Water Research Vol. 16, pp. 31 to 36, 1982
Printed in Great Britain

0043-1354/82/010031-06$03.00
Pergamon Press Ltd

AN INTERACTIVE DATA MANAGEMENT SYSTEM FOR RIVER WATER QUALITY DATA

ROLF A. DEININGER
School of Public Health, The University of Michigan, Ann Arbor,
MI 48109, U.S.A.

(*Received March* 1981)

Abstract—Water quality data banks are needed to document the status and trends of water pollution in a country. Examples of such systems are the STORET system in the U.S., the NAQUADAT system in Canada and the EIS system in Scandinavia. All of these systems require trained personnel to help in the formulation of the inquiry and the actual querying of the system.

By contrast, what is described in this paper is an on-line, interactive data management and analysis system which allows the user the direct search, update, retrieval and analysis of the data from a computer terminal. The user addresses the system in a high level language closely resembling English and has complete control over building, updating and querying the individual data banks. Almost all statistical operations can be performed on the data starting from histograms, distributions, correlations to regression, discriminant, component and spectral analysis. Commands for producing camera-ready graphs on graphic terminals are available.

The system is implemented on The University of Michigan Computer System and can be accessed through local telephone numbers in more than 100 cities in the U.S. and Canada and from the major European capitals via the TELENET system. The operation of the system is illustrated on a small sample data base on the Ohio river provided by the Ohio River Sanitation Commission (ORSANCO).

INTRODUCTION

Regional water quality data banks are of great importance if one wishes to determine trends in water pollution over time. The general rationale for creating such systems is that data on water quality are generated by many agencies, municipalities and individual researchers, but are sometimes difficult to find. The basic idea is therefore to store them at one central location and to share them with all legitimate users. Central data banks ensure that a user will have access to all data that have been collected, that the cost of coding the data is incurred only once and that the reproduction and retrieval of the data is rather inexpensive.

Examples of such systems are the STORET system in the U.S., the NAQUADAT system in Canada and the EIS system in Scandinavia. In all of these systems the user submits a request for data on specific forms and receives within a hopefully short period a printout of the data requested. Specially trained personnel is necessary to query the system and help in the formulation of the inquiry.

In contrast to the above, what is described here is an on-line, interactive system which allows a user the direct search, retrieval and analysis of water quality data from a computer terminal in his own office.

THE COMPUTER SYSTEM

The computer system used in this study is The University of Michigan central computing facility which consists of an AMDAHL 470/V8 with a central memory of 16 Megabytes (16×10^6) and on-line disk storage of 8 Gigabytes (8×10^9). It is a time sharing system and allows simultaneous access to about 500 users which can access the computer system over the telephone network from a variety of terminals and/or the central stations on campus. Connections of The U of M system to TELENET allow access through local telephone numbers in about 100 cities in the U.S. and Canada and from the major capitals of Europe. Extensive software exists which has been used in this study and the system has been put together using various existing components. Most of the software is specific to the local system and can not be easily implemented on other computer systems, but similar software packages exist at other computer installations. Since the system is dedicated to research purposes, it is not generally available for commercial use.

THE SYSTEM

Rather than describing the system in abstract terms, it is best to illustrate the operation of such a system by way of examples. In the examples following, actual conversations with a small demonstration data base of water quality data on the Ohio River are shown. This small data base has values on 9 selected parameters of water quality from 35 monitoring stations in the Ohio River Basin and was obtained from the Ohio River Basin Sanitation Commission (ORSANCO). Figure 1 is a sketch of the location of

