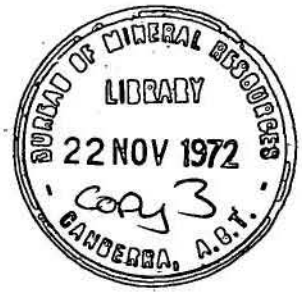

COMMONWEALTH OF AUSTRALIA

DEPARTMENT OF
NATIONAL DEVELOPMENT

BUREAU OF MINERAL
RESOURCES, GEOLOGY
AND GEOPHYSICS

Record 1972/100

A COMPUTER PROGRAM FOR PLOTTING CUMULATIVE FREQUENCY CURVES AND CUMULATIVE FREQUENCY GRADIENT CURVES

by

S. Henley

Record 1972/100

A COMPUTER PROGRAM FOR PLOTTING CUMULATIVE FREQUENCY CURVES AND CUMULATIVE FREQUENCY GRADIENT CURVES

by

S. Henley

CONTENTS

SUMMARY

A computer program has been developed to plot cumulative frequency curves for untransformed or logarithmically transformed data against a probability scale designed such that data with respectively normal or lognormal distributions appear as straight line plots. A cumulative frequency gradient curve is also plotted, which has similar appearance and significance to a histogram but avoids some of the problems associated with arbitrary choice of class intervals.

1. CUMULATIVE FREQUENCY CURVES

The principal advantage of plotting cumulative frequency curves on probability paper, over the more conventional use of histograms, is that one may immediately identify a distribution by inspection as normal (or lognormal) or otherwise (Tennant & White, 1959). Though not theoretically correct (Jizba, 1959; Krumbein & Graybill, 1965), normal and lognormal distributions very often approximate closely the true distributions (Shaw, 1961) and are empirically useful.

Manual plotting of cumulative frequency curves requires the use of 'probability paper', in which one scale is linear or logarithmic and the other scale is derived from the normal distribution function; i.e. values close to the median plot closer together than those at the extremities of the distribution, and so compensate for the greater density of values close to the median.

In the computer program, the co-ordinate on the normal-distribution axis must be calculated. A subprogram ERF is available through the CSIRO SRLIST (Sub-Routine LIST) from which the normal distribution function may be computed via the entry point PROBNORM. However, as this gives a value of the probability of values occurring below a known co-ordinate, it is the inverse of what we require. A function subprogram, XINVNDF, has been written to obtain an approximate value of the co-ordinate from a given probability level, using a 'half-difference' method, which converges to seven-figure accuracy within about 20 iterations.

The program sorts the data values for each variable into ascending order, and each value is plotted according to its position in the list (the empirical probability of values in the population occurring below the given value) - the probability co-ordinate is given by the relation: $y = \text{XINVNDF}\ (i/n+1)$, for the ith sample of n samples.

Figure 1 illustrates the form of the plotted cumulative frequency curve, using CaO data from a suite of pelitic sedimentary rocks.

2. CUMULATIVE FREQUENCY GRADIENT CURVES

The cumulative frequency curve is, by definition, the integral of the probability density function, of which the most commonly used form is the histogram. The histogram, however, is merely an approximation to the actual function, and particularly with small numbers of samples (less than 30 or 40), the choice of class intervals and class limits critically affects both the appearance of the graph and the chi-square derived from it. In many studies, cumulative frequency curves have been derived from histograms, and are defined by few points. The computer program now available, however, plots every point on the cumulative curve, and it would seem logical to attempt construction of a more meaningful 'histogram' by the reverse process: rather than integrating from the density curve, we differentiate from the cumulative curve. In more precise terms, in fact, we plot the gradient of the cumulative frequency curve, calculated by:

