EXPLORATORY DATA ANALYSIS: »SOFT» STATISTICS FOR ARCHAEOLOGISTS

TAPIO SEGER

Isonnevantie 13 A 4, SF-00300 Helsinki 30

After a few years of development and testing in various fields, Exploratory Data Analysis (Tukey, 1977) can now be considered a totally fresh and powerful branch of statistics as well as a minor revolution in statistical thinking. Traditional »mainstream» statistics are mostly concerned with Probability Theory, ie. trying to confirm probabilities of events, and on the other hand, use »heavy» multivariate methods, the mathematics of which are beyond the scope of non-experts. In contrast, EDA is totally aimed at exploration. The idea is simply to obtain and graphically display all the information a group of data (a »batch» in EDA jargon) can produce of itself. EDA is also *relatively* easy to use even by hand-calculation.

Another important advantage of EDA is its independence of the Normal Law in theory as well as in practice. In other words, it is of little relevance for the analysis, whether the data is distributed »normally» or in any other way; whether the data is a proper sample in the statistical sense or not etc. This makes it an extremely suitable tool particularly in archaeology, in the material record of which a Gaussian distribution is, at least in my experience, a rarity, and where groups of data are not generally proper statistical samples of anything. However, EDA is still rather unknown in Europe and, to my knowledge, has not so far been applied to archeological problems anywhere.

Other general features of EDA are its ability to identify stray-values, here called *outliers,* and its overall in-built resistance to them, the latter due to the fact that EDA operates with median-based concepts and routines. In traditional confirmative statistics, even such every-day methods as the arithmetic mean and standard deviation are easily thrown off course by one or two extraordinarily high or low values (in contrast to the mean, the median is always the mid-value or the average of two mid-values, however much the extreme values differ from the rest of the batch). This resistance allows outliers to be included in the analysis, while in mainstream statistics they are generally omitted (eg. Bølviken *et al.,* 1982), which naturally reduces the available information.

Although EDA routines can be carried out by hand, access to a proper computer is obviously of great help. The most important routines of EDA were computerized recently (Velleman and Hoaglin, 1981), and are included in the MINITAB (TM) statistical package (Ryan *et al.* 1982). Most of the following analysis was carried out using this package.

All in all, in my opinion EDA methods are highly recommended as the initial steps of numerical analysis in any field. Exploration comes first, confirmation after there is something to confirm (Tukey 1977, vi—vii). In problems of archaeology and the humanities in general, EDA is usually quite sufficient in itself, in some extreme cases

Table 1. The six analyzed hoards, the number of coins and that of dated coins in them as well as the *terminus post quem* values of each hoard.

Hoards	coins *n*	dated coins *n*	*t.p.q.*
Geta Svedjelandet	107	105	842
Jomala Hammarudda	157	149	857
Saltvik Bertby	859	741	890
Östergeta Västergård	28	25	954
Saltvik Åsgårda	81	75	958
Finström Emkarby	84	77	958
Total	1316	1172	—

perhaps combined with advanced analysis of variance and/or some multivariate method. Interpretation of the results, as with any method, is of course left to the archaeologist's common sense and general knowledge of the subject.

The Viking Period Oriental coin finds from Åland were chosen as example material. Of the total of 1346 coins published by Granberg (1966), single finds, grave finds and hoards with less than 10 coins were omitted (all of these are, however, well inside the range of the analyzed material). This leaves six largish hoards with altogether 1316 coins, of which 1172 are dated (Table 1; also coins dated to the accuracy of a decade or the reign of the ruler who had them minted are accepted). One is a Byzantine and all the others Islamic or pre-Islamic silver coins. In the analysis each coin is assumed to date to its (earliest possible) minting year and each hoard to that of its youngest coin. This is a construction quite commonly used in numismatical studies (eg. Blackburn and Metcalf, 1981, *passim*), and the only sensible basis for a statistical analysis of coin deposits (see also Sarvas, 1972, 10—12).

