AD-A254 908 OBTAINING MEDIANS AND OTHER QUANTILES WITH LIMITED MEMORY SPACE(U) DEPARTMENT OF DEFENCE CANBERRA (AUSTRALIA) ANALYTICAL STUDIES UNIT B K MCMILLAN

UNCLASSIFIED MAY 92 ASU-TN-4(MEDIANS 1) DODA-AR-007-148 NL

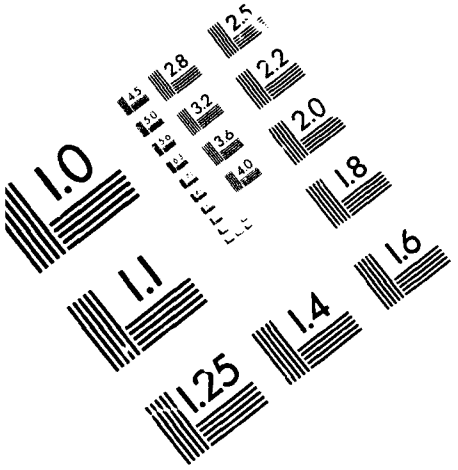
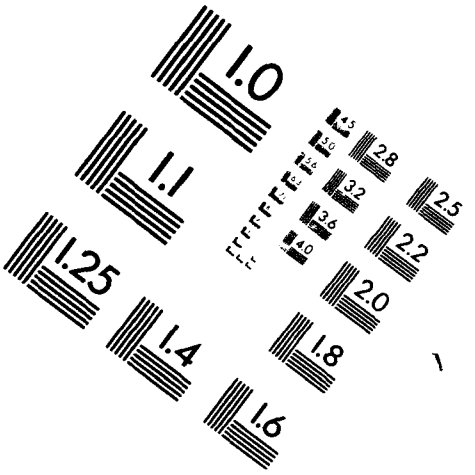
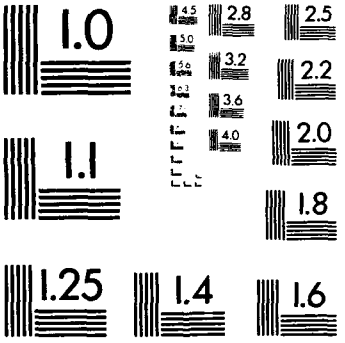
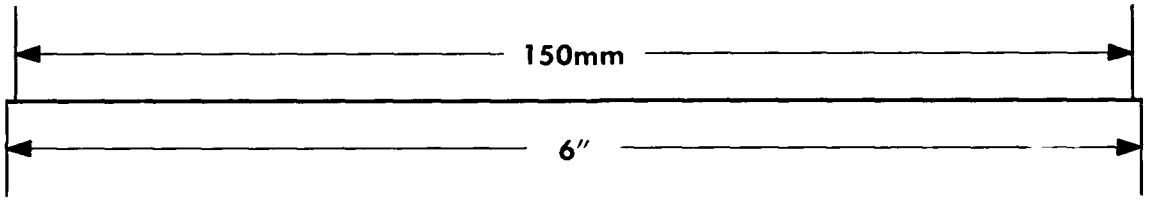
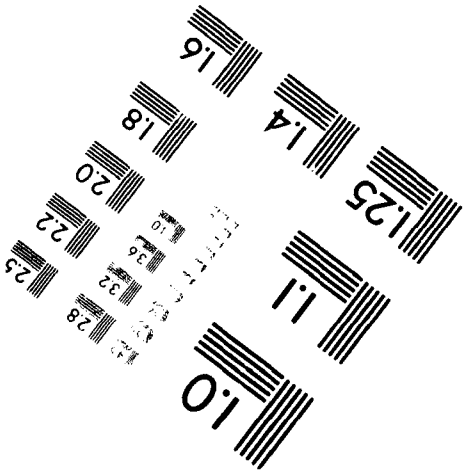


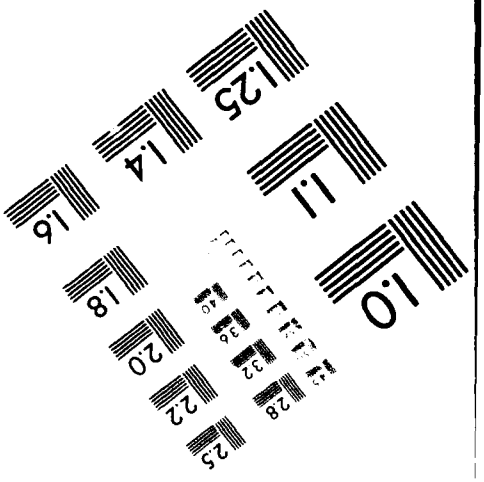
IMAGE EVALUATION TEST TARGET (MT-3)











PHOTOGRAPHIC SCIENCES CORPORATION
770 BASKET ROAD
P.O. BOX 338
WEBSTER, NEW YORK 14580
(716) 265-1600

AR-007-148

(2)

