Presented at the SIGGRAPH-ACM Fourth
Annual Conference on Computer Graphics
and Interactive Techniques, San Jose, CA,
July 20 - 22, 1977

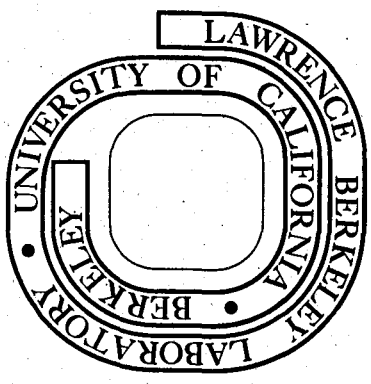
# INTERACTIVE ANALYSIS AND DISPLAY OF TABULAR DATA

William H. Benson and Bernard Kitous

July 1977

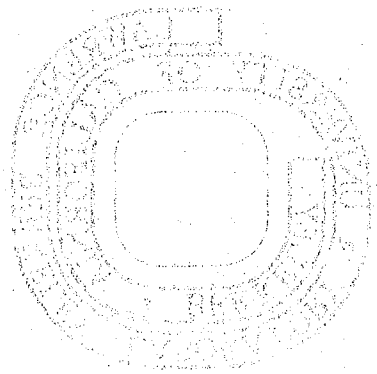
Prepared for the U. S. Energy Research and
Development Administration under Contract W-7405-ENG-48





## LEGAL NOTICE

*This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.*



# INTERACTIVE ANALYSIS AND DISPLAY OF TABULAR DATA

William H. Benson
Bernard Kitous

Computer Science and Applied Mathematics Department
Lawrence Berkeley Laboratory
Berkeley, California 94720

Abstract

A program for simple data analysis and report design is described. The design emphasizes flexibility, ease of use, and rapid interactive response. These considerations are discussed in relation to the choices that may be made in the analysis and display design process. The analysis may be directed and monitored at several points - data selection and calibration, binning, choice of data scaling, choice of graphic variable, and scaling of the graphic variable. Table rows and columns can be re-organized by operators such as ranking, sequencing and grouping, and re-computed from arithmetic combinations of existing rows and columns. Where the raw data represents different cases scored over the same attributes, profile tables can be computed in a systematic fashion. Interactive report design is supported by a variety of page layout and chart annotation directives, which can be used to embellish and adjust the default line, bar and pie charts. The program can be used interactively as well as driven from a prepared script, and uses a device independent graphic system. Although intended to be used primarily with a graphics terminal, at least half the actual use has been conventional report generation on both alphanumeric and graphic terminals.

## Introduction

Most people have trouble assimilating even small amounts of data in tabular format. It is widely recognized that the familiar graphic representations - bar charts, pie charts, line graphs - such as found in newspapers and magazines as well as technical journals are more effective and efficient in conveying relationships in the data.

This paper describes a program (CHART) for simple data analysis and report design for presentation and publication. The basic goal is to represent numerical data tables in a graphic display format so that data analysis and report design can be done in a visual fashion. The program was developed for use in management information reporting. In this environment tables, graphs and charts are typically worked up and drawn by hand, or produced by custom programs with rigid formats designed for a particular report or application. Most of the considerations discussed by Schneider [8] about the social science computing environment hold here. Reports must be clearly labeled and adjustable in format and capable of being displayed in a variety of graphic forms. Typical analysis tasks include grouping and ranking of rows and columns and computing totals, subtotals, averages and percents. A graphic software system for this purpose should be device independent and mask as much as possible technical details about programming languages, operating systems and hardware.

## Design framework

A pragmatic approach has been taken towards meeting these needs.

1. Tabular data is entered either from the terminal or from a previously prepared script. Since there is not as yet a clean interface to a data base management system, it is not possible to describe the data in terms of variables in the data base, such as is done in report generation systems [3, 6]. Consequently the tabular data is not considered to have additional structure beyond that of a rectangular array. Row and column labels are considered to be part of the table and occupy row and column positions.

2. The program is designed for interactive use. Users can experiment with different graphic representations, make adjustment to titles, labels, grids, etc., and perform simple analytic tasks. Since display features are specified independently of each other, modifications to a given display are made by a sequence of incremental changes.

