DCSUG Newsletter

Fourth Quarter Meeting

When: Tuesday, December 1, 1998
Time: 8:30a.m. to noon
Place: Bureau of Labor Statistics
                  Postal Square Building
                  Room G440
                  First Street North East
                  Washington DC

Agenda:		
8:30 - 9:00		Registration and refreshments
9:00 - 9:15		Announcements
9:15 - 10:15		Marge Scerbo, Center For Health Programming & Development -UMBC. ***A Health Care Claims Data Mart: Construction & Exploitation***
10:15 - 10:30		Break
10:30 - 11:30		Andrew Karp, Sierra Information Services, Inc. ***Indexing and Compressing Data Sets: How, Why, and Why Not***
11:30 - Noon		User-to User Questions and Answers Moderators - DCSUG officers

Visitors must use the visitor's entrance on First Street NE. Directions for those taking Metro: Take the Red Line to Union Station. Exit the station via the Amtrak exit, but do not go up to the train level. Instead, after going up to and out through the farecard gates, turn to the left where there is an exit to the street (First Street NE). Directly across from the exit are the employee and visitor's entrances to the building. After checking in with security, take the main elevators down to the G level. **BRING PICTURE ID!**

If you have special needs and are planning to attend this meeting, please contact any DCSUG officer no later than 11/23/98

SAS® is a registered trademark of the SAS Institute, Cary, NC

✱ ✱ ✱ ✱ ✱ ✱ ✱ ✱ ✱ ✱

☎☎☎☎☎☎☎ **If you are not currently a paid member of the DCSUG and are planning to attend the meeting, please Email Karen Dennis at dennisk1@westat.com or call 301-294-3876 by Nov 24th and leave your name. Security at the Bureau of Labor Statistics is very tight and having your name on a list of possible attendees will make your entrance to the building easier.** ☎☎☎☎☎☎☎

President's Message
Peter Knapp

Two years ago I presented my first SAS paper at NESUG. Boy was I nervous and stiff. A few months later I presented it at a DCSUG meeting. This time things went much more smoothly. I wished I had presented at DCSUG first. This year I got it right. Before presenting my paper at NESUG, I presented it at the last DCSUG meeting. I'm glad I did. The advice, encouragement, and feedback I got at DCSUG helped my presentation at NESUG greatly. I recommend that anyone who wants to present a paper at a regional or international SAS conference first practice at DCSUG. You'll be glad you did!

Indexing and Compressing Data Sets: How, Why, and Why Not

Andrew Karp, Sierra Information Services, Inc.

Abstract

This paper discussed two important tools in managing SAS data sets: indexing and compression. Appropriate applied indexes will speed the retrieval of observations from SAS data sets, and therefore often a useful tool to extract subsets of observations from large SAS files. Compression is used to make data sets 'smaller' by removing blanks and repeating characters from character variables. Although both tools are often useful when working with SAS data sets, there are times when improperly applied indexes will increase the amount of time needed to extract observations from a data set, and situations where compressing a SAS data set will actually make it bigger! This tutorial will show you how and when you might want to consider using indexes or compression with our data sets, as well as point out important pitfalls to consider before applying them.

Andrew H. Karp is President of Sierra Information Services, Inc., a San Francisco-based SAS Institute Quality Partner providing strategic data analysis and data management services, as well as customized SAS System software training, to clients in many industries. Andrew is a 15 year SAS System software user who has presented over 45 invited papers at SAS user group meetings across North America, as well as in the United Kingdom, New Zealand and Australia. He is a graduate of The George Washington University in Washington, DC and enjoys returning to the National Capital Area to participate in DCSUG meetings.

A Health Care Claims Data Mart: Construction and Exploitation

Stuart B. Levine, SAS Institute, Inc.
Marge Scerbo, UMBC

Abstract

Health care claims data are complicated animals. The eccentricities of this data include differences in the types of claims files and the services which populate those files, variations in the type of providers who service the population and the information needed about each type of provider, and finally, the actual populations of health care recipients and all of their individual specifications.