Before the analysis, a few simple concepts, some of them familiar, others probably not, have to be defined (Tukey, 1977, ch. 2):

extremes = the maximum and minimum values of a batch,
median = the mid-value or the average of two mid-values of the ordered data,
hinges = essentially quartiles (see Tukey 1977, 32—33), ie. the mid-values between median and the extremes,
range = upper extreme — lower extreme,
H-spread = upper hinge — lower hinge,
inner fences = (lower hinge) — 1.5 × (H-spread) and (upper hinge) + 1.5 × (H-spread),
outer fences = (lower hinge) — 3 × (H-spread) and (upper hinge) + 3 × (H-spread),
adjacent values = those closest to the inner fences but still inside them,
out values = those outside the inner fences but still inside the outer ones,
far-out values = those outside the outer fences.

The first step in EDA is usually a *stem-and-leaf display* (Tukey 1977, ch. 1; Velleman and Hoaglin 1981, ch. 1) of the data (Fig. 1). It combines the essential features of standard tables and histograms, as well as identifies potential outliers. The latter (especially those far-out) should first be checked in any analysis against the possibility that they might be sampling, input or calculating errors. In case they are none of these, they must be considered indicating significant deviations from the overall behaviour of the batch. The display is generally trimmed at the inner fences, and low outliers, if any,

```
STEM-AND-LEAF DISPLAY OF 'TOTAL'
LEAF DIGIT UNIT =  1.ØØØØ
1  2 REPRESENTS 12.

      LO   559, 582, 585, 588, 596, 6ØØ,6ØØ, 6ØØ, 617, 62Ø, 62Ø, 62Ø, 624,
           624, 672, 675,

 17   67   9
 2Ø   68   Ø99
 21   69   1
 26   7Ø   28899
 33   71   ØØØ1235
 37   72   5559
 45   73   11677799
 56   74   11122357899
 81   75   Ø1111456666777777777777788
 15Ø  76   ØØ11223333344444444567778888888888888888888888888888888888888888*
 238  77   ØØØØØØØ1111122222222233333333334444555555566666667777777777777788*
 29Ø  78   ØØØØØØØØ1111111222222333333444466667777777888889999
 352  79   ØØ1111125555555556666666666666666677777777778888888889999999
 48Ø  8Ø   ØØØØØØ1111222222222222222333333333333333333333444444444444444444*
(114) 81   ØØØØØØØ11111222222233333344444444444445566666666666666666666666666*
 578  82   Ø111122233334444559
 558  83   ØØØ123345555577777789
 537  84   12223666777777888888889
 514  85   ØØ11122222333333444444444444555666666777777777777777777778999999*
 444  86   ØØØØØØØØ1111222222222222222222222222233333333344444444444444444444*
 211  87   ØØØØØØ111112222233333333344444444
 177  88
 177  89   ØØØ355555566677889
 159  9Ø   ØØ11124555567778999
 14Ø  91   ØØØØØØ1111222222333333334445566678 9
 1Ø5  92   ØØØ12223666777899
  88  93   12222333333334444444445556666888999
  53  94   ØØØØØ12223355555556668899
  28  95   ØØØØ1111112222224444444445588
```

Fig. 1. A stem-and-leaf display of the coins (see text). The heading specifies the unit (1 year) with an example, »1 2 represents 12» (and not .12, 1.2 etc.). The »*»'s in the rightmost spaces of some lines signify overflow from the computer screen (the depths still provide a complete count).

are listed after the heading »LO» above the display and potential high outliers (none here) below it marked »HI».

As to the actual display, the second column from the left contains the »stem» part of the data items, while the rows to the right of it contain their »leaf» digits. Thus, the first line consists of the number (minting year AD of a coin) 679, the second reads 680, 689, 689, the third 691 etc. The column to the far left informs how many items there are on each line, cumulatively counted from both edges of the ordered data towards the median, while the line containing the latter is shown with the actual count of leaves on it within parenthesis. In other words, the display preserves the numerical information while presenting it as horizontal histograms. It is easy to see (Velleman and Hoaglin 1981, 1):

a) how wide the range is,
b) where the values are concentrated,
c) how nearly symmetric the batch is,
d) whether there are gaps with no items in them and
e) whether there are values straying markedly from the rest.