$g = (k-1/(n(V_{i+k-1}-V_i)$, where n is the number of samples, V_i and V_{i+k-1} are the ith and (i+k-1)th values, g is the abscissa of the ith plotted point, and k is a positive integer. The ith point is plotted at an x co-ordinate of:

$$x = \sum_{j=1}^{i+k-1} V_j/k$$

To obtain the most accurate value of the gradient, a value of 1 must be used for k, resulting in n-1 plotted points. Random clustering of data values caused both by sampling error and by 'stepping' effects in analytical data , results in graphs that are very spiky (Figure 2), but smoothing may be effected by averaging the gradient over a number of points. The choice of k will, of course, depend on several factors, such as analytical precision, the extent of local clustering, and the degree of smoothing required (k=n-2 would give a rather flat straight line joining two points near the mean). For general purposes, k=n/10 would seem a reasonable value to adopt, giving a rather smooth graph with 9n/10 plotted points, as opposed to the conventional 10 to 20 classes in a histogram.

3. PROGRAM OPERATION

Data may be supplied in any format, with one restriction: data values for each sample must be preceded by an identification field of two words in the format (A8,A1) - i.e. 9 alphanumeric characters, required for an 8-character registered number and a 1-character fraction code. Data may be supplied for up to 40 variables and 400 samples, though these limits are arbitrary and may be changed simply by replacing the DIMENSION card. The product of NV (number of variables) and NS (number of samples), however, must not exceed about 21 000.

The data cards are preceded by 3 data control cards: the first defines NV and NS for the data set in question; the second defines the input format of data cards; the third give details of the first variable to be plotted. The data may be followed by any number of cards of type 3, for the same or different variables.

Deck structure

The complete deck consists of the program deck in FORTRAN or binary cards, with associated system control cards, followed by the data control cards and data deck. The last card in the deck is an 'end of document' card to terminate running of the job.

FORTRAN program deck

1. Job card punched from col.1. *JOB, chargecode, CUMUPLOT, time limit
2. Equip card punched from col.1. *EQUIP, 1=PL, 2=PL.
3. Fortran card (or Kwiktran card on the CSIRO computer) punched from col.1. *FTN,X,L (or *KTN,X,L)
4. Program deck in FORTRAN, listed below.

5. Scope card, punched from col. 10 SCOPE
6. Load card, punched from col.1 *LOAD
7. Run card, punched from col.1 *RUN, time, maximum no. of lines of print.

Binary program deck

1. Job card (as above)
2. Equip card (as above)
3. Binary program deck
4. Run card (as above)

Data deck to follow immediately either the FORTRAN or Binary program deck.

Card (a) Columns 1-3, right justified. NV, number of variables.
4-6, right justified. NS, number of samples.

Card (b) Variable format card. FORTRAN format definition, enclosed in brackets. The first two words must be a 9-character identification field in A8,A1 format.

Card (c) Columns 1-3, right justified. Index number of variable to be considered.
4-6, -1 if logarithmic transformation required, +1 if not.
7-14 Name of the variable (alphanumeric)
15-24 Estimated minimum value of the variable (must be positive if logarithmic transformation is requested)
25-34 Estimated maximum value of the variable.

(d) Data cards in the specified format, for NS samples.

If plots are required for further variables or for the same variable with different values inserted in card type (c), any number of repetitions of card (c) may follow the data cards.

End-of-job card. Punch from col. 1. *EOD

Output

A set of plots of cumulative frequency curves is obtained on unit 1, while a set of cumulative frequency gradient curves is obtained on plotter unit 2. The scales are annotated on the former (if the data are logarithmically transformed, the logarithm value is plotted); on the latter the horizontal scales are the same, while the vertical scale is simply a proportional linear scale with no meaningful units. Output on the lineprinter gives the minimum, median, and maximum for each plotted variable.

4. REFERENCES

JIZBA, Z.V., 1959 - Frequency distribution of elements in rocks. Geochim. Cosmochim. Acta., 16, pp. 79-82.

KRUMBEIN, W.C., & GRAYBILL, F.A., 1965 - AN INTRODUCTION TO STATISTICAL MODELS IN GEOLOGY. New York, McGraw-Hill.

SHAW, D.M., 1961 - Element distribution laws in geochemistry. Geochim. Cosmochim. Acta., 23, pp. 116-134.