ASU Technical Note 4 (MEDIANS 1)

OBTAINING MEDIANS AND OTHER QUANTILES WITH LIMITED MEMORY SPACE

B.K. McMILLAN

Copy No. 7

May 1992

AR-007-148

ASU Technical Note 4 (MEDIANS 1)

MAY 1992

OBTAINING MEDIANS AND OTHER QUANTILES WITH LIMITED MEMORY SPACE

B.K.McMILLAN

ABSTRACT

This paper provides some basic methodologies for determining medians and other quantiles of data. They are applicable to data more numerous than conventional techniques can manage efficiently. The origins of the work are in ionospheric propagation assessment, but the results are generally applicable. Because these problems push computing facilities to their limit, the final implementation depends on the available facilities. Thus a complete solution cannot be presented, although important factors and techniques have been identified and potential effort savings estimated.

CONTENTS

INTRODUCTION.

Generally medians and quantiles are unpopular statistics because they can be difficult to evaluate in a few lines of coding, and because they are not as tractable mathematically as averages. Nevertheless there are occasions when they are important. In particular the median may be the most appropriate central measure if there is a significant but variable amount of skewness in the distributions underlying sampled data.

This paper looks at some ways to determine medians (the mid point of the data) and other quantiles. Because of the background to the problem, the emphasis is on very large numbers of comparatively small data sets.

It is assumed that computing resources will be an important element of the problem, and as a consequence a detailed methodology is not presented. Instead, a number of different approaches are suggested that may be used in different ways according to the nature of the resource limitations. That is, users must produce their own solutions based on particular problem and computing resource details, but with insight gained from the options presented here.

The paper also indicates that finding quantiles takes much more effort than means. Users may wish to reconsider the use of quantiles in their problem in this light.

BACKGROUND.

In analysis of Over the Horizon Radar (OTHR) capability, ionospheric behaviour patterns were needed. As a consequence, data relating to ionospheric radio wave reflectivity were collected over a period of time for a large number of geographical locations and at a number of radio frequencies. Past analyses have shown that reflectivity varies with a number of different factors including frequency, time of day, time of year and time of sunspot cycle, and that other elements such as clutter, noise etc can be important for OTHR operations.

Ionospheric behaviour is stochastic, with a distribution that is often skew. A convenient central measure of its behaviour for OTHR modelling purposes is its median, rather than the average or mode. Since ionospheric behaviour is an important determinant of OTHR performance, the median is appropriate for related operational elements.

In the OTHR problem there were very many data subsets (each a defined location, frequency, time etc), but not many measurements in a subset (up to 150, each in the range 0-255).

Overall, the quantity of data made existing computing facilities inadequate to allow median determination for all subsets in one reading of the data. Data are still being collected, so future studies may have an even bigger problem. It is this aspect that has lead to the work presented here.

Data points were recorded on fewer than 20 magnetic tapes (allowing only serial access), and in general the data for one subset was spread throughout the tapes. Also, labour and tape drive availability were such that it was important to keep down the number of times the data needed reading.

In summary, the problem was to determine medians of a large number of relatively small data sets, using limited computing facilities.

OVERVIEW.

The general problem of finding quantiles can be categorized by problem size. The definition of size depends on computing facilities available, calculation algorithm requirements and the space needed to store a datum (e.g. integer data require less space than real data). The categories are:

(a) Small. All data may be held in the available storage space. This category is outside the scope of this paper since memory space is not a limitation. Nevertheless, the methods identified here can be used on these problems and could save computing time.

(b) Medium. Some storage is available for all subsets, but in general there is not enough to allow all medians to be found on a single pass of the data.

(c) Large. There are more subsets than storage.

Datum space required may affect categorisation both directly, and in the choice of algorithm. If it is possible to typify observations by bytes or integers rather than reals (e.g. use the exponent rather than the observation), this should be done. Datum space may be subdivided as:

(a) byte;

(b) integer (2 bytes);

(c) real 1 (4 bytes);

(d) real 2 (8 bytes).

The OTHR problem was large with byte sized data. Problems with real data can be harder to handle.

Three generally different approaches are presented. The first has been included for reference purposes, when comparing overall efficiencies. It is the approach that normally would be adopted for small single problems because of its conceptual simplicity. The second may be used when there are few observations per subset - typically under 50. The third is most generally applicable but may present problems in its final stage. The methods are not totally exclusive, and in some ways can be combined. The last two are multi-pass methods that use quantile estimates which are improved with re-readings of the data (passes). Both require an initial estimate, and its production is dealt with separately.

Because the last two methods produce quantile estimates, it may be possible to reduce the total effort if the user can accept an approximation. For example, if the median is to be used simply as a measure of central location, a good estimate should be sufficient.

Some practical experience of the proposed methods was gained using randomly selected data sets, but no memory management techniques were examined.