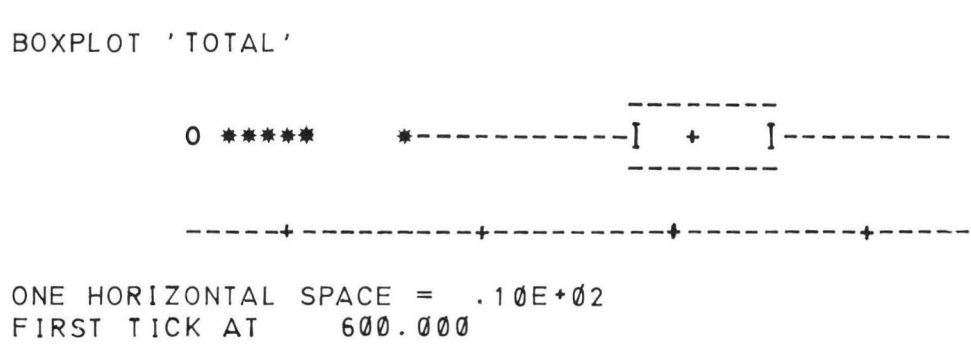

Fig. 2. A boxplot of the coins. The code »+» marks the median, the »I» 's mark the hinges, the »whiskers» run from the hinges to the adjacent values, while »*»'s signify the out values and »O»'s the far-out ones.

However, in order to easily obtain a general idea of the batch with a quick look, it should be displayed in a strongly summarized form. The method to do this is the *boxplot* (Tukey 1977, 39—48; McGill *et al.* 1978; Velleman and Hoaglin 1981, ch. 3), which by omitting as many details as possible, clearly shows the general behaviour of the batch (Fig. 2). Even a summarized diagram like this can transmit a surprising amount of information; practically half of the values are located inside the box, a quarter of these to the left of the median and *vice versa,* while half of the remaining ones lie to the left of the lower hinge and the other half to the right of the upper hinge. Moreover, boxplots are effective in visually displaying potential outliers. One defect of the method is the fact that boxplots cannot clearly detect possible bi- or multimodal tendencies in a batch, but this is no problem, as other routines, starting with the stem-and-leaf display, are able to do that.

In order to compare the six hoards with each other they are next displayed as *simultaneous boxplots* on the same axis in Fig. 3. The most striking feature is that they are clearly clustered into two separate groups with minimal overlap. Although eg. the hoard from Saltvik Bertby may well have been deposited not until the early 10th century, we may speak of three 9th century hoards and three 10th century ones for convenience. Another obvious fact is that the ranges of the 9th century hoards are much wider than those of the 10th century ones, partly because of the low outliers.

One step further in one-variable analysis is the *letter-value display* (Fig. 4; Tukey 1977, 53—54; Velleman and Hoaglin 1981, ch. 2; Ryan *et al.* 1982, 128—129). The »letter values» are the midpoints between the previous letter value(s) (starting with the median) and the edges of the ordered data, and are defined by their »depth» (from the nearer edge of the ordered data). The median (M) can thus be found at depth (M) = $(n+1)/2$, the hinges (H) at depth $d(\mathrm{H}) = \mathrm{int}(d(\mathrm{M})+1)/2$, the »eights» (E) at int $d((\mathrm{H})+1)/2$ etc., »int» naturally standing for the integer part function. Remaining letter values are found by continuing the routine and have no proper names, but are marked with the letters D, C, B, A, Z, Y, X etc. As an extended summary, the letter value display shows any inconsistencies and leanings towards high or low values (most clearly with tendencies in the »mids» and »spreads») in more detail than a boxplot.