TENNANT, C.B., & WHITE, M.L., 1959 - Study of the distribution of some geochemical data. Econ. Geol., 54, pp. 1281-1290.

APPENDIX: Listing of program CUMCURVE

5,5DAD 05/09/72

```
      PROGRAM CUMCURVE
      DIMENSION ID(2),V(40,400),FMT(10),WORDS(5)
      DIMENSION YY(400)
      COMMON K
      DATA (WORDS=32HCUMULATIVE FREQUENCY CURVE FOR  )
      READ 10,NUMV,NUMS
      READ 10,K,LOGUNIT
      ANUMS=NUMS
      IF (LOGUNIT.EQ.0) LOGUNIT=60
      IF (NUMS/2.EQ.ANUMS/2.0) 3,4
    3 MED=1  $  GO TO 5
    4 MED=0
    5 MN=NUMS/2
      READ 20,FMT
      READ 10,INDVAL,LOG,WORDS(5),XMIN,XMAX,K
   10 FORMAT (2I3,A8,2F10,I6)
   20 FORMAT (10A8)
      PRINT 25,NUMS,NUMV,INDVAL,WORDS(5)
   25 FORMAT (1H1,*DATA READ FOR*,I3,* SAMPLES,FROM DETERMINATIONS OF*,
     1I3,* VARIABLES*/* VARIABLE NO.*,I3,1H,,A8,*USED FOR THIS ANALYSIS*
     2)
      DO 30 I=1,NUMS
      READ FMT, ID,(V(J,I),J=1,NUMV)
   30 PRINT 35,ID,(V(J,I),J=1,NUMV)
   35 FORMAT(1H ,A8,A3,15F7.2/(13X,15F7.4))
   31 DO 311 I=1,NUMS
  311 YY(I)=V(INDVAL,I)
      CALL SINGSORT (YY,NUMS)
      IF (MED.EQ.1) 32,33
   32 AM=YY(MN)+YY(MN+1)
      AM=0.5*AM  $  GO TO 34
   33 AM=YY(MN)
   34 PRINT 341, YY(1),AM,YY(NUMS)
  341 FORMAT(* MIN=*,F10.4*  MEDIAN=F10.4*  MAX=*F10.4)
      CALL PLOT (1.,1.,2,1)
      CALL PLOT (-1.,-8.,1,1)
      CALL PLOT (0.,5.,3,1)
      CALL PLOT (0.,-5.,4,1)
      CALL PLOT (8.,-5.,4,1)
      DO 37 I=1,11
      F=I-6  $ E=0.7*F
      Q=PROBNORM(E)*100.0
      ENCODE (5,36,IQ) Q
   36 FORMAT(F5.2)
      CALL PLOT (0.,F,3,1)
      CALL PLOT (-0.2,F,4,1)
      CALL PLOT (-1.0,F-0.1,3,1)
   37 CALL TEXT(IQ,5,2,1)
      PRINT 375, WORDS(5),XMIN,XMAX
  375 FORMAT (1H0,A8,* VALUES FROM=*,F10.5,* TO=*,F10.5)
      IF (LOG) 40,40,45
   40 XLO= ALOG10(XMIN)
      XHI= ALOG10(XMAX)
      PRINT 405,XLO,XHI
  405 FORMAT (* LOG TRANSFORMED DATA = LOG X FROM=F10.5,* TO=*,F10.5)
   42 DO 41 I=1,NUMS
      IF (YY(I).LE.0.0) YY(I)=XMIN
  441 YY(I) = (ALOG10(YY(I))-XLO)*8.0/(XHI-XLO)
   41 CONTINUE
      GO TO 50
```

```
   45 XLO= XMIN
      XHI= XMAX
      DO 46 I=1,NUMS
   46 YY(I)=(YY(I)-XLO)*8.0/(XHI-XLO)
   50 DO 55 I=1,5
      F=I-1
      Q=XLO + F*(XHI-XLO)/4.0
      ENCODE (7,51,IQ) Q
   51 FORMAT(F7.2)
      CALL PLOT (F*2.,-5.,3,1)
      CALL PLOT (F*2.,-5.2,4,1)
      CALL PLOT (F*2.-0.8,-5.5,3,1)
   55 CALL TEXT(IQ,6,2,1)
      AI=0.0