Efficiency

Efficiency is a term that needs to be defined in relation to a task. In the OTHR problem, there was only one tape drive and one operator, available for some of the time on week days only. As a consequence only a small number of tape changes could be made per day and the total elapsed time became significant, even for calculating means. In commercial computing services, it is common for a charge to be made for each tape change. *In these contexts, algorithm efficiency can be approximated by the number of scans of the data, rather than by the amount of computing.* In other contexts, some other measure may be better.

This definition of efficiency means that memory required by an algorithm for an individual data subset is important. It will dictate the number of data subsets that can be handled in one pass, and thereby affect the number of groups of subsets needed to find quantiles of every subset. Because the amount of storage space available in the computer is comparatively large, that used for the program can be ignored. Occasional comments as to computational efficiency are made, and refer primarily to the standard measures of computing as appropriate to the task.

Compared to calculation of averages, quantile determination is very inefficient:- it requires more complex programming, more memory and may need more passes of the data. To calculate an average, only one pass of the data is needed, along with one counter and one accumulating variable.

TERMINOLOGY, DEFINITIONS AND SYMBOLS.

The following terminology, definitions and symbols are used in this paper:

Median: The value of the midpoint of a data set after it has been sorted into (data) order. For an even number of points in the data it is the average of the middle two.

Quantile: The value of a point (or the average of the two nearest points) at a given fraction of the data set after it has been sorted. The median is the middle (50%) quantile.

Location: The position of a value in the sorted data relative to some point. The median has *location* n/2 relative to the start of the data. If it is not otherwise specified, the *location* is relative to the start.

Cell: A data subset for which a quantile is required.

Pass: A single reading of the whole database to obtain those data for selected cells.

b: The number of bytes of storage needed per cell. The symbol is subscripted with 1, 2 or 3 to indicate method.

k: The number of passes of the data set required to determine the quantile for an individual cell. The symbol is subscripted with 1, 2 or 3 to indicate method.

M: The amount of computer memory available for storage of all intermediate and ongoing values relating to individual data subsets.

N: The number of cells.

n: The number of observations in a cell.

n_a: The average number of observations per cell.

r: The number of individual points from a cell that may be recorded.

S: The number of passes needed to determine all quantiles. The symbol is subscripted with 1, 2 or 3 to indicate method.

t: The number of bytes needed to store one datum. Generally t is 1, 2, 4 or 8.

QUANTILE INITIAL ESTIMATION (FOR METHODS 2 AND 3).

As indicated in the overview, the second and third methods are essentially iterative, and require a starting point. This initial estimate can be produced in a number of ways, depending on the problem size and the amount of data overall. With up to 20 data tapes per tape drive as in the OTHR problem, at least 5% (1/20) of the data should be readily available at any one time, and this can be used to produce a sensible starting value with no loss of efficiency. That is, the first tape(s) can be used to produce initial estimates and can be rewound automatically as required. If there is much less data readily available for an initial estimate, the detailed capabilities and limitations of the problem and computing resources will need to be assessed for the best solution.

The initial quantile estimate can be taken as the quantile of the first few points of the cell data set. Any of the methods described in this paper can be used to determine this initial point, but methods 2 and 3 will still require a starting estimate even for this initial phase. It can be any datum, or possibly the mean or an appropriate quantile of the first few points.

Bounds can be similarly estimated, or perhaps may be chosen as the highest and lowest points of the initial sample. Where the quantile is near one end of the sample range, it may be better to find the standard deviation of the sample and use it to determine values that are unlikely to be violated. For example, the tenth quantile may be needed, and the sample may have 20 points. The initial estimate may then be the second point, and the initial bounds may be that value plus or minus two standard deviations of the 20 points.

The median lies between the mean and the mode, so either of these values can be used to approximate the initial median value if they are already available or are the most convenient to find. In general however, a more accurate initial estimate will require fewer scans to obtain the result.

Other Considerations.

There may be other important factors to be considered which could affect the method used to produce an initial estimate. For example:

(a) If the form underlying the distribution of the data is known, it may enable quantiles to be estimated in some more reliable way.

(b) If it is known that the data are serially correlated then the initial estimate may need to allow for some bias.

(c) Availability of other storage may permit a bigger initial sample to be used without loss of efficiency.

Accuracy and Validity of Initial Median Estimates.

For large n, it has been shown (S.S. Wilks, Mathematical Statistics, 1947) that the median of a sample is an unbiased estimator of the median of the parent population. It has a variance of:

$$1/(f(\text{median})^2.8.n)$$

where f(median) is the probability density value of the median. In a Normal distribution the median has about a 25% greater standard deviation than a similar estimate of the mean. For skew distributions with one mode this percentage may decrease.

Using means estimation as a parallel, it may be inferred that the initial estimate of the median should not be seriously biased, provided that n is not small and that there is no correlation in the sequencing of data points. That is, if the estimate is biased, it is unlikely to be a large fraction of the standard deviation. As an example, the 5% of data suggested above applied to the OTHR problem with about 150 observations represents 7 or 8 points on a distribution ranging from 0 to 255. Assuming a near Normal distribution spanning this space with 6 standard deviations, and that the estimate process is reasonable for this value of n, then the standard deviation of the median estimate would be around 13 or 5% of the total range. It should be within 10% of the actual value with quite high probability.