The final method for one-variable data is the *rootogram* (Tukey 1977, 543—661; Velleman and Hoaglin 1981, ch. 9; Ryan *et al.* 1982, 136—137). The Gaussian distribution is one of the best known statistical features in existence, and the techniques measure deviations of the data from the Gaussian »normal». The routine first divides the data into bins (intervals) and displays it with a Gaussian distribution fitted to it. The fitting is based on *double root residuals* (Valleman and Hoaglin 1981, 265—267):

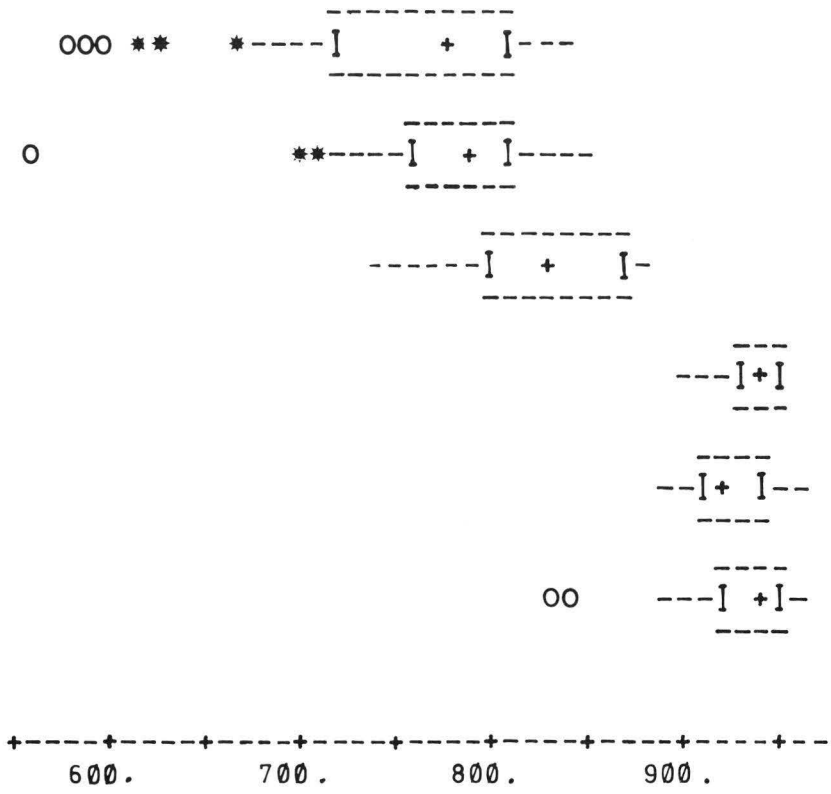

Fig. 3. The six hoards displayed as simultaneous boxplots on the same axis. The hoards from top to bottom are: Geta Svedjelandet, Jomala Hammarudda, Saltvik Bertby, Östergeta Västergård, Saltvik Åsgårda, Finström Emkarby.

LVALS 'TOTAL'

DEPTH	LOWER	UPPER	MID	SPREAD
N= 1172				
M 586.5	819.000		819.000	
H 293.5	791.000	867.000	829.000	76.000
E 147.0	768.000	907.000	837.500	139.000
D 74.0	757.000	934.000	845.500	177.000
C 37.5	730.000	945.000	837.500	215.000
B 19.0	689.000	951.000	820.000	262.000
A 10.0	620.000	954.000	787.000	334.000
Z 5.5	598.000	954.000	776.000	356.000
Y 3.0	585.000	955.000	770.000	370.000
X 2.0	582.000	958.000	770.000	376.000
1	559.000	958.000	758.500	399.000

Fig. 4. A letter-value-display of the data. The leftmost column shows the depth of the letter values from the nearer edge of the ordered data, the second column the values of the lower and the third those of the upper letter values of the same depth. The »mids»[1]are the average of the two, while the »spreads» are upper minus lower letter value of the same depth.

$$\begin{aligned} \text{DRR} &= \sqrt{2 + 4(\text{observed})} - \sqrt{1 + 4(\text{fitted})} && \text{if observed} \neq 0 \\ &= 1 - \sqrt{1 + 4(\text{fitted})} && \text{if observed} = 0, \end{aligned}$$

this in order to stabilize variance. The fit is found using the median and hinges of the batch in order to resist extraordinary values or bin counts (for details see Velleman and

