cirlin 2

1963 CENSUS ESTABLISHMENT SIZE DISTRIBUTION DATA

ABSTRACT: This data set contains the 1963 Census statistics by establishment employment size classes for the 2-digit and 4-digit Census manufacturing industries. Data is also included for the total of all industries in the 1-digit SIC coded "1". Some of the data in some of the size classes is aggregated with other size classes in order to protect confidential firm information.

JCI:
```
//GO.FT08F001 DD DSN=ESTSIZ63,DISP=OLD,UNIT=TAPE6250,
// DCB=(RECFM=FB,LRECL=80,BLKSIZE=3520),
// LABEL=(02,SL,,IN),VOL=SER=DATABK DTA063
                  38
```

DATA SET ORGANIZATION: Each record is uniquely associated with a particular size class in a particular industry. Records are sorted according to the following major to minor sequence:

1. Census SIC
2. Size class code

The sort is in ascending sequence, with the following qualifications:

1. The 2-digit codes precede the 4-digit codes. The data set begins with an industry with SIC=1 that represents all manufacturing industries.
2. SIC 19 follows all the 2-digit codes; all SIC's in the 1900's follow the 4-digit codes.

For each industry, there are 11 records--one record for each of 10 size classes, plus an additional record for the industry totals. The size class codes are as follows:

Size class	Code
Industry totals	0
1-4	1
5-9	2
10-19	3
20-49	4
50-99	5
100-249	6
250-499	7
500-999	8
1,000-2,499	9
2,500 and over	10

 # OF RECORDS: 11 records per industry, 446 industries, and hence 4906 total records

FORMAT: (I4,2X,I2,1X,I1,6F10.0)

VARIABLES:

1. Census industrial code (SIC)
2. Size class code 10 possible
3. Aggregation code (see note 1 below)
4. Number of establishments
5. Number of employees
6. Payroll ($ thousands)
7. Value added " "
8. Value of shipments " "
9. Capital expenditures " "

SOURCE: U.S. Bureau of the Census. Census of Manufactures: 1963, Vol. 1, Summary and Subject Statistics, Table 3, "Selected Statistics by Employment Size of Establishments: 1963". U.S. Government Printing Office: Washington, D.C., 196x.

NOTES: 1. Aggregation--As noted earlier, some size classes are aggregated with others to prevent disclosure of confidential firm information. In the Census report, such aggregated data is recorded in the column for the smallest size class of the group and is indicated by underscoring. The subsequent size classes in the grouping are coded by the letter "D", and are left blank in this data set. Our corresponding aggregation codes are:

1. All values of the variables 5 through 9 of this size class are aggregates of the data of this and one or more subsequent size classes. (Indicated in the original source by a column of five underscored figures.)

3. The data for variables 5 through 9 of this size class is included i in a previous aggregation. (Indicated in the original source by a column of five D's.)

2,4. Analogous to codes 1 and 3 for variable 9 (capital expenditures) only. (Indicated in the report by an underscoring and a subsequent D for capital expenditures.)

5. All values in variables 5 through

9 of this size class are aggregates of the data of this and one or more subseqent size classes. In addition, the data for variable 9 (capital expenditures) is included in a previous aggregation. (Indicated in the report by a column of four underscored figures, and a D for capital expenditures.)

If both codes 3 and 4 are valid, code 3 is recorded.

2. For some size classes there are no data observations. Variables 3 through 9 in the corresponding records contain zeroes. Missing values are recorded as zeroes, as are suppressed variables.

3. The original size distribution data was summed and checked against the industry totals. A number of errors were found. The Bureau of the Census was able to correct most of these errors including all the major ones.

4. a version of this data set has been prepared that provides estimates of the value of the non-disclosed variables. A description of the exact disaggregation procedure is available in another document. The JCL for the revised data set is given below:

```
//GO.FT08F001 DD DSN=SIZ63MCD,DISP=OLD,UNIT=TAPE6250,
// DCB=(RECFM=FB,LRECL=80,BLKSIZE=7280),
// LABEL=(03,SL,,IN),VOL=SER=DATABK
```

5A copy of this data set is kept on cards at the RPCBP, at GSM, UCLA.

cirlin 3

1963 CENSUS ESTABLISHMENT SIZE DISTRIBUTION DATA, DISAGGREGATED

ABSTRACT: This data set contains the 1963 Census statistics by establishment employment size classes for the 2-digit and 4-digit Census manufacturing industries, but differs from the original data in one important respect. In the original data (available in another data set), the data is often aggregated to avoid revealing confidential firm information. We have developed a technique for disaggregating the aggregated data--the technique is described in separate documentation--and this data set contains our estimates of the true values of the aggregated variables.

JCL:
```
//GO.FT08F001 DD DSN=SIZ63MCD,DISP=OLD,UNIT=TAPE6250,
// DCB=(RECFM=FB,LRECL=80,BLKSIZE=7280),
// LABEL=(~~03~~39,SL,,IN),VOL=SER=~~DATABK~~DTA063
```

DATA SET ORGANIZATION: Each record is uniquely associated with a particular size class in a particular industry. Records are sorted according to the following major to minor sequence:

1. Census SIC
2. Size class code

The sort is in ascending sequence, with the following qualifications:

1. The 2-digit codes precede the 4-digit codes. The data set begins with an industry with SIC=1 that represents all manufacturing industries.
2. SIC 19 follows all the 2-digit codes; all SIC's in the 1900's follow the 4-digit codes.