Being able to sort through all of this data (and there is usually a lot of it!) and answer queries from insurers, providers, analysts, etc., and other people within your organization can be a real challenge. In the past, the solution was to have people submit requests for one or more programmers to extract data and run analyses. The problem was twofold: programmers were overwhelmed with requests, and worse, the data and/or analyses requested were not the correct solution to the problem.

This paper will highlight how this challenge was met through the development of a Data Mart. This includes the steps taken in the design of the Mart, its base tables and MDDBs (MultiDimensional DataBase), the definition of SAS/EIS objects that can exploit MDDBs, and other specific customizations that were developed to allow one organization, the UMBC Center for Health Program Development and Management, to address these data access issues.

Stuart B. Levine is a Senior Applications Developer for the Consulting Services Department of SAS Institute. His expertise is primarily in DATA step programming, SAS/AF (including FRAME technology), SAS/EIS, SAS/MDDB Server and application design. He holds a Bachelor's degree in Accounting from the University of Dayton. Stuart has been a SAS user for over 12 years and with SAS Institute for over 5 years.

Marge Scerbo is Senior Analyst at the Center for Health Programming and Development at the University of Maryland Baltimore County. This group was established as the state's policy and research arm for the Maryland Medicaid system. SAS has been a major part of her life for the past many years (double digits now!). Marge has participated in local, regional and international user groups as an organizer and speaker.

Marge has been studying karate for over 4 years and has just earned her black belt. She promises not to toss anyone who asks a question!

Questions and Answers

Moderated by DCSUG Officers

Bring your SAS programming or procedure questions to DCSUG and its members. During this open session, you can ask questions of and get answers from experienced SAS users. All are welcomed to participate

Desktop Special Interest Group

by David Kiasi-Barnes

DCSUG's Desktop Special Interest Group will meet in the Spring (Date to be announced) at Westat, Inc. 1650 Research Blvd. in Rockville at 7:30PM. If you have any suggestions for future meeting topics, speakers, or alternate locations, or if you'd like to be added to our group's mailing list so you can get announcements of upcoming meeting events, please contact David Barnes at (301) 350-4752 or Yesvy Gustasp at (301) 589-4530. Also, if you need a ride from the Metro, please call either one of us.

Directions to Westat: Take I-270 8 miles from the Capital Beltway to exit #6B - Route 28 (toward Darnestown). Once on 28, go past the traffic light at the Shell station and turn right at the next traffic light, which is Research Blvd. The 1650 building is less than 1/4 mile on the left. You can park in any non-reserved space. Signs will be posted in the lobby to direct you to the meeting room. Westat has many buildings on Research Blvd. so make sure you're at 1650.

If using Metro, take the Red line towards Shady Grove and get off at the Rockville station. Take Ride-On bus #54 in the direction of Lakeforest Mall. The bus goes out on 28 past I-270 and the Shell station and then turns onto Research Blvd. Get off of the first stop on Research and cross the street. It takes 8-10 minutes from the Metro station to Westat. The #54 bus runs every half hour.
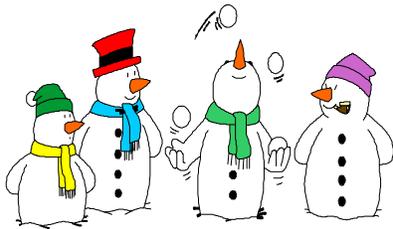
DCSUG 1998 Steering Committee
OFFICERS

CHAIR -- Peter Knapp
(202) 482-1359   (202)482-1388(fax)
peter_knapp@ita.doc.gov

SECRETARY -- Larry Altmayer
(301) 457-2581   (301) 457-1777 (fax)
laltmaye@census.gov

TREASURER -- Arlene Siller
(301) 436-8522 x188   (301) 436-5452 (fax)
abs2@cdc.gov