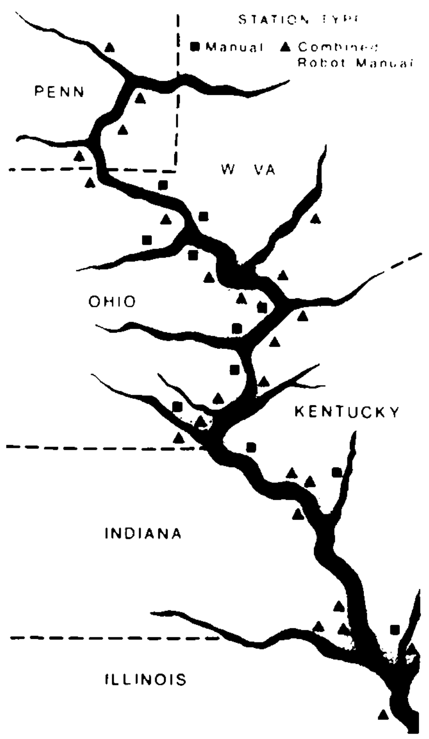


Fig. 1

the monitoring stations in the six states. The examples following show the three major functions of the system—the search and retrieval of data, the analysis of data, and finally the graphing of data.

Search and retrieval of data

The data storage and retrieval part of the system uses a data base management system called MICRO(1) which allows a user to direct inquiries in a language closely resembling English. It is very user oriented, friendly, forgiving and allows a user to recover from the mistakes. In the examples following, the user queries are printed in lower case letters and the system response is printed in capital letters.

```
*READY:
-print in orsanco all field names
*STATION              STA
*DAY                  DAY
*MONTH                MON
*YEAR                 YR
*FECAL.COLIFORMS      COLI
*PH                   PH
*BOD                  BOD
*NITRATE              NO3
*PHOSPHATE            PO4
*TEMPERATURE          TEMP
*SUSPENDED.SOLIDS     SSOL
*TOTAL.SOLIDS         TSOL
*DISS.OXYGEN          DO
```

Example 1

```
*READY:
-print in orsanco all categories of station
*S.HEIGHTS        S.HI
*ADDISON          ADD
*WINFIELD         WIN
*GALLIPOLIS       GALL
*HUNTINGTON       HUNT
*KENOVA           KENO
*S.PITTSBURG      PITT
*OAKMONT          OAK
*BEAVER.FLLS      BEAV
*LIVERPOOL        LIV
*PIKE.IS          PIKE
*SHADYSIDE        SHAD
*WILLOW.IS        WILL
*MUSKINGUM        MUSK
```

Example 2

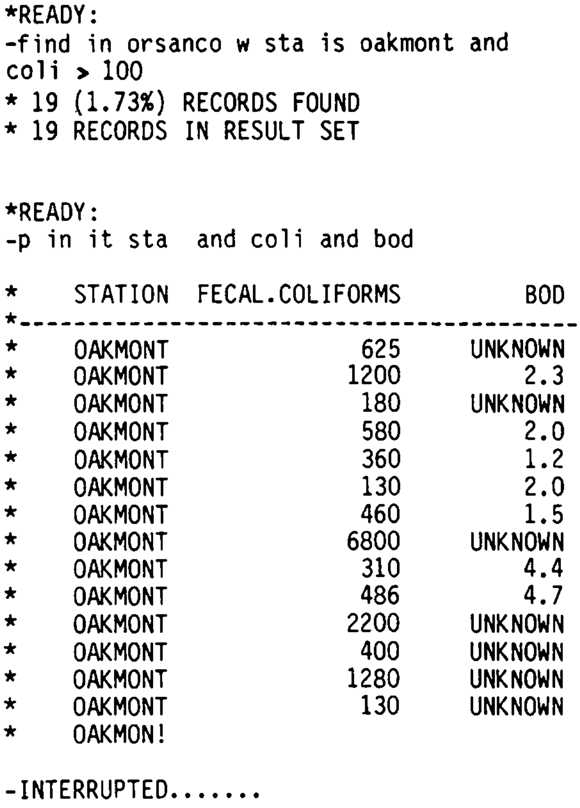
To search for data and retrieve them, the user should not have to remember specific codes but should be prompted by the system. After signing on the user may simply issue a command to show the names and abbreviations of all parameters. Example 1 shows the nine parameters of water quality, the station where the sample was taken, and the date of the sampling.