FIRST METHOD.

Read all the data for a subset, sort it into order and determine a quantile directly (e.g. the median as the middle point, or as the average of the two middle points). The method is illustrated in figure 1. Data for as many cells as possible are read in, and the process is repeated until all cell quantiles have been found.

Storage and output of quantiles are a problem that is specific to the particular configuration of computing facilities so can not be considered in depth, but may be to tape or disk at the end of each pass of the data. In the latter case, available memory may decrease between passes, thereby increasing the total number of passes, but will be ignored.

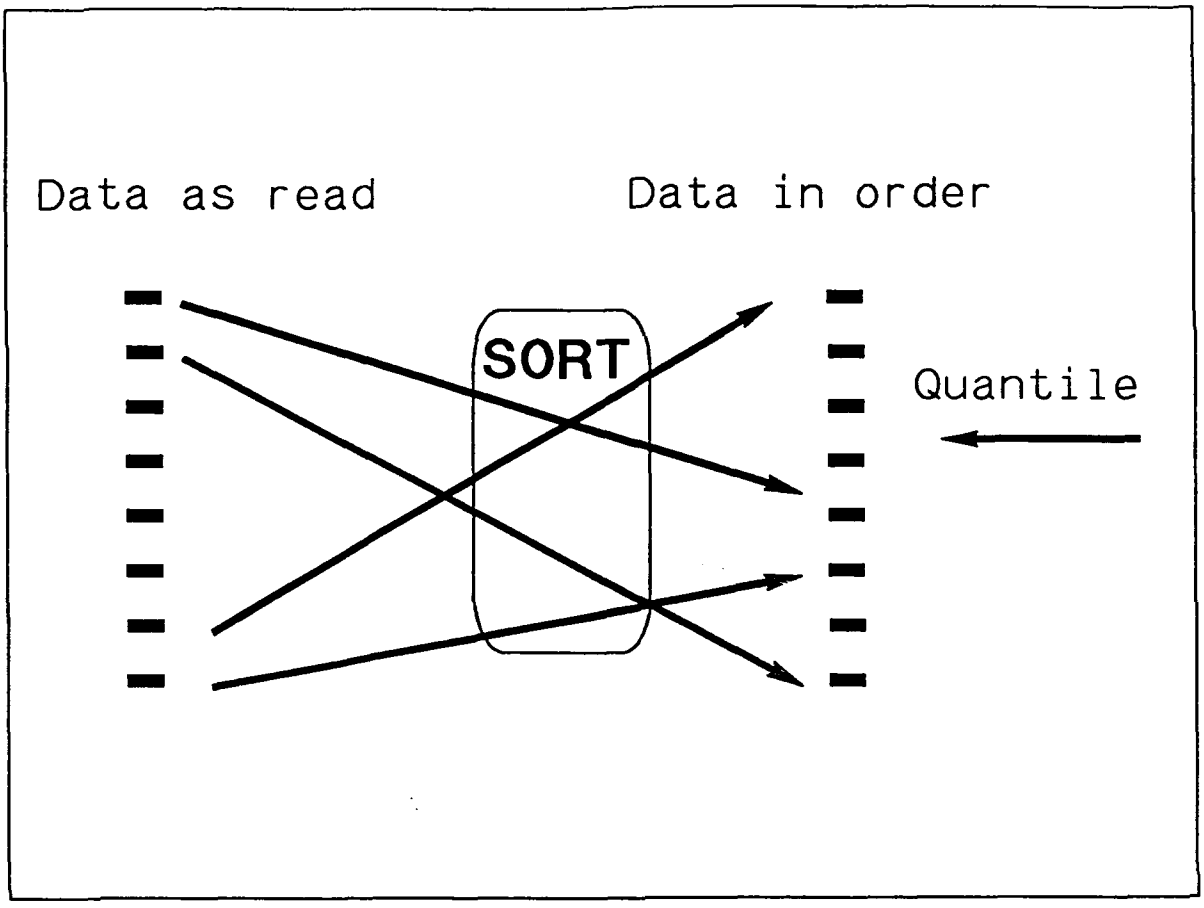


FIGURE 1. METHOD 1.

The efficiency of the algorithm may be assessed as the number of passes of the data set. It is not less than the number of cells multiplied by the number of bytes needed to store the average number of observations plus a counter to give the actual number, all divided by the amount of memory available. In practice it is likely to be greater than this, because of the need for computational efficiency in storing the numbers. Thus allowing one byte for the counter:

$$S_1 >= N(n_a.t + 1)/M$$

Computationally there can be considerable work involved in sorting, usually in proportion to n^{372}.

By way of comparison, the efficiency of mean calculation is approximately t+1 bytes for the sum of all observations in a cell and a counter, giving an efficiency of:

$$N(t+2)/M$$

which is almost n_a times as efficient as S_1.