SAS LIAISON -- Linda Atkinson
(202) 694-5046   (202) 694-5718 (fax)
atkinson@econ.ag.gov

OTHER COMMITTEE MEMBERS

Karen Dennis
(301) 294-3876   (301) 294-2034 (fax)
dennisk1@westat.com

Frank Fry
(202) 452-2666   (202) 452-6433 (fax)
ffry@frb.gov

Michael Raithel
(202) 927-0675   (202) 927-1896 (fax)
maraithel@mcimail.com

Mike Rhoads
(301) 251-4308   (301) 294-2040 (fax)
rhoadsm1@westat.com

Howard Schreier
(202) 482-4180   (202) 482-4614 (fax)
Howard_Schreier@ita.doc.gov

Douglas McAllaster
(703) 697-7619
mcalldl@hqda.army.mil

DESKTOP SPECIAL INTEREST GROUP

CHAIR -- David Kiasi-Barnes
(301) 350-4752   (301) 350-4785 (fax)
appalt@aol.com

VICE-CHAIR -- Yesvy Gustasp
(301) 589-4530   (202) 633-2668 (fax)
Yesvy.Gustasp@mix.cpcug.org

Calendar of Meetings and Events

This feature is meant to give users an idea of some of the activities of area users groups and special interest groups, as well as regional and international groups. If your group would like to be regularly included in this feature, please call Mike Rhoads or one of the DCSUG officers. There is a DCSUG directory with phone, fax, and E-mail contact information elsewhere in this newsletter.

December

1 **DCSUG** DC SAS Users Group, General Meeting
Time: 8:30 am - 12:00 noon
Site: Bureau of Labor Statistics
2 Massachusetts Ave. N.E.
Program: See front page announcement
Contacts:
Linda Atkinson at (202) 694-5046
atkinson@econ.ag.gov
Peter Knapp at (202) 482-1359
peter_knapp@ita.doc.gov

March

9 **DCSUG** DC SAS Users Group, General Meeting
Time: 8:30 am - 12:00 noon
Site: Bureau of Labor Statistics
2 Massachusetts Ave. N.E.
Program: TBA
Contacts:
Linda Atkinson at (202) 694-5046
atkinson@econ.ag.gov
Peter Knapp at (202) 482-1359
peter_knapp@ita.doc.gov

April

11-14 **SUGI** SAS Users Group International Conference
Site: Fontainebleau Hilton Resort and Towers
Miami Beach FL
Contacts:
Mic Lajiness at (616) 833-1794
Michael.S.Lajiness@am.pnu.com

Other Groups

CENSUG Census SAS Users Group
Contact: Rick Denby at (301) 457-3208

MDSUG Maryland SAS Users Group
Contact: Mr. R.H. Miller, (410) 740-4229
refa09a@prodigy.com

NCHSSUG National Center for Health Statistics SAS Users Group
Contacts: Arlene Siller at (301) 436-8522
abs2@cdc.gov
Linda Tompkins at (301) 436-7022

NIHSUG National Institutes of Health SAS Users Group
Contact: Ray Danner at (301) 496-6037

Mark Your Calendar

SAS Talk

By Ian Whitlock

A colleague came with a question about PROC SORT. She wanted to sort approximately 17 million records with just two variables, and the job was bombing. What could be done? First I suggested more space, but when her boss suggested that there was plenty of space, I decided to go and look at the problem.

Her "real" problem was to make one record for each person/day covered. The input file had about 7 thousand records with an ID and a relevant date interval for the person. The problem was that there were multiple records per person, and intervals could overlap; hence she wanted a count of the number of intervals containing the given day. While the sort was executing, I thought about the problem.

My colleague's solution was straight forward and simple - generate one record for each day in each interval for each ID, sort by ID and DAY, and then count the days in a DATA step outputting at LAST.DAY. This is how she converted a problem about 7 thousand observations into one about 17 million. Instinctively, I thought the sort shouldn't be needed. Surely one could store the end points of the intervals in temporary arrays and then write out the records for an ID at LAST.ID.