3. The user communicates with the program in a language which resembles a kind of fractured English. A keyword directed syntax allows the user to converse in terse sentences composed from numbers and a small vocabulary of English words. Each sentence changes the state of the data values or the display and gives visual feedback. No graphic input is used. This implies that the program can be directed from a text file as well as from a terminal (including alphanumeric terminals).

4. Any display produced by the program is completely specified by a short list of sentences. The description may be saved on a script file so that the results from one session at the terminal can be the input to the next. The form of the display can be saved independently and used with different data.

5. A device independent graphics package is used. Any picture produced should look as similar as possible on different devices. Hardware characters are used to reduce plotting time. Since hardware characters vary in size from one device to another, plotting space for text is allocated in terms of character size.

6. Choices may be made throughout the analysis and design process. The analysis may be directed and monitored at many different points - data selection and calibration, binning, scaling statistical data to display data, choice of graphic variable and chart type, and scaling of the graphic variable. Rows and columns of the table can be re-organized by operators such as ranking, sequencing and grouping, and re-computed from arithmetic combinations of existing rows and columns. Titles and row and column labels may be entered and positioned, bars and pie slices shaded and labeled with data values, axes labeled, etc. There are format options for missing data and hierarchical labels.

7. A transcript of each session, minus the graphic output, is printed and routed to the system designer. This provides valuable information about patterns of usage, difficulties with the program, proposals for new features, and bugs. User comments entered at the terminal are seen in context. For the user, the last thirty commands are saved and may be displayed in a brief history log. In addition to on-line access to part of the users' manual, users remote from the computer center can be helped by expanding the interaction to include on-line monitoring and consulting.

Chart types

The graphic forms which can be produced include matrix displays, bar charts, line graphs, pie charts, and tabular reports.

A matrix display is the direct representation of a numerical data table in graphic form. That is, the display looks like a matrix or table with rows and columns but where the numbers in each table cell have been mapped into a graphic variable. The graphic variable used is the variation of size (of a circle, angle, or horizontal or vertical bar). Bertin [2] has shown that any graphic variable (such as size, value shading, shape, color, orientation or texture) can represent categorical data and a few (texture, shading and size) can convey relations of order between items. However only variations of size can adequately express quantitative relations (numerical ratios) between data items. Ratios could be estimated using value shading or texture although with much less spontaneity and precision.

Each table cell is replaced by a graphic item such as a circle or bar. The numerical data is mapped into graphic items according to specific rules of scaling which can be modified by the user. The arithmetic sign is always preserved. Negative values are shown both by shading bars and circles, and extending bars in the opposite direction from positive bars. Since matrix displays preserve the positional characteristics of a table, the user can return to the more familiar tabular format to re-interpret patterns observed in the display. To avoid clutter and improve perception of patterns in the data, labeled axes are normally omitted. When there is one or only a few rows or columns, a bar chart can support labeled axes, shaded bars, and values attached to the bars. Line graphs where curves from each row of the table are superposed and plotted to the same scale permit ready comparison of horizontal profiles. Pie charts show the division of a whole into parts. Conventional tabular reports have the advantage of familiarity and preciseness, and can be displayed on alphanumeric terminals lacking graphic capablilities.

## The interactive cycle

An interactive session at the terminal is seen as a cyclical process consisting of data selection and calibration, analysis, and display.

### Data selection

Data selection begins with input of a raw data table. Tables are prepared in a stylized tabular format similar to that used by SPSS [7], but more suitable for entry at the terminal. Column headers are entered first, followed by successive rows consisting of a label and a number in free field format for each column. A table may be entered at the terminal or from a previously prepared script. Once established, a table can be partitioned and attention restricted to a subset. A set or range of rows and columns may be masked from sight and later restored in whole or in part. Individual table cells can be changed - for example erroneous data or outliers - and new rows and columns can be computed from arithmetic combinations of old ones. Since the raw data may need to be augmented or partially replaced in unpredictable ways, an interpreter for vector arithmetic computes arbitrary arithmetic expressions of row or column vectors element by element. For instance population densities could be calculated from a table of populations and areas for a set of regions. An auxiliary command backs up to the last table as it was originally entered.

Analysis

It is expected that one looks at raw data tables, time series data aside, primarily for size effects. For example table cells may be compared according to which is the largest, or rows put in rank order on the basis of a particular column.