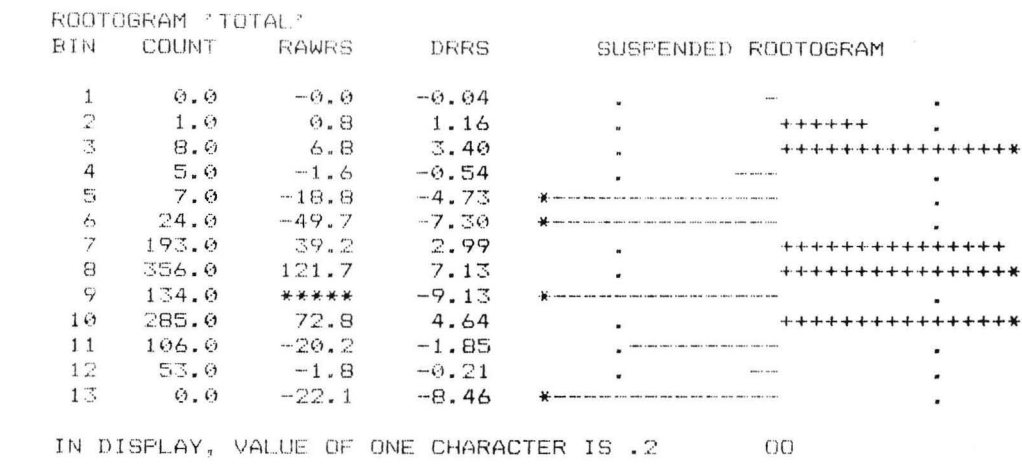

Fig. 5. The (suspended) rootogram of the total distribution of coins. The display shows the number of bins (intervals), the bin counts, »raw residuals» (original observed values minus fitted values) and double root residuals (see text) of the data. The graphic display is a plot of the latter and the plotting character is the sign of them (— or +). The vertical zero-line (ie. »the perfect Gaussian fit») passes the display between the double »0»'s. Two vertical dotted lines are plotted at —2 and +2, giving approx. 95 % confidence limits to the DRRs of the data. The »*»'s indicate DRRs beyond —3 and +3.

Hoaglin 1981, 267—274). The graphic display, called the *suspended rootogram,* is a plot of the DRRs (Fig. 5).

So far we have been concerned with one-variable analysis. Other methods are needed for bivariate (*x-y*) analysis. The basis of these methods in EDA is the standard scatter plot. A plot of »raw» values can of course submit information, but it is possible to go much further with a few techniques special to EDA. The first of them is *transformation* (re-expression) of the raw data values (Tukey 1977, chs. 3 and 6; Velleman and Hoaglin 1981, sections 2.4 and 5.8). The most common options (in increasing order) are $\ldots -1/y^3$, $-1/y^2$, $-1/y$, $-1/\sqrt{y}$, $\log(y)$, $\sqrt{y}$, [y^1 (no transformation)], y^2, y^3 ... Going lower or higher is necessary only extremely seldom. The aim of re-expressing either x or y values or, generally preferably both, to the same power, is to straighten out the plot, ie. to make the x-y relationship linear in case it is curved (as it usually is). Here only the re-expression of y values is needed as we are dealing with a sequence, in other words, the x values are evenly distributed (decades), in which case transformation causes no apparent change.

The method to establish which re-expression is the most effective in a particular case, is to fit two lines to the data, one to the left and the other to the right half of the scatter (if done manually, it is possible to use only three representative data points) (Tukey 1977, 171—203; Velleman and Hoaglin 1981, 135—142). When, after repeated transformations (in the latter case, only the three points need to be re-expressed), the slopes of the two lines are as close to each other as possible, the best re-expression is found. However, here a linear model is not too well suited to the rather complex pattern of all of the six hoards. So, in order to show an example functioning really well, only the three 9th century hoards are analyzed at this stage. Logs (base 10 logarithms) were found to be the perfect re-expression for them.