      DO 60 I=1,NUMS
      BI=AI
      AI=I
      AI=AI/ANUMS
      X=YY(I)
      Y=XINVNDF(0.5*(BI+AI)) / 0.7
      CALL PLOT (X,Y,3,1)
   60 CALL TEXT (1H*,.0,1,1)
      CALL PLOT (-0.5,-6.0,3,1)
      CALL TEXT (WORDS,40,2,1)
      CALL PLOT (-1.0,8.0,3,1)
      CALL CUMUGRAD(YY,NUMS,WORDS(5))
      CALL PLOT (-1.0,8.0,3,2)
      READ 10,INDVAL,LOG,WORDS(5),XMIN,XMAX
      IF (EOF,60) 99,31
   99 CALL EXIT
      END
```

```
      FUNCTIONXINVNDF(P)
      J=1
      TOL=0,00000001
      IF (P,EQ,1,0) GO TO 100
      IF (P,LT,0,5)10,20
   10 J=-1 $ P=1,0-P
   20 X=0,0
      DIFF=2,0
   35 PTEST=PROBNORM(X)
      PP=PTEST-P
      IF (X,GT,3,99999999) RETURN
      IF(ABS(PP)-TOL) 90,90,40
   40 IF(PP) 50,90,60
   50 X=X+DIFF
      GO TO 70
   60 X=X-DIFF
   70 DIFF=DIFF*0,5
      GO TO 35
   90 XINVNDF=SIGN(X,J)
      RETURN
  100 XINVNDF=4,0
      RETURN
      END
```

```
      SUBROUTINE CUMUGRAD (Y,N,TITLE)
      DIMENSION Y(N)
      DIMENSION WORDS(5)
      COMMON K
      DATA (WORDS=32HCUM, FREQ, GRADIENT CURVE FOR    )
      WORDS(5)=TITLE
      IF (K,EQ,0) K=N/10
      M=N-K
      R=K
      A=10,0/M
      CALL PLOT (1.,1.,2,2)
      CALL PLOT (-1.,-8.,1,2 )
      CALL PLOT (0.,10.,3,2)
      CALL PLOT (0.,0.,4,2)
      CALL PLOT (8.0,0.0,4,2)
      CALL PLOT (0.,0.,3,2)
      L=3 $ LL=0
      DO 50 I=1,M
      L=MAX0(L,LL)
      LL=4
      IH=I
      II=I+K-1
      AA=A*(K-1)
      SUMXY=SUMX=SUMY=0,0
      DO 1 J=I,II
      SUMXY=SUMXY+J*Y(J)
      SUMY=SUMY+J
    1 SUMX=SUMX+Y(J)
      SSXY=SUMXY-SUMY*SUMX/R
      SSXS=SUMX**2 /R
      V=SSXY/SSXS
      IF DIVIDE CHECK 5,10
    5 V=10,0
   10 X=SUMX/R
   24 IF (V,GT,10,0) PRINT 25,X,V
   25 FORMAT(2F10,5)
      IF (V,GT,10,0) V=10,0
      CALL PLOT (X,V,L,2)
   50 CONTINUE
      CALL PLOT (-0,5,-1,0,3,2)
      CALL TEXT (WORDS,40,2 ,2)
      CALL PLOT (-1,0,10,0,3,2)
      RETURN
      END
```

```
      FUNCTION ERF(X1)
C
C *ERF* EVALUATES THE ERROR FUNCTION, THE NORMAL DISTRIBUTION
C  FUNCTION, OR THE SIGNIFICANCE LEVEL OF THE NORMAL DISTRIBUTION
C  BY THE FUNCTION CALLS *ERF(X)* FOR THE ERROR FUNCTION, *PROBNORM(X)*
C  FOR THE NORMAL DISTRIBUTION AND *SIGNIF(X)* FOR THE SIGNIFICANCE
C  LEVEL.
C THE ROUTINE IS THE SAME AS *C3 CSIR PRNORM* WITH EXTRA ENTRY
C  POINTS TO MAKE IT MORE VERSATILE WITH ONLY A SLIGHT LOSS OF SPEED.
C