In many cases it may be inappropriate to compare the data on the basis of size, for example when the raw data is not homogeneous or when shape rather than size effects are of interest. Suppose several cities are to be compared on a cost of living basis. Table 1 shows consumer price indexes for major categories such as food, housing, etc. Figure 1A gives a global idea of where costs are greatest. Fuel is high everywhere but all other costs are fairly similar. Here it would be useful to compare the relative deviations from the average for each category rather than comparing absolute index values. The relative deviations describe vertical profiles for each price index category. Figure 1B shows how much each city differs from the average. Negative differences are shown by shaded bars. Seattle and Honolulu are consistently below average; Washington D.C. and New York are above average in each column. The rest are mixed.

Frequently, a table represents a set of different entities scored over the same attributes. Consider a cross tabulation of age distribution by county. Although counties

may differ widely in population, they may be compared directly if the raw counts for each age group within a county are replaced by the percent each count is of the total county population. Similarly age groups could be compared across counties by taking percent of total in the other direction.

The two operations illustrated, difference and proportion, can be combined. For example economic indicators are often described by percent change from previous month. The differences between consumer prices in successive months are divided by the earlier of each pair to compute rate of change profiles.

Since obtaining profiles is an analytic step of general utility, profile tables can be derived from the raw data in a systematic fashion. There are two steps to the normalization process in which size effects cancel out to allow comparisons across attributes.

1) define a reference row or column
2) compare each row or column in the table to the reference

Several choices are available for each step. The reference may be computed by summing or averaging over rows or columns. A particular row or column in the table may be designated, or a new one, such as a threshold, entered at the terminal. Or instead, for each row or column the previous one may be specified as the reference. Comparisons with the reference may be made either by taking differences, proportions, or first differences and then proportions. Typically each row, say, of the profile table is obtained from the corresponding raw data row by subtracting the row average or dividing by the row sum element by element.

Several operators which manipulate entire rows and columns at a time can be used to re-organize a table. These include

1) ranking - the rows or columns can be ranked in ascending or descending order on the basis of a particular column or row.

2) permutation - a particular sequence of rows and columns can be specified; pairs of rows or columns can be switched.

3) grouping - the table can be partitioned into blocks by declaring groups of consecutive rows and columns. When plotted as a matrix display, bar chart, or tabular report the groups are shown separated by blank rows or columns. A

reference row or column, for example, is shown in a group by itself set off from the rest of the table. In Figure 1A the row average is shown as a reference. Row and column categories are immediately recognized since the groups are perceived spontaneously as separate entities. Figure 1B shows cities grouped according to whether they are above average, below average, or mixed with regard to cost of living indicators.

These operators are applied simultaneously to the profile table and the raw data table from which profiles have been derived. Thus a meaningful organization of a profile table can be interpreted back to the raw data, and vice versa.

Display

The data structure consists of a set of rectangular arrays, one each for the profile, raw, and display data tables. The numerical data in the raw or profile tables is mapped into display data normalized between 0 and 1 in absolute value. These normalized values can be expressed directly by graphic items. The mapping is performed in several steps. At each step the user may intervene to choose from among a limited set of alternatives:

1) bin the data - this is a preliminary step to scaling. The data is distributed as equally as possible into however many bins are wanted, from one to the number of distinct data values. The distribution is adjusted so that tied values are always in the same bin. The average value is taken within each bin. This step, which reduces data resolution, is optional.

2) scale the data - the numerical data values are scaled prior to being plotted by one of several transformations described below. All of the transformations preserve the sign and relative order of the data values.

Ranking preserves only the order of the data values. These are scaled in equal steps from the minimum to the maximum. The minimum takes the smallest graphic item and the maximum the largest, so that the entire range of the graphic variable is used. The intermediate values are equally distributed within this range in rank order. Ranking allows comparison of all data values even though disproportionately large values may be present.

Relative scaling preserves intervals between values as well as order. Ratios are in general not preserved, since the zero value may not lie within the data range. This

transformation may be misleading, but can be useful when extreme values are present or when zero would be an extreme value. The minimum is shown by the smallest possible graphic item and the maximum by the largest. Relative scaling uses the entire range of the graphic variable. However when the data range already includes zero, absolute scaling is used instead. Relative scaling is used in Figure 1A. The largest values are recognized spontaneously.