If the user does not remember the names of the monitoring stations, the command above (Example 2) will display the names and abbreviations of the stations. The printing of all the names was interrupted to save space and not print all 35 lines.

If one is looking for, say, BOD values, and particularly those which are larger than 5 mg l^{-1}, the first command below (Example 3) extracts from the data

```
*READY:
-find in orsanco where bod is greater
than 5
* 33 (3.00%) RECORDS FOUND
* 33 RECORDS IN RESULT SET

*READY:
-print in it station and bod and do

*    STATION         BOD  DISS.OXYGEN
*------------------------------------
*  S.HEIGHTS         5.4         12.4
*  S.HEIGHTS         7.5         10.1
*S.PITTSBURG         5.7      UNKNOWN
*S.PITTSBURG         6.5         12.0
*BEAVER.FLLS         5.5          4.2
*BEAVER.FLLS         5.6          6.9
*BEAVER.FLLS         6.8         13.7
*BEAVER.FLLS         7.7          9.4
*BEAVER.FLLS         7.1         11.1
*    PIKE.IS         6.5          8.5
*!

*INTERRUPTED......
```

Example 3



```
*READY:
-find in orsanco w sta is oakmont and
coli > 100
* 19 (1.73%) RECORDS FOUND
* 19 RECORDS IN RESULT SET


*READY:
-p in it sta  and coli and bod

*    STATION  FECAL.COLIFORMS          BOD
*---------------------------------------------
*    OAKMONT               625     UNKNOWN
*    OAKMONT              1200         2.3
*    OAKMONT               180     UNKNOWN
*    OAKMONT               580         2.0
*    OAKMONT               360         1.2
*    OAKMONT               130         2.0
*    OAKMONT               460         1.5
*    OAKMONT              6800     UNKNOWN
*    OAKMONT               310         4.4
*    OAKMONT               486         4.7
*    OAKMONT              2200     UNKNOWN
*    OAKMONT               400     UNKNOWN
*    OAKMONT              1280     UNKNOWN
*    OAKMONT               130     UNKNOWN
*    OAKMON!

-INTERRUPTED.......
```

Example 4

base those records which show a BOD of more than 5 (a total of 33 records) and the second command prints the actual values for the station, the BOD and the dissolved oxygen. Most of the commands may be abbreviated. For example, W stands for where and P for print.

Example 4 shows how one can search for data at a specific station. Suppose the user is interested in fecal coliform counts of greater than 100 at the station Oakmont, the two commands below extract and print the requested data. Again, only a partial listing of the 19 records is shown.

The command XTAB (Example 5) produces a cross tabulation of the data on the parameter specified, in this case the month. A simple command like this shows that the samples taken are rather unevenly distributed over the year. The difference between March and October are rather prominent.

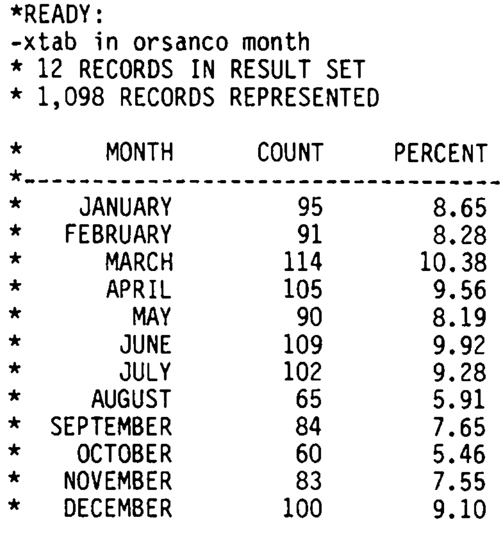
New variables can be computed rather easily using the COMPUTE command (Example 6). Suppose it is known that the relation between turbidity and suspended solids is known to be: TURB = 10 + 0.5189 * SSOL, the command below shows how this would be computed and how the new values would become part of the data set.

Example 7 shows a rather compound search inquiry, namely to find the records where the total solids are known and the total solids exceed the suspended solids. The particular example shows that 5% of the records show that the total solids are less than the suspended solids. A partial listing is shown.