SECOND METHOD.

This is a multiple pass method. A first estimate of the quantile and the bounds within which it should lie is made. Each pass then records those observations nearest the quantile estimate and at the finish either reduces the range between the bounds or determines the quantile exactly. Figures 2 illustrate the computer memory resources needed for a cell, and give an indication of how they are used as well as the position at the end of a pass.

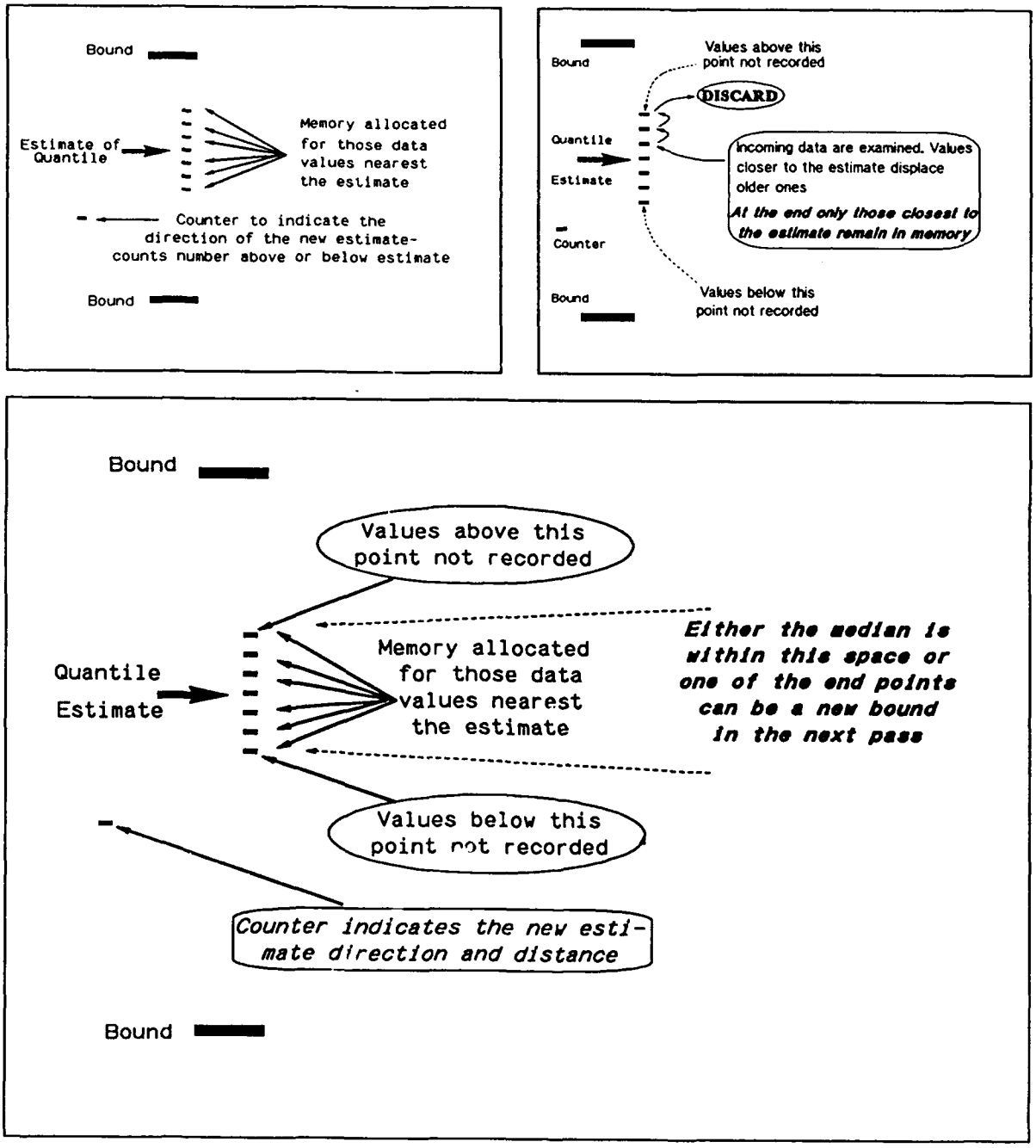


FIGURE 2. A PASS IN METHOD 2.

In general the number of data points between the bounds will be greater than the number that can be recorded, at least until the last pass. During the pass, those values nearest the estimate are stored, older close values being displaced (and lost) as later closer values are input. Also kept is the ongoing quantile *location* relative to the initial estimate, this being simply a count of the number of values greater (or less) than the estimate. The process stops when either the median is found or the bounds are sufficiently narrow for the user.

It is important that the bounds are true bounds, so the initial estimates *locations* in the data must be found in the first pass. This can be done using two of the central range elements as counters, effectively reducing the central range by two in the first pass.