Absolute scaling extends the data range, if necessary, to include zero. This has the effect of preserving ratios between graphic items, but may use only a small part of the space available in each matrix cell to show the data variation. That is, the lower part of the range of the graphic variable may be unused.

Adjustments to the last two rules are usually made when a labeled scale is displayed on the chart. For absolute scaling the range is further extended to reach "pleasing round numbers". Since users frequently want to specify minimum, maximum and division points on a labeled scale, these parameters for relative scaling can be set directly.

3) select graphic item - horizontal bars, vertical bars and circles translate numerical data by size. Profiles along rows are typically shown by vertical bars; along columns by horizontal bars. The circle display is unbiased as to direction so that extreme data values are spontaneously perceived regardless of where they are in the table. Line graphs, pie charts, and tabular reports are also specified at this point. If no graphic item has been selected the normal graphic feedback from all but a few of the other commands is suppressed.

4) scale graphic variable - a further transformation can be optionally applied to the normalized display data to scale the graphic variable. Three functions are provided to enhance the low, middle, or high ranges of the data values. [5]

a) exponential base 10 (Fechner's rule) to enhance the high end of the distribution
b) square root - to enhance the low end
c) a composite of the first two rules to enhance the middle range

The interval [0,1] is mapped onto itself. Both sign and order are preserved. These functions can be used to match perceived variation in the graphic variable with the actual variation in the data, or to distort the graphic variable so

as to enhance discrimination in the region of interest while retaining a global picture of the data.

Interactive report design

An auxiliary program is available to interface with a data retreival system. This program processes a script identical in form to that used by CHART, replacing data descriptors by data elements but leaving all else unchanged. Since the form of a display can be saved (as command sentences) for use with new data, report design needs to be done only once for a set of similar data. In practice reports are usually produced in several steps - (1) design the report using typical data; (2) save the form of the display on a script; (3) replace the table values by descriptors in the data base using a text editor; (4) run the auxiliary program to rewrite the script, substituting values from the data base for the data descriptors; and (5) run CHART using the rewritten script.

The user has a choice of several standard graphic forms to represent a table and can experiment with different formats within each form. Since the program is oriented towards interactive use, it is expected that the report design process proceeds from a default standard display to the final design by a series of incremental changes. The display is structured accordingly so that the visual components may be modified independently of each other. The display is constructed from the outside in with titles surrounding row and column labels surrounding labeled scales. Individual title lines can be entered and positioned selectively at the four outside edges of the display. Labels and labeled scales can be placed at the left, right, top, or bottom. The remaining space is used to display the data. Space for the titles, labels, and scales is allocated in terms of character size. This helps achieve device independence and recognizes the importance of text to the chart reader. Some other features that can be modified are grids, width and shading of bars, numerical formats, and placement of labels on line graphs and pie charts.

Tabular data in the real world is often incomplete - parts of rows and columns may be empty because no data for these positions is available. The approach taken here is to key these positions to explanatory messages within the program and to display one or more characters from the message at the corresponding matrix cell. Asterisks, for example, could be displayed at empty cells and referred to a footnote, starting with an asterisk, which identifies or explains the missing data. The messages are implemented by

assigning title lines to missing data. Empty table cells are filled from a list of distinguished numbers, too large to be confused with real data, where each number on the list corresponds to a particular title line. Numbers of this magnitude are uniformly treated as missing data and ignored in scaling and all other computations. These huge numbers may be entered by the user, or computed by the program. For example the calculation row3=row1/row2 will result in no data for those positions where row2 has zeroes.

Device independence

The graphic systems environment consists of a variety of interactive terminals and off-line hard copy devices and a set of FORTRAN subroutine packages, one for each device type, which supports device independent displays. The subroutine packages, collectively known as GRAFPAC[4], are low level drivers supporting standard graphic functions. These include plotting points, lines and characters, erase screen, and defining windows and viewports. These functions are described by identical parameters and calling sequences among the GRAFPAC modules.

The module for each device interprets only the parameters supported for that particular device. For example, modules for devices with only one character size will ignore that specification. Each device is given a normalized plotting space from [0,1] in x and y so that pictures drawn within these limits are guaranteed to plot on any device. The applications program specifies a linear mapping from a data space to the normalized page space. A further mapping is made within each module to the raster space used by the particular device. Characters and lines may be clipped to appear only within the viewport specified in the normalized page space coordinates.