C MODIFIED BY N.R.PUMMEROY, CSIRO DIV. OF COMPUTING RESEARCH, 17/2/71.
C
                          NTRY=-1 $ XT=X1 $ X1=X1*1.414213562   $ GO TO 9
      ENTRY SIGNIF    $ NTRY= 1 $ XT=X1 $ X1=-ABS(X1)           $ GO TO 9
      ENTRY PROBNORM $ NTRY= 0
C
 9    X=X1
 10   Z=1.0
      IF(X)30,20,40
 20   ERF=0.5
      GO TO 100
 30   X=-X
      Z=-Z
 40   IF(X.GT.6.)GO TO 90
 60   IF(X.GE.2.)GO TO 80
 70   X=X*X*.25
      Z=       ((((((((( .00012481800 *X-.00107520405)*X+.00519877502)*X
     $                   -.01919829200)*X+.05905403564)*X-.15096875136)*X
     $                   +.31915293269)*X-.53192300731 )*X+.79788456059)*X1
      GO TO 90
 80   X=X*0.5-2.0
      Z=((((((((((((((-.00004525566 *X+.00015252929)*X-.00019953813)*X
     $                   -.00067690499)*X+.00139060428)*X-.00279462082)*X
     $                   -.00203425487)*X+.00654979121)*X-.01057627301)*X
     $                   +.01163044738)*X-.00927945334)*X+.00539357911)*X
     $                   -.00214126874)*X+.00053531085)*X+.99993665752)*Z
 90   ERF=(Z+1.0)*.5
 100  IF(NTRY)110,130,120
 110  X1=XT $ ERF=ERF*ERF-1.0 $ RETURN
 120  X1=XT $ ERF=ERF*ERF
 130  RETURN
      END
```

```
      SUBROUTINE SINGSORT(A,N)
C     CACM ALGORITHM 343  R.C. SINGLETON   17,9,68
C     TO SORT INTEGERS, USE    INTEGER A,T,TT
      DIMENSION A(N)
      DIMENSION IU(16),IL(16)
      M=1
      I=II=1  $  J=N
    5 IF(I.LT.J) 10, 70
   10 K=I
      IJ=(J+I)/2
      T=A(IJ)
      IF(A(I).GT.T)  21, 20
   21 A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
   20 L=J
      IF(A(J).LT.T)  41, 40
   41 A(IJ)=A(J)
      A(J)=T
      T=A(IJ)
      IF(A(I).GT.T)  31, 40
   31 A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      GO TO 40
   30 A(L)=A(K)
      A(K)=TT
   40 L=L-1
      IF(A(L).LE.T)  51, 40
   51 TT=A(L)
   50 K=K+1
      IF(A(K).GE.T)  61, 50
   61 IF(K.GT.L)  71, 30
   71 IF(L-I.GT.J-K)  81, 60
   81 IL(M)=I
      IU(M)=L
      I=K
      M=M+1
      GO TO 80
   60 IL(M)=K
      IU(M)=J
      J=L
      M=M+1
      GO TO 80
   70 M=M-1
      IF(M.EQ.0)RETURN
      I=IL(M)
      J=IU(M)
   80 IF(J-I.LT.11)  11, 10
   11 IF(I.NE.II)   6, 5
    6 I=I-1
   90 I=I+1
      IF(I.NE.J)  12, 70
   12 T=A(I+1)
      IF(A(I).GT.T)  13, 90
   13 K=I
  100 A(K+1)=A(K)
      K=K-1
      IF(T.GE.A(K))  14, 100
   14 A(K+1)=T
```

```
      GO TO 90
      END
```

