	100.00	127.92	99.99	99.99
23,000 - 24,000.....	1.600	64.686	.49269	100.00	101.55	99.94	99.94	100.00	108.74	100.03	100.03
24,000 - 25,000.....	1.632	66.318	.49604	100.00	101.28	100.00	100.00	100.00	127.62	100.01	100.01
25,000 - 26,000.....	1.602	67.919	.50873	100.00	101.49	100.03	100.03	100.00	109.09	99.96	99.96
26,000 - 27,000.....	1.523	69.443	.49250	100.00	100.68	99.98	99.99	100.00	126.04	100.02	100.03
27,000 - 28,000.....	1.507	70.949	.50559	100.00	101.01	99.94	99.93	100.00	110.34	99.97	99.97
28,000 - 29,000.....	1.505	72.454	.48230	100.00	100.87	100.18	100.18	100.00	122.68	100.14	100.15
29,000 - 30,000.....	1.275	73.729	.49770	100.00	101.41	99.89	99.89	100.00	131.13	100.05	100.05
30,000 - 32,000.....	2.760	76.489	.50213	100.00	101.87	100.13	100.16	100.00	126.21	100.08	100.07
32,000 - 34,000.....	2.446	78.935	.49140	100.00	101.93	99.89	99.89	100.00	108.97	100.01	100.01
34,000 - 36,000.....	2.361	81.296	.49466	100.00	101.12	100.05	100.04	100.00	124.52	100.02	100.01
36,000 - 38,000.....	2.087	83.383	.47785	100.00	101.48	99.94	99.95	100.00	107.36	100.05	100.05
38,000 - 40,000.....	1.909	85.293	.49733	100.00	100.04	100.05	100.02	100.00	110.20	99.99	99.98
40,000 - 45,000.....	3.911	89.204	.47720	100.00	100.98	100.12	100.14	100.00	115.35	100.04	100.02
45,000 - 50,000.....	2.749	91.952	.47954	100.00	107.22	99.36	100.15	100.00	104.55	99.66	99.97
50,000 - 75,000.....	5.594	97.546	.37152	100.00	108.41	97.69	100.41	100.00	105.24	99.98	100.03
75,000 - 100,000....	1.256	98.802	.40110	100.00	207.39	81.97	97.17	100.00	107.90	92.24	98.97
100,000 - 200,000...	.904	99.705	.31082	100.00	184.68	43.30	98.52	100.00	102.69	64.73	97.36

Note: The two adjoining interpolated values were obtained by using the published cumulative AGI size distribution excluding two size classes that are to be estimated.

Our initial attack of over 20 years ago, in which we attempted to fit parametric forms globally, was a failure; but, as we have seen, a good guess on the parametric form can be quite helpful when fitting wide intervals and sparse data. Clearly, for example, the Pareto did very well in the upper tail of the AGI distribution. This brings us full circle from global parametric approaches, which don't work, to local parametric approaches which do. It suggests we look very hard at our data to see if other parametric forms might work (even better than the Pareto). Certainly the suggestions in Hoaglin et al. [19] for heavy-tailed distributions are worthy of study with tax data.

While improvements in the basic constraint equations can be made and are being worked on, methods for dealing with open-ended classes really cannot be effectively approached using the methods being developed unless strong distributional assumptions are made. We have already looked at this problem and will talk briefly in a later paper at these meetings [20] about a James-Stein approach to smoothing the open-ended interval.

One final point, the discussant, Bob Fay, suggested that the sensitivity of the interpolation to sampling error be investigated. While we do not feel this is that important in a number of our applications because of the stratified samples we use, nonetheless, the challenge is appropriate and deserves study and we thank him for it as well as his other helpful comments.

ACKNOWLEDGMENTS

In addition to Bob Fay, the authors would like to thank Beth Kilss and Wendy Alvey for their editorial assistance in the preparation of this paper. Special thanks also go to Dorothy Farmer and Nancy Dutton for their typing.

NOTES AND REFERENCES

[1] Aitchinson, J. and Brown, J.A.C. (1957), The Lognormal Distribution with Special Reference to Its Uses in Economics, Cambridge University Press. See, also, Aigner, D.J., and Goldbeyer, A.S. (1970) "Estimation of Pareto's Law from Grouped Data," Journal of the American Statistical Association, vol. 65, no. 330, pp. 712-723.

[2] Oh, H. L. (1977), "Osculatory Interpolation with a Monotonicity Constraint," 1977 American Statistical Association Proceedings, Section on Statistical Computing.