The system is efficient for interactive use since the display list (including hardware characters) for each individual device type is generated directly. The device types required must be selected before program execution but only the corresponding modules need be loaded. There are mechanisms for switching between device types and between hardware and software characters during execution.

An alternative approach to device independence is also supported. An intermediate file generated by a module for a virtual device can be interpreted by a post processor which has been linked with any of the other modules. Since pictures from the intermediate file can be plotted and overlaid in any order, page layout and composition could be

approached in this way.

No graphic input is used, avoiding the issue of device independence.

Discussion

The first design underestimated the role of text in favor of providing a variety of graphic forms. Since novel graphic forms do not seem to be readily accepted, the most familiar graphic representations were emphasized. Experience with users indicates several ways in which textual material is important.

The data must be clearly identified. In a practical case employment data, which can be qualified with many conditions, often requires many words in each row and column label in addition to titles identifying the overall subject. Additional titles may be needed to guide interpretion of the graphic display. Use of hierarchical labels has helped reduce the quantity of text.

Although graphic displays are effective and efficient in conveying relationships in the data, it still seems essential to be able to refer back to the numbers. Users may feel the need to validate or check the consistency of visual impressions, be able to verbalize results, or feel more secure with the precision of numerical data. Consequently attention is also given to the display of numbers as text. In particular, since matrix displays are a direct representation of a numerical table, one can be readily compared with the other.

Although the program is intended to be used primarily with a graphics terminal, at least half the actual use has been conventional report generation on both alphanumeric and graphic terminals. In part this reflects greater familiarity with tabular reports and limited availability of graphic terminals. Here device independence has been especially useful. Since tabular reports are plotted rather than printed, they may be displayed on an alphanumeric terminal by using the corresponding GRAFPAC module.

Future developments will emphasize integration into a larger information system [1].

References

[1] Austin,D.M., Kranz,S.G., and Quong,C. An overview of the LBL socio economic environmental demographic information system. LBL-3699. March, 1975.

[2] Bertin,J. Semiologie Graphique. Gauthier-Villars, Paris, (1967),69.
[3] Gibson,T.A., and Ting,P.D. A new report generator. Proc. ACM-PACIFIC-75, San Francisco, Ca., (April 1975),119-126.
[4] GRAFPAC Users Guide, Graphics Research Group, Lawrence Berkeley Laboratory, July 1976.
[5] Kitous,B. Interactive matrix displays and management information reporting - a feasibility assessment. (Ph.D. thesis) LBL-5310. (June, 1976), 118-120.
[6] Mendelssohn, Rudolph C. The development and uses of table producing language. U.S. Department of Labor, Bureau of Labor Statistics, Report 435, 1975.
[7] Nie,Norman, et al. Statistical Package for the Social Sciences , 2nd edition, McGraw-Hill, N.Y., 1975.
[8] Schneider, Edward J., Barge, Sylvia, and Marks, Gregory A. Graphics for social scientists. Computer Graphics Vol. 10 No. 2 (Summer 1976), 125-131.

This work was done with support from the U.S. Energy Research and Development Administration and the Department of Labor.