For each industry, there are 11 records--one record for each of 10 size classes, plus an additional record for the industry totals. The size class codes are as follows:

Size class	Code
Industry totals	0
1-4	1
5-9	2
10-19	3
20-49	4
50-99	5
100-249	6
250-499	7
500-999	8
1,000-2,499	9

2,500 and over 10

 # OF RECORDS: 11 records per industry, 446 industries, and hence 4906 total records

FORMAT: (I4,2X,I2,1X,I1,6F10.0)

VARIABLES:

1. Census industrial code (SIC)
2. Size class code
3. Aggregation code (see note 1 below)
4. Number of establishments
5. Number of employees
6. Payroll ($ thousands)
7. Value added " "
8. Value of shipments " "
9. Capital expenditures " "

SOURCE: U.S. Bureau of the Census. Census of Manufactures: 1963, Vol. 1, Summary and Subject Statistics, Table 3, "Selected Statistics by Employment Size of Establishments: 1963". U.S. Government Printing Office: Washington, D.C., 196x. This data is available on magnetic tape; the JCL of the data set containing the raw data is:

```
//GO.FT08f001 DD DSN=ESTSIZ63,DISP=OLD,UNIT=TAPE6250,
// DCB=(RECFM=FB,LRECL=80,BLKSIZE=3520),
// LABEL=(~~02~~ 38,SL,,IN),VOL=SER=~~DATABK~~ DTA063
```

NOTES: 1. Aggregation--As noted earlier, all aggregated data has been disaggregated. This data set retains the aggregation coding used in the original data set in order to indicate where the disaggregation took place.

1. All values of the variables 5 through 9 of this size class are aggregates of the data of this and one or more subsequent size classes. (Indicated in the original source by a column of five underscored figures.)

3. The data for variables 5 through 9 of this size class is included i in a previous aggregation. (Indicated in the original source by a column of five D's.)

2,4. Analogous to codes 1 and 3 for

variable 9 (capital expenditures) only. (Indicated in the report by an underscoring and a subsequent D for capital expenditures.)

5. All values in variables 5 through 9 of this size class are aggregates of the data of this and one or more subsegent size classes. In addition, the data for variable 9 (capital expenditures) is included in a previous aggregation. (Indicated in the report by a column of four underscored figures, and a D for capital expenditures.)

If both codes 3 and 4 are valid, code 3 is recorded.

2. For some size classes there are no data observations. Variables 3 through 9 in the corresponding records contain zeroes. Missing values are recorded as zeroes.

3. The original size distribution data was summed and checked against the industry totals. A number of errors were found. The Bureau of the Census was able to correct most of these errors, including all the major ones.

The Census Size Distribution Disaggregation Procedure for Undisclosed Data

Census rules on preserving firm confidentiality complicate use of the Census size distribution data. In size class cases where confidentiality is an issue, the Census does not disclose values of variables within those size classes. Instead, the variable totals in two or more adjacent size classes are grouped together and given as one total. The one exception to this aggregation procedure is the number of establishments which is given for each size class in all industries. In order to estimate plant economies of scale for the largest number of industries possible, the grouped data must be disaggregated and assigned to the appropriate size classes. The method we utilize follows that of Horowitz (1971).

Values of each disaggregated variable are derived by multiplying the aggregated variable by a fractional weight unique to each size class, which sums to one over all of the size classes. The weight consists of the number of establishments in each size class multiplied by the geometric mean of its employment class boundaries divided by the summed value over all of the aggregated size classes. If size classes M to N are aggregated, the estimated value of the variable in question X(i) for size class i, is:

$$X(i) = \left[\frac{(EST(i))(G(i))}{\sum_{i=M}^{N} (EST(i))(G(i))} \right] AGG(X)$$

where,

EST(i) = the number of establishments in size class (i)

G(i) = the geometric mean of the upper and lower bounds of the number of employees in size class (i). For example, for a size class containing establishments employing

250 to 499 employees, $G(i) = \sqrt{(250)(499)} = 353$

$X(i)$ = estimated value of variable X in size class (i)

$AGG(X)$ = the aggregated value of the undisclosed variable, X, in size classes M through N.

The bracketed expression on the r.h.s. is the weight which we call $W(i)$.

By inspection,

$$\sum_{i=M}^{N} W(i) = 1 \quad \text{and}$$

$$\sum_{i=M}^{N} X(i) = AGG(X)$$

There is a major problem with the above procedure: the weighting fraction is undefined when the data in the largest size class is suppressed because the largest size class is open-ended and G(i) cannot be calculated. We resolve this by allocating to the next smaller size classes their share of the aggregated variable and placing the remainder or residual value in the largest size class. More formally, let N be the largest size class, then

$$X(i) = (W(i))(AGG(X)) \qquad i \neq N$$

$$X(N) = AGG(X) - \sum_{i=M}^{N-1} (W(i))(AGG(X)) = \left(1 - \sum_{i=M}^{N-1} W(i)\right) AGG(X)$$

Since G(i) is not known for the largest size class, W(i) for other than the largest size class is redefined as:

$$W(i) = \frac{(EST(i))(G(i))}{EMPL(T)}$$

where,

$EMPL(T)$ = total employment for size classes M through N.