[3] A discussion of Census methods is found in Spiers, E. (1977) "Estimation of Summary Measures of Income Size Distribution From Grouped Data," 1977 American Statistical Association Proceedings, Social Statistics Section, pp. 252-257. Incidentally, the Census Bureau in its CPS income series has been using ungrouped data for quite some time in making percentile estimates. See U.S. Bureau of the Census (1986), Money Income of Households, Families and Persons in the United States: 1985, Series P-60, no. 156, Washington, DC.

One final point, as noted by the discussant: the new procedure advocated here, which is based on means, may not be appropriate for the CPS because of data problems such as rounding. See for example, Scheuren, F. (1980), "Appendix II: Placement of Survey Wage Class Intervals in the Presence of Rounding Error," Studies from Interagency Data Linkages, Report No. 3, U.S. Department of Health and Human Services, Social Security Administration, pp. 241-275. The original 1977 Karup-King approach may have merit in a CPS context, however, especially in long intervals away from the median.

[4] The approach we took was a special case of what would eventually be called the EM algorithm. See Dempster, A.P.; Laird, W.; and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," Journal of Royal Statistical Society, Ser. B, 39, pp. 1-38.

[5] The practical problem of looking at changes in income and poverty over time, which started us on this quest, was partially "solved" in Scheuren, F. (1973), "Ransacking CPS Tabulations: Applications of the Log Linear Model to Poverty Statistics," Annals of Economic and Social Measurement, 2/2, pp. 159-182. (The demise of OEO ended our remaining interest, at least at that point.)

[6] Oh, H.L. and Scheuren, F. (1976), "Some Preliminary Results from a Validation Study of the Estate Multiplier Procedure," 1976 American Statistical Association Proceedings, Social Statistics Section, pp. 650-654. Also, see Scheuren, F. and Oh, H.L. (1976), "Some Preliminary Synthetic SSA Cohort Survival Rates, 1950-1971," Unpublished SSA working paper, U.S. Social Security Administration.

[7] Greville, T.N.E. (1944), "The General Theory of Osculatory Interpolation," Transactions of the Actuarial Society of America, 45 (112): 202-265, Part II.

[8] Shryock, H.S., and Siegel, J.S., et al. (1971), The Methods and Materials of Demography, vol. 2, pp. 681-701, U.S. Bureau of the Census, Washington, DC.

[9] Akima, H. (1970), "A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures," Journal of the Association for Computing Machinery, vol. 17, no. 4, pp. 589-602.

[10] Freeman, H. (1965), Finite Differences for Actuarial Students, University Press, Cambridge.

[11] Gastwirth, J. and Glauberman, M. (1976), "The Interpolation of the Lorenz Curve and Gini Index From Grouped Data," Econometrica, vol. 44, pp. 479-483.

[12] Gastwirth, J. (1972), "On the Estimation of the Lorenz Curve and Gini Index," Technical Report No. 147, Department of Statistics, Johns Hopkins University, Baltimore.

[13] This statement assumes, among other things, that we can treat the income distribution as having all of its mass lying on an interval of finite length. In some cases, this could not be justified; hence, an

extrapolation procedure (such as fitting a Pareto to the upper class), rather than an interpolation method, would be required.

[14] Vedder, R. and Frenze, C. (1987) "Latest Data Attest Cuts in Top Rate Play Robin Hood," Wall Street Journal, p. 34.

[15] Gastwirth, J. (1985) L., Comment on "Measurement of Economic Distance Between Blacks and Whites" by H.D. Vinod, Journal of Business and Economic Statistics, vol. 3, no. 4, pp. 405-407.

[16] Krieger, A.M. (1983), "Bounding Moments From Grouped Data and the Importance of Group Means," Sankyha, Ser. B, 45, pp. 309-319.

[17] Krieger, A.M., and Gastwirth, J.L. (1984), "Interpolation From Grouped Data for Unimodel Densities," Econometrica, 52, pp. 419-426.

[18] Cooil, Bruce (1987), "On Empirical Procedures for Estimating the Tail of a Quantile Function," presented at the 1987 Annual Meeting of the American Statistical Association.

[19] Hoaglin, D.C. (1983), "Letter Values: A Set of Selected Order Statistics," in Understanding Robust and Exploratory Data Analysis, eds. D.C. Hoaglin, F. Mosteller, and J.S. Tukey, New York: John Wiley, pp. 33-57. Also see Hoaglin, D.C. (1985), "Summarizing Shape Numerically: the g-and-h Distributions," in Exploring Data Tables, Trends, and Shapes, eds. D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: John Wiley, pp. 461-513.

[20] Scheuren, F. and McCubbin, J. (1987), "Piecing Together Personal Wealth Distributions," Statistics of Income and Related Administrative Record Research: 1986-1987, Internal Revenue Service, 1987 (in this volume).