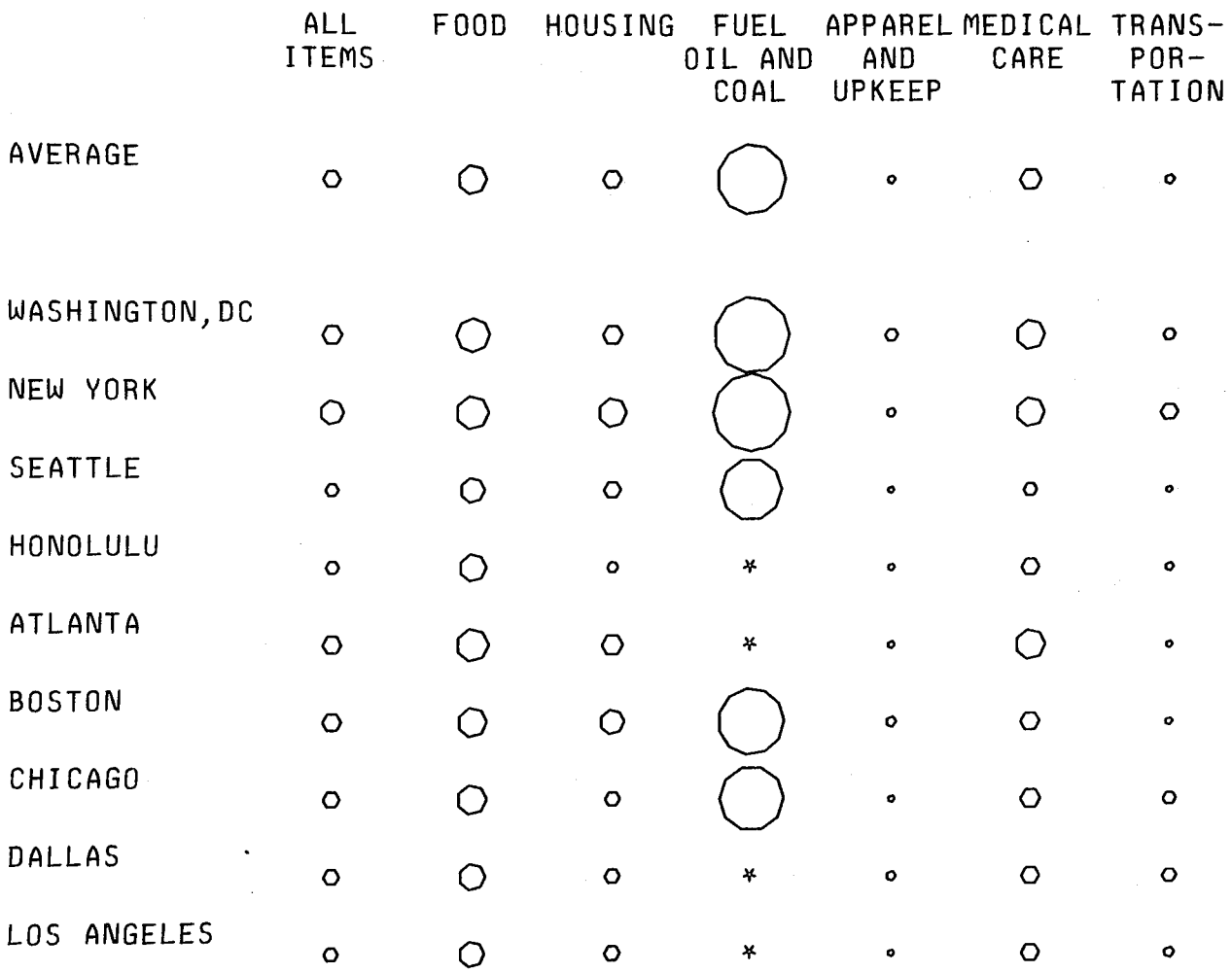
CONSUMER PRICE INDEXES FOR SELECTED METROPOLITAN AREAS
1974 AVERAGES (1967=100)
(SOURCE - U.S. STATISTICAL ABSTRACTS)

| | ALL ITEMS | FOOD | HOUSING | FUEL OIL AND COAL | APPAREL AND UPKEEP | MEDICAL CARE | TRANS-POR-TATION |
|---|---|---|---|---|---|---|---|
| WASHINGTON, DC | 150 | 167 | 150 | 220 | 141 | 161 | 139 |
| NEW YORK | 155 | 166 | 161 | 222 | 136 | 161 | 146 |
| SEATTLE | 141 | 156 | 146 | 202 | 131 | 142 | 125 |
| HONOLULU | 142 | 159 | 139 | * | 133 | 147 | 135 |
| ATLANTA | 148 | 166 | 152 | * | 134 | 162 | 132 |
| BOSTON | 149 | 161 | 155 | 208 | 138 | 149 | 134 |
| CHICAGO | 146 | 162 | 144 | 206 | 133 | 150 | 141 |
| DALLAS | 145 | 158 | 144 | * | 137 | 148 | 142 |
| LOS ANGELES | 142 | 156 | 144 | * | 132 | 148 | 138 |