Here is one solution given that no person had more than 100 records, together with test data and the original code involving the sort.

```
/* create test data */

data match ( keep = id start stop ) ;

   do id = 1 to 10 ;

      do i = 1 to ceil ( ranuni ( 8683403 ) * 100 ) ;
         start = ceil ( ranuni (0) * 10000 - 5000 ) ;
         stop = start + 500 + ceil (ranuni (0) * 500);
         output ;
      end ;
   end;
run ;
/* ------------ method without sort -------------- */
proc sort data = match ;
   by id start stop ;
run ;

data matched2 ( keep=id day n_match ) ;

   retain lastday . ;
   array starts (100) _temporary_ ;
   array stops  (100) _temporary_ ;

   set match ;
   by id ;

   if first.id then
   do ;                    /* initialization */
      do i = 1 to dim (stops) while (starts(i) ^= .) ;
         starts ( i ) = . ;
         stops  ( i ) = . ;
```

```
            end ;
            indx = 0 ;
            lastday = . ;
        end ;
                         /* save periods    */
        indx + 1 ;
        starts ( indx ) = start ;
        stops ( indx ) = stop ;
        if lastday < stop then lastday = stop ;

        if last.id ;        /* ready to output */

        nperiods = indx ;
        do day = starts ( 1 ) to lastday ;
            n_match = 0 ;
            do i = 1 to nperiods ;
                if starts ( i ) <= day <= stops ( i ) then
                    n_match + 1 ;
            end ;
            if n_match > 0 then output ;
        end ;
    run ;

    /* --------------- method with sort ------------- */
    data matchday(keep=id day);
        set match;
        length day 4;

        do i = start to stop;
            day = i;
            output;
        end;
    run;

    proc sort data=matchday;
        by id day;
    run;

    data matched1(keep=id day n_match);
        set matchday;
        by id day;
        retain n_match;
        if first.day then n_match=0;
        n_match=n_match+1;
        if last.day;
    run;

    proc compare data = matched1 compare = matched2 ;
    run ;
```

The last step is used to show that the results are the same. How did the execution times compare? Boost the number of IDs and test it for yourself.

The moral here - be suspicious when the problem drastically changes size; and ask - is there a better algorithm? Try to develop an intuition for what is necessary and how to approach a programming problem. Don't be afraid of slightly more complex code. It can be more fun and execute much faster.

SAS is Y2k Compliant, Are You?

By Ian Whitlock

Everybody seems to be somewhat aware of the 2000 programming problem in general, but I rarely see signs of it in the SAS programs that I have a chance to read. I wonder how many SAS programmers would recognize 2000 problems in SAS code and know how to fix them.

Here is a little test. The program below is divided into three parts

1. Read a flat file (not given) into a SAS data set

2. Report monthly counts for multiple years

3. Write out a flat file of monthly counts for multiple years

Here are some problems to think about.

- What line(s) tip you off to a problem?

- Can the parts be fixed independently?

- What is the minimal number of changes that are required?

In answering the above you are not allowed to change the input file or the output file since these files interface with other systems. First assume all date related values are legal (non missing) and represent date between January 1, 1910 and December 31, 2009.

```
/* Read data from IN,
   sum and report monthly counts
   write monthly counts to OUT
*/

filename in  'c:\fixedlayout\input.dat' ;

filename out 'c:\fixedlayout\output.dat' ;

/* -------------- Part 1 --------------- */

data proj ;

   infile in ;

   input mon 2. day 2. yr 2. count 6. ;

run ;

/* -------------- Part 2 --------------- */
proc summary data = proj nway ;
   class yr mon ;
   var count ;
   output out = summary sum=total ;
run ;

proc transpose data = proj out = tproj prefix = cnt ;
  by yr ;
```

```
      id mon ;
      var count ;
   run ;

   title "Monthly counts for each year
   proc print data = tproj ;
   run ;
   /* -------------- Part 3 --------------- */
   proc sql;
     select min ( yr ) , max ( yr )
         into : firstyr , lastyr
        from proj ;
   quit ;
   data _null_ ;
      array cnt (&firstyr:&lastyr, 12) _temporary_ ;

      if eof then link wrtout ;

      set proj end = eof ;

      cnt ( yr, mon ) + count ;
   return ;

   eof:
      file out noprint ;
      do yr = &firstyr to &lastyr ;
         put yr 2. @ ;
         if mod ( yr, 4 ) = 0 then
            put '*' @ ; /* leap year */
         else
            put ' ' @ ;
         do mon = 1 to 12 ;
            put cnt ( yr , mon ) 8. @ ;
         end ;
         put ;
      end ;
   return ;
   run ;
```