At the end of a pass, the quantile can be determined directly if the *location* identifies an element of the central range, otherwise the bounds can be narrowed by replacing one of them by its furthest extreme in the central range. The new median estimate can be simply the mid point of the bounds, or a better estimate may be found using the interpolation techniques of method three. If the initial estimates of the bounds are wrong, then one new bound needs to be estimated and one old bound can be used (but as the opposite bound). Clearly, the larger the central range the more likely it is that the quantile will be found but the fewer cells that can be dealt with at a time. Also the better the initial estimates the fewer the passes needed.

Efficiency.

Memory required by each cell for this approach is a counter (of the total data set), a pointer, and t bytes each for the bounds and elements of the central range. If one byte is allowed for the pointer and the counter it should allow about 250 observations in a data set or 500 if an extra byte is allowed for the counter and suitable programming practices are followed. Its efficiency is therefore:

$$S_2=(3+(2+r)t)k_2.N/M$$

and the efficiency relative to method 1 is:

$$S_1/S_2=(n_a.t + 1)/(3+(2+r)t)k_2$$

which can be seen to be greater than

$$n_a/(5+r)k_2$$

With n_a at around 150 and $(5+r)k_2$ at about 50 (e.g. a central range of 7 and no more than 4 passes) the advantage is a factor of about three. Large values of t will increase this advantage (at t=3 it rises to nearly four), and the trade off

between r and k_2 may need to be explored for the particular underlying distributions. If n_a is somewhat smaller, r and k_2 should also be smaller. Often, a nine element central range will find a quantile of a 50 point dataset in one pass.

Discussion.

With discrete data, it is probable that there will be several observations of each value near the median estimate. In this case, the number of central numbers held can be increased by maintaining a counter with each central value. Reducing the number of central values held to compensate for these counters will leave the efficiency measures unchanged, but a much larger number of central observations should be maintainable, and this may reduce the number of passes required per cell - thereby improving efficiency, but by an amount determined by the data structure.

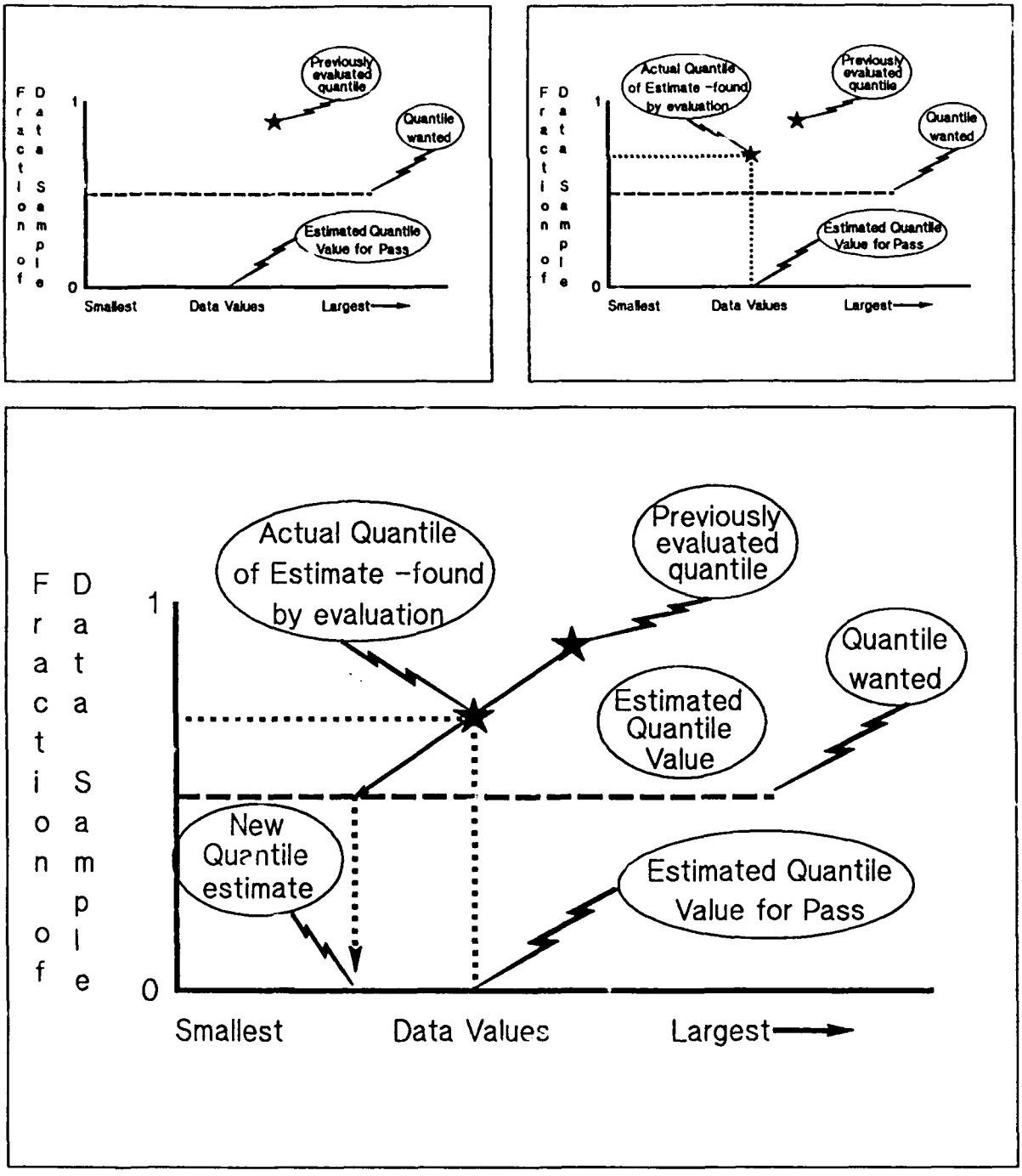
Accelerator methods can also be used that keep more than one median estimate, so that an improved median estimation process can be applied at the end of each pass. This will not be discussed here as there are strong similarities with method 3 below.