One of the great advantages of the system is the capability to extract subsets of data. In Example 8 the data from the station Joppa (in Illinois) are extracted and named ILLI. A second set of data is extracted from the stations in Pennsylvania and named PENN. The COMBINE command then combines those two sets into one data set named ILLIPENN. A listing of the sets confirms that a new temporary data set ILLIPENN is available, and a command SAVE would save it permanently on the disk system. The asterisks behind the data set names show which sets are active at the moment (other data sets like NORSDATA are inactive).

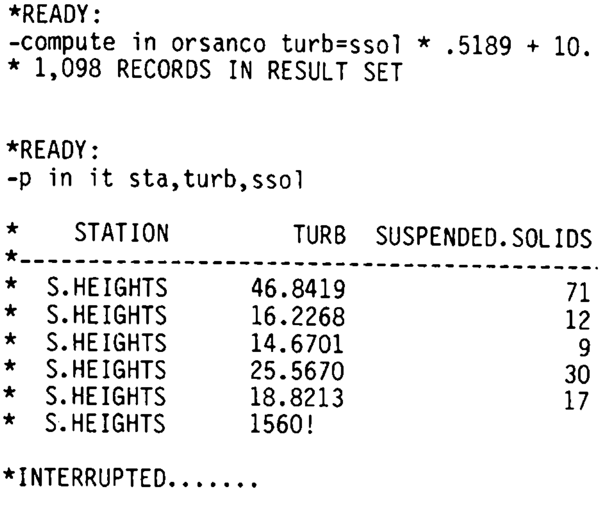
Statistical analysis of data

Once a user has found the data of interest, statistical analyses might be wanted. This is accomplished best using an existing software system called MIDAS



```
*READY:
-xtab in orsanco month
* 12 RECORDS IN RESULT SET
* 1,098 RECORDS REPRESENTED

*       MONTH      COUNT     PERCENT
*-----------------------------------
*     JANUARY         95        8.65
*    FEBRUARY         91        8.28
*       MARCH        114       10.38
*       APRIL        105        9.56
*         MAY         90        8.19
*        JUNE        109        9.92
*        JULY        102        9.28
*      AUGUST         65        5.91
*   SEPTEMBER         84        7.65
*     OCTOBER         60        5.46
*    NOVEMBER         83        7.55
*    DECEMBER        100        9.10
```

Example 5



```
*READY:
-compute in orsanco turb=ssol * .5189 + 10.
* 1,098 RECORDS IN RESULT SET


*READY:
-p in it sta,turb,ssol

*    STATION         TURB  SUSPENDED.SOLIDS
*-------------------------------------------
*  S.HEIGHTS      46.8419                71
*  S.HEIGHTS      16.2268                12
*  S.HEIGHTS      14.6701                 9
*  S.HEIGHTS      25.5670                30
*  S.HEIGHTS      18.8213                17
*  S.HEIGHTS      1560!

*INTERRUPTED.......
```

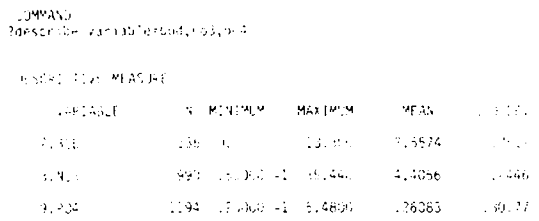
Example 6

```
*READY:
-find in orsanco where tsol is not unknown -
+and tsol less than ssol
* 56 (5.10%) RECORDS FOUND
* 56 RECORDS IN RESULT SET


*READY:
-print in it sta and tsol and ssol

*    STATION  TOTAL.SOLIDS  SUSPENDED.SOLIDS
*-------------------------------------------
*  S.HEIGHTS           140               406
*  S.HEIGHTS           225               437
*    OAKMONT            73               180
*BEAVER.FLLS           166               209
* HUNTINGTON           114               212
* HUNTINGTON           287               443
* HUNTINGTON           122               150
* BELLEVILLE           162               178
* GALLIPOLIS           256               340
* GALLIPOLIS           171               258
* GALLIPOLIS           129               154
*   WINFIELD           107               115
*  LIVERPOOL           136               708
* CINCIN!

*INTERRUPTED.......
```