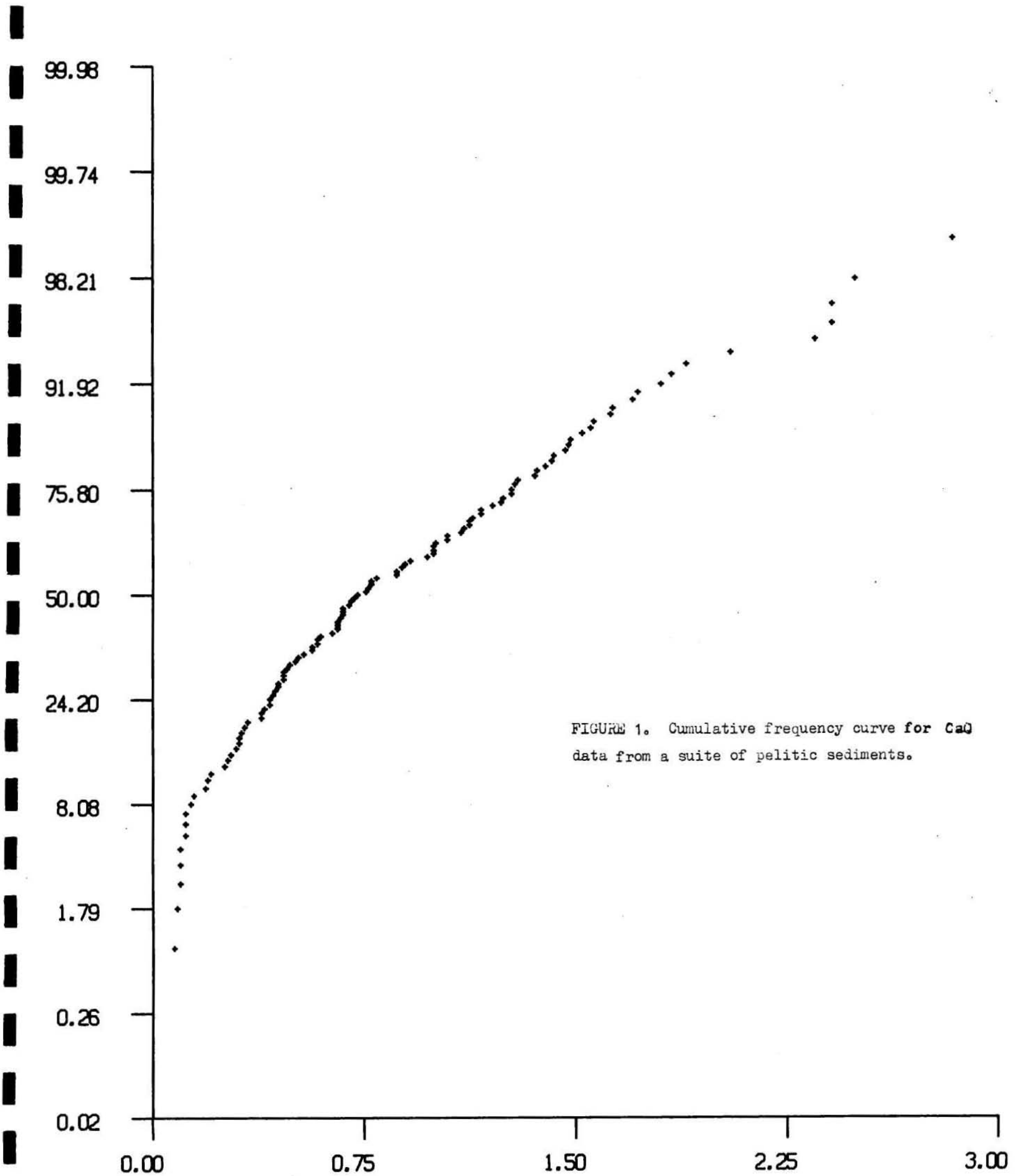

FIGURE 1. Cumulative frequency curve for CaO data from a suite of pelitic sediments.