Because of the randomness of the processes involved, some cells will need only one pass, and others will require several passes. The mechanics of tracking available memory are beyond the scope of this paper, but unless some form of memory management is applied then the cell requiring the most passes will determine the number needed for all cells involved in a pass.

There is less computing effort required for this process than the first method, but the numbers of decisions and options is quite high.

THIRD METHOD.

This is also a multiple pass method. Estimates of the quantile are made, and each pass finds the actual position of the estimate in the data (its *location*) by counting the number of points that are above it or below it. At the end of a pass, the most recent estimates (of known *location*) are used to produce a new quantile estimate. The method is illustrated in the three parts of figure 3, which show the position at the start of a pass, then the position at the end of a pass and finally the new quantile estimation.



FIGURES 3. A PASS IN METHOD 3.

On the first pass of the data a nearby point must also be *located* so that two known points can be used for the interpolation. The second point can be chosen as another nearby quantile or value (e.g. a fraction of a standard deviation away), and may be selected in the initial estimate process. An important enhancement is detailed later. There is very little computing effort involved in this method.

Because the cumulative count of data points plotted against value is non-decreasing, the process must converge to a narrow interval. Since cumulative counts are generally well behaved, the accuracy of second and subsequent estimates for most distributions should be very good and convergence initially rapid.

When using discrete data, the quantile estimates should always be taken to a whole or half data interval.

Factors Affecting Process Termination.

While producing excellent estimates of the quantile, the approach has problems in producing an exact answer. Because the cumulative count is a step function, then with discrete data there are likely to be large steps in the vicinity of quantiles near the median. If the data are not discrete, changes in the estimate may not be large enough to move past even one datum. Consequently a special terminating process may be needed to obtain the exact quantile. Some of the material in this section should be modified if enhancements to the basic methodology are adopted.

When evaluating estimate *locations* with discrete data, tests for the *location* count should be in opposite directions, and should include those at the value of only one estimate. For example count those observations that are greater than the lower quantile estimate (and lower than the upper one), or else count those that are not greater than the lower quantile. Thus two estimates of the same value will have *locations* that differ by the number of observations of that value. Providing that they span the mid-point, the process is finished. Alternatively two differing estimates having the same *location* should be on the quantile *location*, and so should have a quantile value between them. In practice, there might be no observations at an estimate so an identifier can be used to record their presence. For example the high order bit of the evaluation counter could be set to one when a suitable observation is found, or if practical the sign of the estimate changed.

With continuous data, when an estimate has been found to be reasonably accurate, some of method two can be used. That is, the next pass need only record a few point values nearest that estimate in the direction of the actual quantile in order to complete the process. At least two passes are therefore needed, and more generally three for a cell.

Efficiency.

Memory required per cell for this approach is three counters and t bytes for each quantile estimate (2) apart for the terminating pass when more may be needed. Space for the last pass depends on data type and spread as well as estimate accuracy produced and required. If it is possible to store

the most accurate quantile estimates somewhere, and then use twice as much storage per cell for the terminating pass, this would be equivalent to two normal passes for the final pass.

Absorbing the two terminating passes into k_3 and assuming one byte per counter, efficiency of the method can be estimated as:

$$S_3=(3+2t)k_3.N/M$$

Comparing with method 1 gives:

$$S_1/S_3=(n_a.t+1)/(3+2t)k_3$$

which can be seen to be greater than

$$n_a/5k_3$$

With n_a at around 150 and k_3 at 4 the advantage is a factor of at least seven. With t at 3 the advantage is over 12.

Enhancing the Basic Method.

Improvements in the accuracy of prediction can be made at the cost of an increase in the number of counters. Extra counters will allow the evaluation of points in between the two quantile estimates, and will provide data for higher order curve fitting for the new estimate. In particular, one more counter will allow for a quadratic curve estimator, and with reasonable data and initial estimates could reduce k_3 to 3. This would increase the advantage to nearly 8 when t is 1 and nearly 17 when t is 4.