* DATA NOT AVAILABLE

TABLE 1

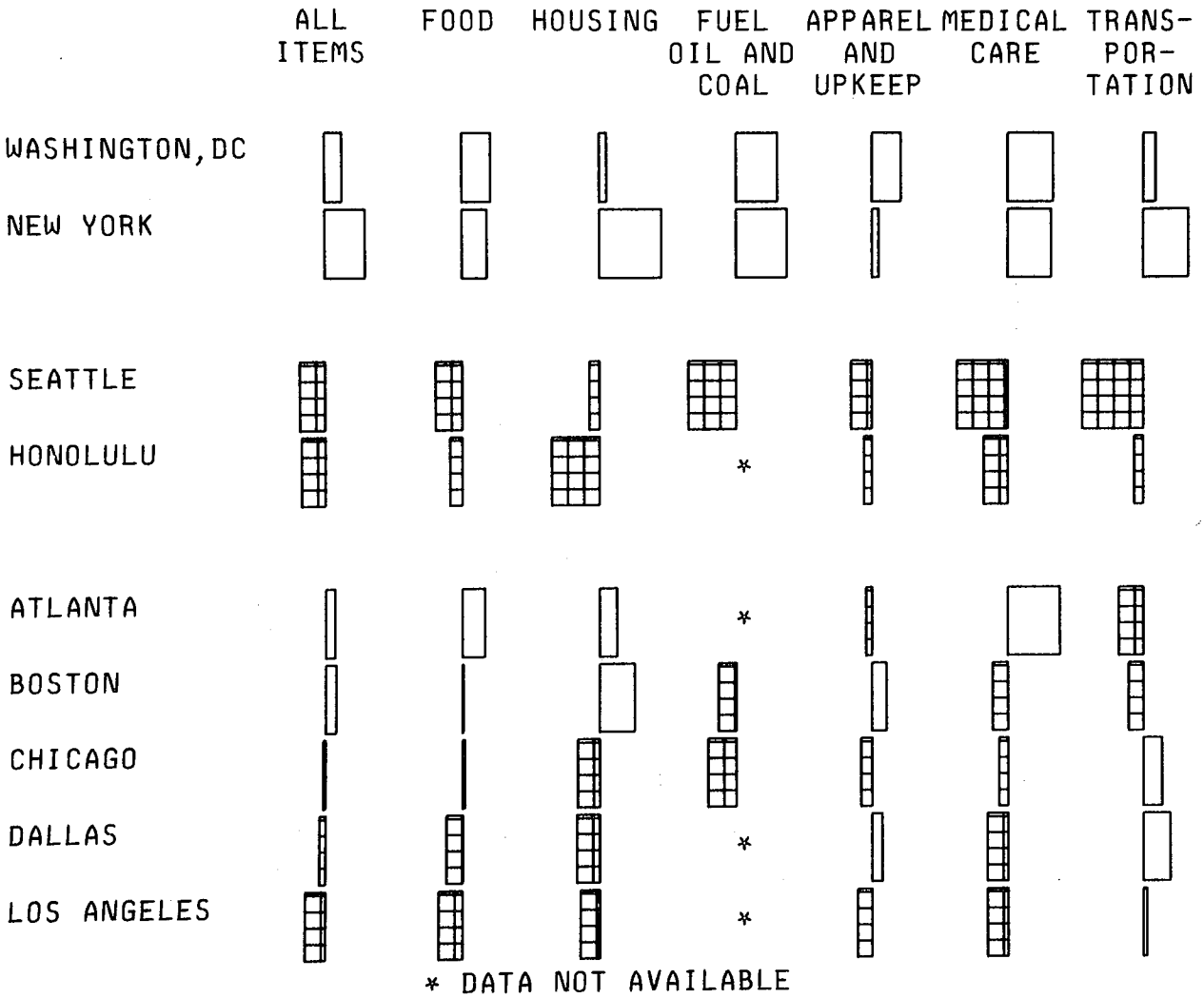
MATRIX DISPLAY REPRESENTATION OF TABLE 1
INCLUDING A HYPOTHETICAL AVERAGE CITY
THE SIZES OF THE DOTS ARE PROPORTIONAL TO THE DATA
THE LARGEST DOTS STAND OUT




* DATA NOT AVAILABLE

FIGURE 1A

VERTICAL PROFILES SHOW DIFFERENCES FROM THE AVERAGE
(SHADED BARS INDICATE NEGATIVE VALUES)
CITIES HAVE BEEN GROUPED ACCORDING TO
ABOVE AVERAGE, BELOW AVERAGE, OR MIXED PRICE INDEXES




* DATA NOT AVAILABLE

FIGURE 1B

This report was done with support from the United States Energy Research and Development Administration. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the United States Energy Research and Development Administration.