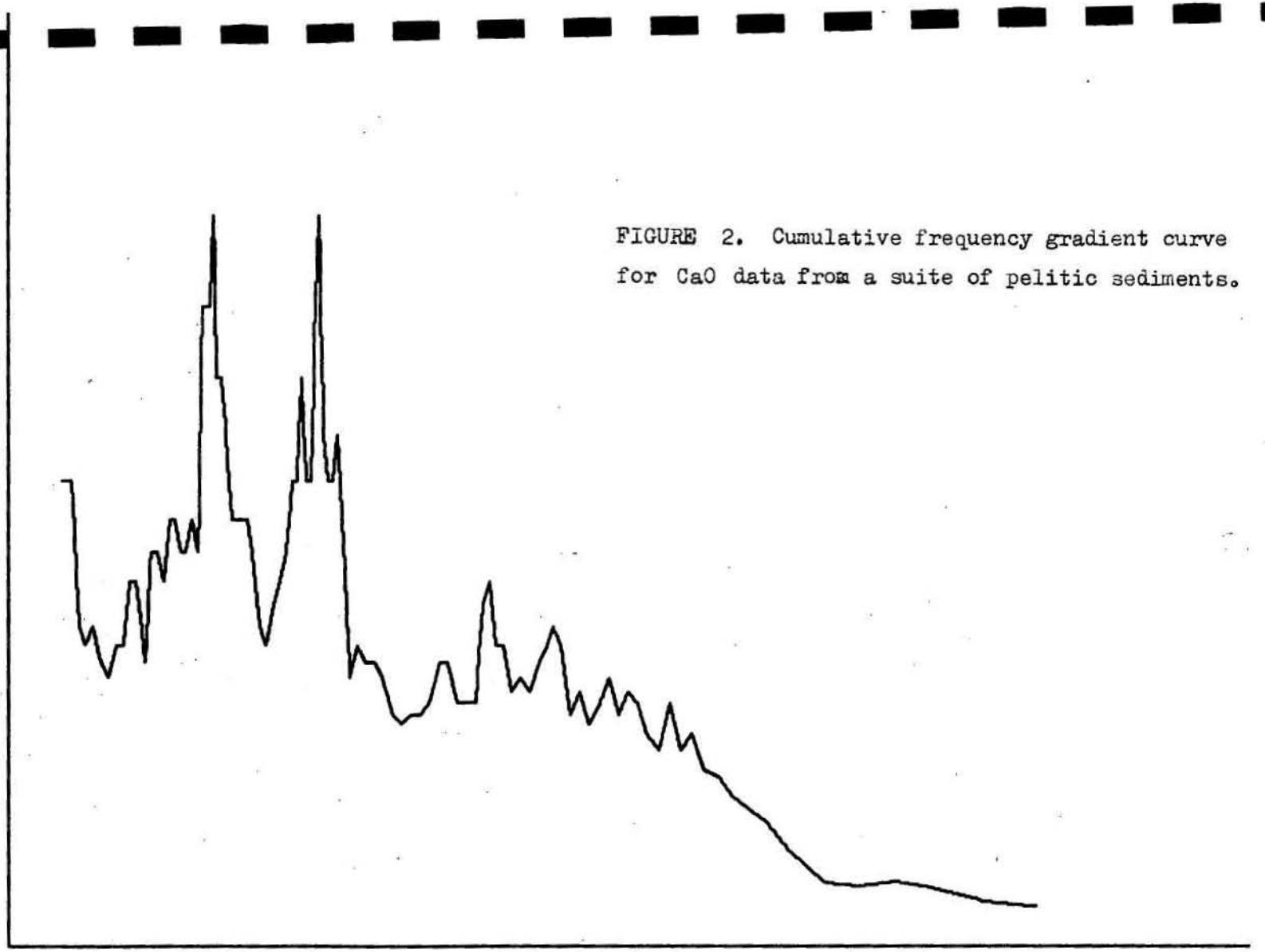

FIGURE 2. Cumulative frequency gradient curve for CaO data from a suite of pelitic sediments.