With large numbers of cells the randomness of the processes involved make it probable that some cells will need additional passes. The excess is likely to be kept small by using quadratic fitting. Adoption of this modification gives:

$$S_3=(4+2t)k_3.N/M$$

$$S_1/S_3=(n_a.t+1)/(4+2t)k_3$$

which can be seen to be greater than

$$n_a/6k_3$$

The evaluation of extra estimates can also be used to reduce the termination problems. For example the outermost estimates can be chosen to estimate points one each side of the central *location*.

Interpolation for New Estimates.

New quantile estimates are calculated from two or three or more old estimates whose *locations* have been found.

Assuming that the old estimates e are *located* at l (subscripted by 1, 2 or 3 etc), and that the new estimate v is required for quantile p, then for a two point interpolation:

$$v=e_1+(e_2-e_1)*(p-l_1)/(l_2-l_1)$$

and for a three point interpolation:

$$v=a+b.p+c.p^2$$

where $c=((e_3-e_2)/(l_3-l_2)-(e_2-e_1)/(l_2-l_1))/(l_3-l_1))$
and $b=((e_3-e_2)/(l_3-l_2)-c.(l_3+l_2))$
and $a=e_2-b.l_2-c.l_2^2$

If any pair of the three points are the same then only a two point interpolation should be used. For higher order interpolation, a higher order polynomial must be found from the old estimates and used to assess v from p.

CONCLUSIONS

A number of conclusions can be drawn. They are:

(a) Compared with calculation of averages, there is much more work in finding quantiles - typically around 10 times more for the dataset sizes used as examples (based on the efficiency measures of this paper). Use of a near estimator or approximating technique may be a consideration. For example using the mean with a multiplier for the median, using a reduced data set, or finding an underlying distribution, estimating its defining parameters for each data set and then determining the quantile.

(b) The specific details of problem and computing facilities have a significant effect on the methodology selected. Properties of disk, tape and other storage should affect the implementation details as should the use to which the results will be put and the type and quantity of data.

(c) The less space needed per datum and quantile estimate, the more efficient the process. If necessary data conversion can be done "on the fly" rather than recasting the original data because compute time is likely to be trivial compared with input effort.

(d) Knowledge of an underlying distribution to the data could be useful, whether or not an alternate suggestion of the first conclusion is used. The knowledge may assist in improving the initial quantile estimate for example.

(e) Memory management may reduce effort substantially if it can be provided using limited memory. For example, at the end of each pass of a tape, store intermediate results, rewind the data tape and begin the pass for a new group of data subsets, without the need for operator involvement.

(f) User requirements should be examined to see if an approximation of the quantile will be sufficient, as this may save considerable work.

(g) The choice of method must ultimately depend on available computing, programming and other resources, and the importance of saving time and effort. Generally method 3 appears to be best in terms of the efficiency measures of this paper because it is likely to use fewer passes overall. However, the problems of the final pass may make it complex and possibly inappropriate. Depending on possible memory management techniques, computing facilities available and data details (e.g. if the average number of points in a cell is small - say under 50), method 2 might be better.

(h) Use of suitable memory management techniques may change the measure of efficiency that should be applied. This in turn may change the relativities of the methods.

DISTRIBUTION

DEFENCE CENTRAL

ASFD	1
DASGFS	2
DASG(HQADF)	3
DOCUMENT EXCHANGE CENTRE	4-11
TECHNICAL REPORTS CENTRE	12

DSTO, SRL, HFRD:

Dr H.GREEN	13
Dr M.GOLLEY	14
Dr B.WARD	15
Dr F.EARLE	16

DSTO, ARL, ASD

Dr P.PRESTON	17
SPARES	18-21

DEPARTMENT OF DEFENCE

Page Classification
UNCLASSIFIED

DOCUMENT CONTROL DATA

1a. A.R. Number	1b. Establishment Number	2. Document Date	3. Task Number
AR-007-148	ASU Tech Note 4	May 1992	

4. Title	5. Security Classification [Use S, C, R, or U]	6. No. Pages
OBTAINING MEDIANS AND OTHER QUANTILES WITH LIMITED MEMORY SPACE	Document U Title U Abstract U	7. No. Refs

8. Author(s)	9. Downgrading/Delimiting Instructions
B.K. McMILLAN	

10. Corporate Author and Address	11. Office/Position responsible for:
Analytical Studies Group (C) Department of Defence Campbell Park Offices CANBERRA ACT 2600	Sponsor Security Downgrading Approval

12. Secondary Distribution (of this document)

13a. This document may be announced in catalogues and awareness services available to

[redacted]

13b. Citation for other purposes (i.e. casual announcement) may be (U) unrestricted or (X) as for 13a
U

14. Descriptors	15. COSATI Group

16. Abstract

This paper provides some basic methodologies for determining medians and other quantiles of data. They are applicable to data more numerous than conventional techniques can manage efficiently. The origins of the work are in ionospheric propagation assessment, but the results are generally applicable. Because these problems push computing facilities to their limit, the final implementation depends on the available facilities. Thus a complete solution cannot be presented, although important factors and techniques have been identified and potential effort savings estimated.