Example 7

```
*READY:
-find in orsanco where station is joppa
* 33 (3.00%) RECORDS FOUND
* 33 RECORDS IN RESULT SET


*READY:
-name it illi


*READY:
-find in orsanco w sta is oak or pitt or beav or s.hi
* 131 (11.93%) RECORDS FOUND
* 131 RECORDS IN RESULT SET


*READY:
-name it penn


*READY:
-combine illi with penn
* 164 RECORDS IN RESULT SET


*READY:
-name  it illipenn


*READY:
-list sets

*DATA SET NAME  STATUS
*NORSDATA        DISK
*ILLIPENN        TEMP*
*PENN            TEMP*
*ILLI            TEMP*
*RCUTDATA        DISK
*ORSANCO         DISK*
*ICDATA2         DISK
```

Example 8



```
 COMMAND
?describe variables=bod,no3,po4

 [illegible] MEASURE

   VARIABLE         N  MINIMUM     MAXIMUM    MEAN     [illegible]
   7,BOD          336  0.          13.[illegible]  2.5574   [illegible]
   8,NO3          990  .[illegible]E-1  15.44[illegible]  4.4056   [illegible]
   9,PO4         1194  .[illegible]E-1  5.4800     .26083   [illegible]
```

Example 9

(2), an acronym for Michigan Interactive Data Analysis System. One simple command moves the data for analysis, and a large number of statistical procedures are available again in conversational form.

Example 9 shows the simple DESCRIBE command which gives the minimum, maximum, mean and standard deviation of the variables specified.

Quite often it is of interest to see if there are correlations between the variables, and the simple command (Example 10) asks for the correlation coefficients between the variables BOD, nitrate and phosphate. The example shows these coefficients which are all statistically highly significant.

If one would like to see how a particular set of data is distributed, a simple histogram serves as a convenient way to display the data. Example 11 below shows a histogram of the BOD data in the range between 0 and 10 mg l^{-1}. It shows that 762 data are missing, that one BOD value is larger than 10 mg l^{-1} and that the distribution is skewed as one would expect.

If one wishes to calculate a regression equation of the BOD as a function of the suspended solids, the commands below show how this would be obtained. Based on 333 pairs out of a 1098 records, the prediction equation would be:

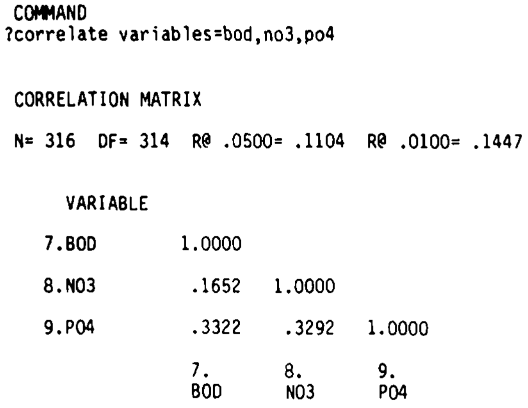
$$\text{BOD} = 2.413 + 0.00247 * \text{SSOL}.$$

The command FINISH stops the program and shows that the entire statistical analysis (Examples 9–12) cost 26 cents, about half a second of computer time (central processing unit) was needed, and that the entire time on the terminal for the statistical part took 4.5 min.