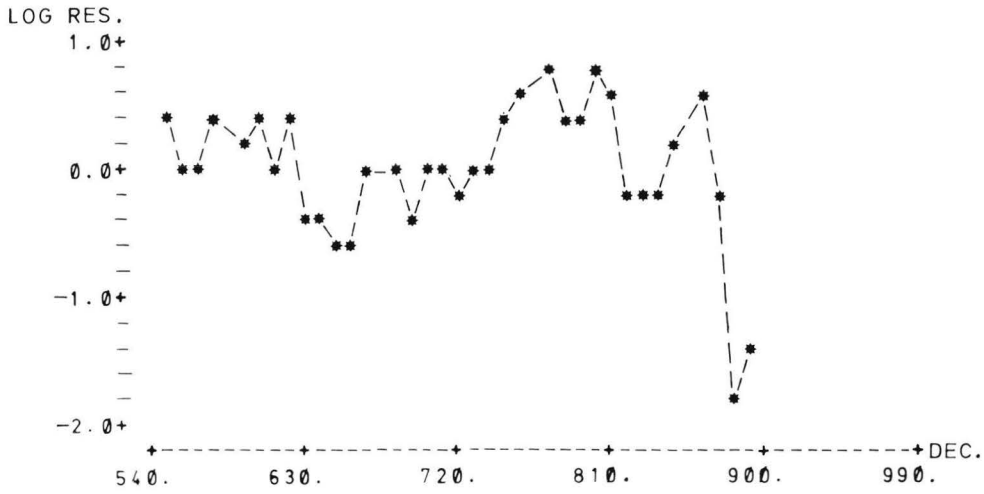

Fig. 6. The (logarithmic) residuals plot of the 9th-century hoards vs. decade (see text).

The fitted values obtained in this way can of course be plotted, but in fact they are rather uninteresting. The main point in finding the linear fit is to display the *residuals* (Tukey 1977, 113—114, 143—159):

data = fit plus residuals, and consequently
residuals = data minus fit.

Plotting the residuals, ie. the differences between the raw data points and the corresponding points fitted by the suitable transformation, effectively brings out the deviations from the fit as if seen through a microscope (Fig. 6), and these correspond to the deviations more or less hidden in the plot of the original raw data. However, to evaluate the importance of the wiggles in the residual plot, comparison with the raw values is always useful.

Another pair of methods for *x-y* analysis is the *smooth-rough* plotting sequence (Tukey 1977, 205—264; Velleman and Hoaglin 1981, ch. 6). Often it is useful to search for patterns more complicated and more general than a straight line. In the same way as in the previous definition it can be stated that:

data = smooth plus rough, and consequently
smooth = data minus rough, and
rough = data minus smooth.

Smoothing in general terms means applying techniques consisting of running medians (and averages) to the data to smooth the wiggles between a few consecutive values at a time in order to obtain a clearer view of the general behaviour of the batch without disturbing detail.

The smoothing routine used here for the six hoards (Fig. 7) is called in short-hand *4253H, twice* (Tukey 1977, ch. 7; Velleman and Hoaglin 1981, 171—177). It is a combination of several smoothing passes using running medians of a varying number of consecutive values and a running weighted average as well as adding the residuals smoothed in the same way to the initial smooth (for details see eg. Velleman and Hoaglin 1981, 171—177).

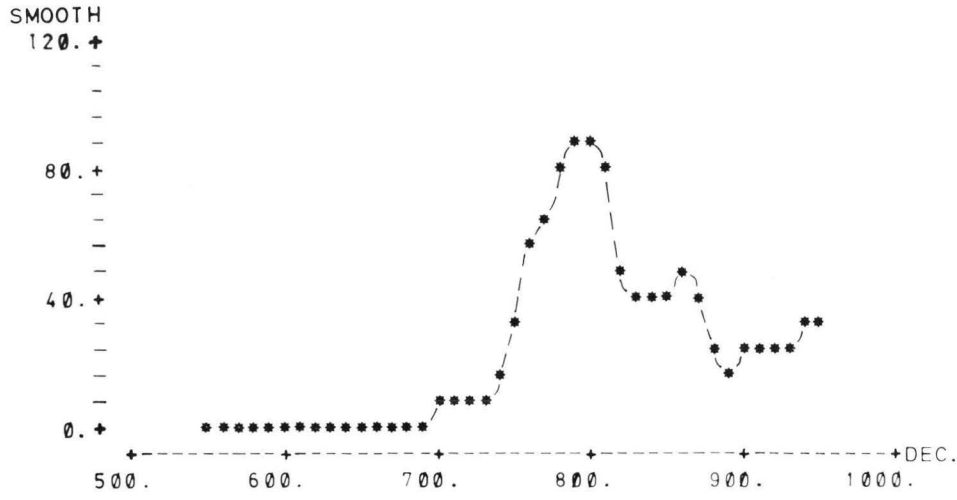

Fig. 7. The smooth of the six hoards plotted vs. decade (see text).