The above expressions are equivalent to the following:

$$EMPL(i) = [EST(i)][G(i)], \quad i \neq N$$

$$EMPL(N) = EMPL(T) - \sum_{i=M}^{N-1} EMPL(i),$$

and

$$X(i) = \frac{EMPL(i)}{EMPL(T)} \times AGG(X),\ i=M,\ldots,N$$

We are assuming in this procedure that the output-labor ratio does not vary from size class to size class.

In certain cases, this method of handling nondisclosure in the upper size classes creates a further problem. In some cases, using the above disaggregation procedure, employment per establishment in the largest size class turns out to be less than the lower bound of employment for the largest size class, i.e., less than 2500 employees.* For these cases, we simply assumed that employment per establishment in the largest size class is 2500 employees. Thus, we estimate total employment in the largest size class (designated N) as follows:

$$EMPL(N) = (EST(N))(2500),$$

and other variables, such as value of shipments or value added, are estimated by:

$$X(N) = EMPL(N)\ \frac{AGG(X)}{EMPL(T)}$$

In these cases we first allocated to the top size class its estimate, X(N), and placed the remainder of the aggregated variable in the next smaller size class.

$$X(N-1) = AGG(X) - X(N)$$

In the exceptional cases in which more than the largest two classes are aggregated and we have to constrain the average employment size to equal 2500 in the largest size class, we use the following procedure: 1) assign

---

*This same problem occurred for some of the smaller size classes in certain industries, but infrequently; and in the smaller size classes, the degree of error in absolute terms is fairly trivial. We left them alone.

to the largest size class the minimum values it needs; 2) assign to the lowest aggregated size classes values according to the formula:

$$X(i) = \frac{(EST(i))(G(i))}{\sum_{i=M}^{N-1} \left(EST(i)\right)\left(G(i)\right)} \times (AGG(X)-X(N))$$

3) put into the next-to-largest size class the residual of the aggregated variable.

There is chronic understatement or overstatement of variables in the largest size class when it involves a disclosure problem. The best we can say is that disclosure problems in the top size classes introduces errors of an unknown magnitude and direction in these size classes. We intend to conduct a simulation study to estimate the degree of error in this disaggregation procedure. The procedures outlined above are now illustrated by way of examples of each case in clarifying the procedure. In the 1967 Census establishment size distribution data, the Census aggregated together size classes 8 and 9 in SIC 2099. Size class 8 is defined as establishments with 500 to 999 employees, inclusive, and size class 9 is defined as establishments with 1000 to 2499 employees, inclusive. Therefore, the geometric mean of the size classes are: $G(8) = 706.8$ and $G(9) = 1580.8$. Size classes 8 and 9 have 7 and 2 establishments, respectively. The value of shipments (VS) for the combined size classes is 357.6. We have sufficient information to perform the computations. The weights are calculated as follows:

$$W(8) = \frac{(7)(706.8)}{(7)(706.8) + (2)(1580.8)} = .61$$

$$W(9) = \frac{(2)(1580.8)}{(7)(706.8) + (2)(1580.8)} = .39$$

The disaggregated estimate of value of shipments is:

$$VS(8) = (.61)(357.6) = 218.1$$

$$VS(9) = (.39)(357.6) = 139.5$$

And now an example in which the largest size class is aggregated. Size classes 9 and 10 are aggregated together in SIC 2511. Neither G(10) nor W(10) are definable. G(9) = 1580.9, as before. Size classes 9 and 10 have 10 and 2 establishments, respectively. Total employment for the two size classes is 21,200, and total value of shipments is 289.1 (in thousands).

$$W(9) = \frac{(10)(1580.9)}{21,200} = .75$$

$$EMPL(9) = (.75)(21,200) = 15809$$

$$EMPL(10) = 21,200 - 15809 = 5391$$

$$VS(9) = \frac{15809}{21200} = 289.1 - 215.6$$

$$VS(10) = 289.1 - 215.6 = 73.5$$

SIC 2221 also has size classes 9 and 10 aggregated together, but the procedure we used in SIC 2511 doesn't work—we get an average employment in the largest size class of less than 2500. So, we set, assuming EST(10) = 1,

$$EMPL(10) = (1)(2500) = 2500$$

Total employment is 26,900, so

$$EMPL(9) = 26,900 - 2500 = 24,400$$

Total value of shipments is 217.4

$$VS(10) = \frac{2500}{26,900} \quad 217.4 = 20.2$$

$$VS(9) = \frac{24400}{26,900} \quad (217.4) = 197.2 = 217.4 - 20.2$$

Horowitz, Ira, "Numbers-Equivalents in U.S. Manufacturing Industries: 1954, 1958, and 1963," Southern Economic Journal, 37 (April 1971), p. 398, note 6.