The graphing of data

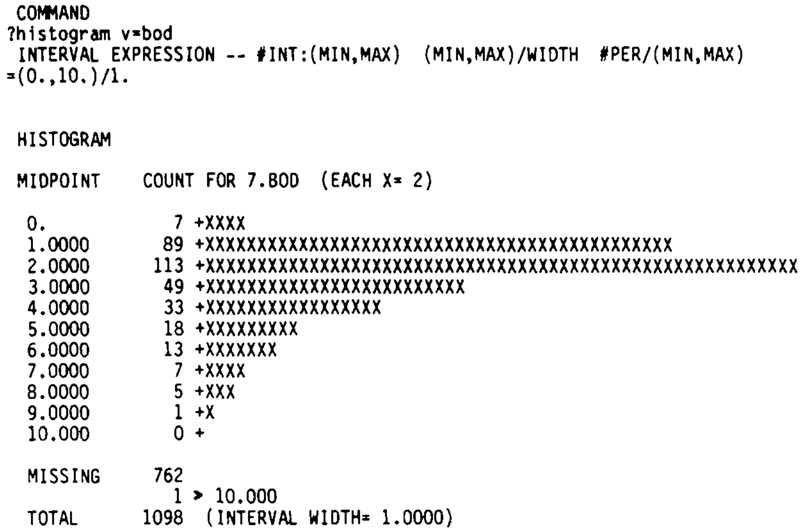
If the user of the system is on a terminal which allows the display of graphics, several commands allow him to plot any of the data he has examined. Example 13 below shows a time series plot of total solids and suspended solids for the station Oakmont. Individual lines can be labeled and drawn differently (solid vs dashed line); if the terminal supports color graphics they can be displayed in different colors. Several different fonts are available to highlight the text.

Figure 2 shows the use of a small microcomputer (an APPLE II) as an intelligent terminal to display



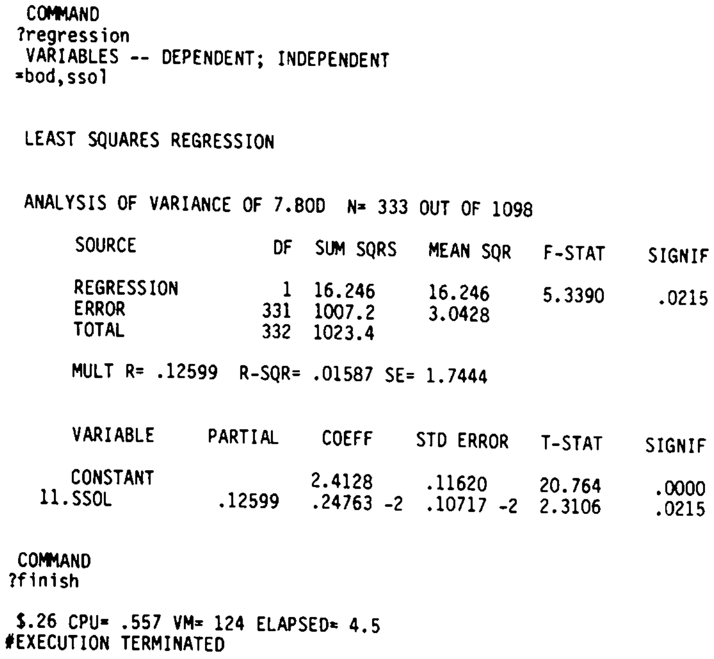
```
 COMMAND
?correlate variables=bod,no3,po4


 CORRELATION MATRIX

 N= 316  DF= 314  R@ .0500= .1104  R@ .0100= .1447


      VARIABLE

    7.BOD        1.0000

    8.NO3         .1652   1.0000

    9.PO4         .3322    .3292   1.0000

                 7.        8.        9.
                 BOD       NO3       PO4
```