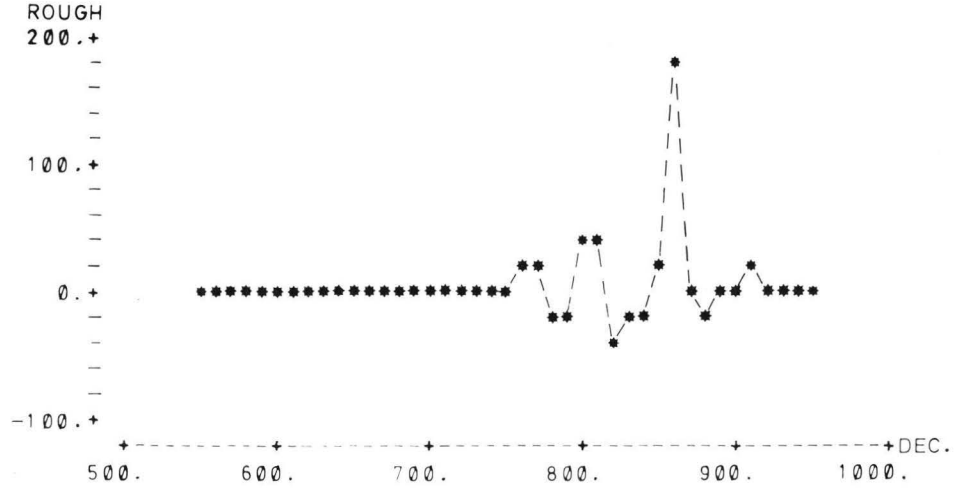

Fig. 8. The rough of the six hoards plotted vs. decade (see text).

Plotting the rough, on the other hand, brings out the details omitted in the smooth, ie. it magnifies the deviations from the latter (Fig. 8). Both should be plotted for a »full» picture of the data.

A few important EDA methods, most of them concerned with data in the form of two-way tables (Tukey 1977, 265—542; Velleman and Hoaglin 1981, chs. 7—8) have to be omitted in this context due to the unsuitability of the example material for these purposes. Moreover, the methodological as well as the theoretical framework of EDA is by no means complete, and new or refined methods are bound to appear in the near future.

By now, I believe, we have learned something about the structure of Viking Period currency in Åland. However, as the main aim of the above is to introduce EDA to archaeologists, the numismatical and historical-archaeological interpretation of the results obtained is best left to another context.

REFERENCES

Blackburn, M.A.S. and Metcalf, D.M. (eds.), 1981. Viking-Age Coinage in the Northern Lands. The Sixth Oxford Symposium on Coinage and Monetary History. Parts i—ii. *BAR International Series 122 (i—ii).*

Bølviken, E., Helskog, E., Helskog, K., Holm-Olsen, J.M., Solheim, L., and Bertelsen, R., 1982. Correspondence analysis: an alternative to principal components. *World Archaeology 14, no.* 1: 41—60.

Granberg, B., 1966. Förteckning över kufiska myntfynd i Finland. *Studia Orientalia edidit Societas Orientalis Fennica XXXIV.*

McGill, R., Tukey J.W. and Larsen, W.A., 1978. Variations of Box Plots. *The American Statistician 32:* 12—16.

Ryan, T.A.Jr., Joiner, B.L. and Ryan, B.F., 1982. *Minitab Reference Manual.* Boston, Mass.

Sarvas, P., 1972. Länsi-Suomen ruumishautojen raha-ajoitukset. *Helsingin yliopiston arkeologian laitos. Moniste n:o 6.*

Tukey, J.W., 1977. *Exploratory Data Analysis.* Addison-Wesley. Reading, Mass.

Velleman, P.F. and Hoaglin, D.C., 1981. *Applications, Basics and Computing of Exploratory Data Analysis.* Duxbury Press. Boston, Mass.