**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 or we could have tried to extend the 1977 results to this new problem. In the end, we did both, 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 \(i=1, 2, \ldots, I\) classes each having estimated proportions

\[
P_i = \hat{P}_i, \quad \text{where} \quad N = \sum_{i=1}^{I} N_i,
\]

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 \(x \in [x_{i-1}, x_i)\), 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 \(p \in (P_{i-1}, P_i)\),

find the income cutoff \(x \in [x_{i-1}, x_i)\), 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

\[(p_1, x_1) \text{ or } (p_2, x_2)\]

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

\[(p_1, x_1, x_2)\]

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.

![Diagram of income distribution classes](image)

The horizontal axis provides, to scale, the four dollar income cut-offs \((x_0, x_1, x_2, \text{ and } x_3)\) 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

\[
F(x) = \left(\frac{x-x_1}{x_2-x_1}\right)F_L(x) + \left(\frac{x_2-x}{x_2-x_1}\right)F_R(x).
\]

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 \(F_L\) (initial interval) or \(F_R\) (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:

\[
\begin{align*}
\text{Area under } dF_L & = \text{Area under } dF_R \\
\text{Area under } dF & = \frac{1}{2} \text{Area under } dF_L + \frac{1}{2} \text{Area under } dF_R
\end{align*}
\]

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.

![Graph](image)

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.

![Graph](image)

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

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

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:

\[
P_L(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3, \quad x \in [x_{i-1}, x_i].
\]

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

\[
\begin{align*}
p_{i-2} &= b_0 + b_1 x_{i-2} + b_2 x_{i-2}^2 + b_3 x_{i-2}^3, \\
p_{i-1} &= b_0 + b_1 x_{i-1} + b_2 x_{i-1}^2 + b_3 x_{i-1}^3, \\
p_i &= b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3, \\
p_{i+1} &= \int_{x_{i-1}}^{x_i} n(t) dt
\end{align*}
\]

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.

### Pareto Fitting

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

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

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.

### Table: Interpolation Interval as Multiple of Lower Class Limit

<table>
<thead>
<tr>
<th>( k )</th>
<th>0.5</th>
<th>1.5</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>7.49</td>
<td>3.86</td>
<td>2.63</td>
<td>2.00</td>
</tr>
<tr>
<td>0.20</td>
<td>10.16</td>
<td>5.32</td>
<td>3.68</td>
<td>2.85</td>
</tr>
<tr>
<td>0.10</td>
<td>20.96</td>
<td>10.94</td>
<td>7.59</td>
<td>5.91</td>
</tr>
<tr>
<td>0.05</td>
<td>41.00</td>
<td>21.00</td>
<td>14.33</td>
<td>11.00</td>
</tr>
</tbody>
</table>

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

\[
F_L(x) = b_0 + b_1x^{-1.5} + b_2x^{-2.5} + b_3x^{-3.5}.
\]

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.

### 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


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. (1987), "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.


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, 272, pp. 159-182. (The demise of OEO ended our remaining interest, at least at that point.)


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.