Example 10



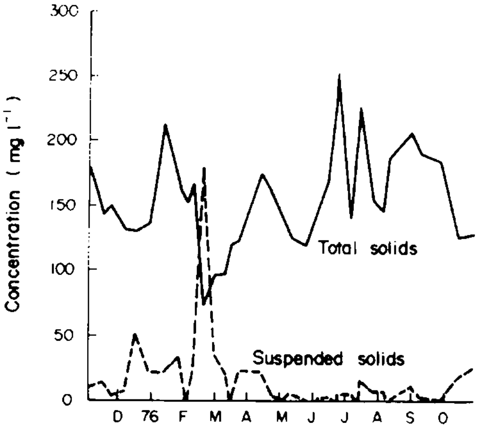
```
 COMMAND
?histogram v=bod
 INTERVAL EXPRESSION -- #INT:(MIN,MAX)  (MIN,MAX)/WIDTH  #PER/(MIN,MAX)
=(0.,10.)/1.


 HISTOGRAM

 MIDPOINT    COUNT FOR 7.BOD  (EACH X= 2)

  0.            7 +XXXX
  1.0000       89 +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  2.0000      113 +XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  3.0000       49 +XXXXXXXXXXXXXXXXXXXXXXXXX
  4.0000       33 +XXXXXXXXXXXXXXXXX
  5.0000       18 +XXXXXXXXX
  6.0000       13 +XXXXXXX
  7.0000        7 +XXXX
  8.0000        5 +XXX
  9.0000        1 +X
  10.000        0 +

  MISSING     762
                1 > 10.000
  TOTAL      1098  (INTERVAL WIDTH= 1.0000)
```

Example 11



```
  COMMAND
 ?regression
  VARIABLES -- DEPENDENT; INDEPENDENT
 =bod,ssol


  LEAST SQUARES REGRESSION


  ANALYSIS OF VARIANCE OF 7.BOD  N= 333 OUT OF 1098

       SOURCE              DF  SUM SQRS   MEAN SQR   F-STAT    SIGNIF

       REGRESSION           1  16.246     16.246     5.3390     .0215
       ERROR              331  1007.2     3.0428
       TOTAL              332  1023.4

       MULT R= .12599  R-SQR= .01587 SE= 1.7444


       VARIABLE     PARTIAL    COEFF     STD ERROR   T-STAT    SIGNIF

       CONSTANT               2.4128     .11620     20.764     .0000
   11.SSOL          .12599   .24763 -2  .10717 -2  2.3106     .0215

 COMMAND
?finish

 $.26 CPU= .557 VM= 124 ELAPSED= 4.5
#EXECUTION TERMINATED
```

Example 12



Example 13

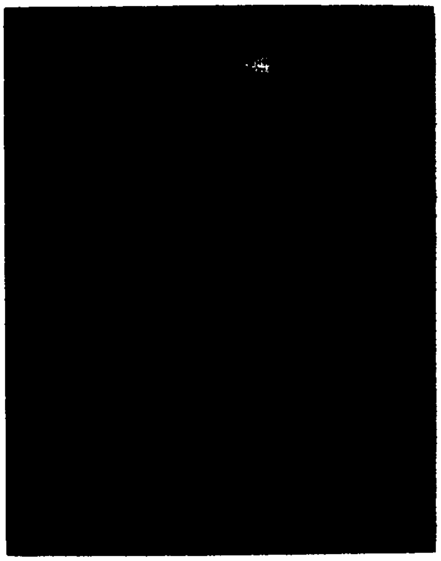


Fig. 2

the data on the authors television set at home. The tape recorder in the back allows the storage and retrieval of programs, data and graphs. Graphs generated through conversations with the system may be transmitted to the terminal system through the modem and acoustic coupler shown to the right of the microcomputer. The picture shown on the television screen is the same as the graph shown in Example 13.

SUMMARY AND CONCLUSIONS

The system described is a very powerful data analysis and management system and allows the analysis of water quality data in conversational form from a computer terminal. Not all the commands available were shown but the reader may have gained an impression of the interactive nature of the system through the many examples. This type of a system can be developed on other computer systems too and will give scientists and administrators direct access to the data of interest without going through intermediary data processing personnel.

APPENDIX

1. MICRO *Information Retrieval System Reference Manual* (Edited by Kahn M. A. *et al.*). University of Michigan, October (1977).

2. MIDAS- *Michigan Interactive Data Analysis System* (Edited by Fox D. J. & Guire K. J.). Statistical Research Laboratory. The University of Michigan, September (1976).

3. GRAF *An Interactive Graphing Package* (Edited by Burling S. & Thomas R.). University of Michigan, February (1979).