Add code to produce a second report giving two numbers - the total count for the first five days of each month and the total count for the last five days of each month.

To participate in the survey, send answers to **whitloi1@westat.com** with the subject "DCSUG 2Yk" by January 15, 1999. All names will be treated confidentially. The code is available at **http://www.ita.doc.gov/industry/otea/dcsug/**.

Election of 1999 DCSUG Officers

Peter Knapp — Candidate for Chair

Peter Knapp interned at the Bureau of Labor Statistics (BLS) where he learned the joys of SAS on an MVS mainframe environment. After contracting with BLS for three years, he joined the International Trade Administration (ITA) at the Commerce Department. For eight years Mr. Knapp has helped administer international trade law at ITA. During this time he has guided SAS processing from a central mainframe shop to an end user PC based system.

Arlene Siller -- Candidate for Treasurer

Arlene Siller is a Computer Specialist with the National Center for Health Statistics and has used SAS for 20 years. Her responsibilities include programming and analysis using SAS/Base and SAS/Stat for National health surveys. She is the SAS Site Representative for her agency and co-chair of the in-house NCHS Users Group (NCHSSUG). She is very involved in providing support for NCHS SAS users, and

enjoys this personal interaction part of her responsibilities the most. She has programmed using the TSO, Wylbur, VMS, DOS, Windows and UNIX operating systems.

Larry Altmayer -- Candidate for Secretary

Larry Altmayer has been with the U.S. Census Bureau for twenty years and is a computer specialist. He has more than nine years SAS programming experience, most recently with macros. He has used SAS and SUDAAN for modeling household survey nonresponse based on demographic person and household variables, and SAS for producing tables of employment percents and unemployment rates for the Current PopulationSurvey. His programming experience has been mainly in the VMS operating system. He has a B.A. in Statistics from the Univ.of Buffalo, and Master of Applied Statistics from Ohio State University.

Linda Atkinson -- Candidate for SAS Liaison

Linda Atkinson is a statistician with USDA's Economic Research Service, and has used the SAS System for 22 years. Her areas of expertise include statistical analysis and econometric modeling. She has a strong interest in user interfaces. She has run jobs and consulted with users on SAS programs under the TSO, CMS, DOS,OS/2 and Windows operating systems, and is the SAS Site Representative for her agency. She has been an Invited Speaker at SUGI and NESUG conferences.

DCSUG 1999 BALLOT

Place an "X" In the Box beside the candidate of your choice. You may write in the name of someone who has not been nominated if you wish.

DCSUG Officers

Chairperson

Description: Lead DCSUG, delegate assignments, chair group meetings, preside over nomination and election of officers.
                    Peter Knapp

Treasurer

Description: Manage the group finances: collect dues and fees, make disbursements, maintain record of income and disbursements. The Treasurer will also prepare a report of income and disbursements for distribution at the fourth quarter general meeting.
                    Arlene Siller

Secretary

Description: Record minutes of group meetings, maintain membership list, prepare meeting notices.
                    Larry Altmayer

SAS Liaison

Description: Represent and increase the influence of DCSUG to the SAS Institute
                    Linda Atkinson

DCSUG Dues

Dues are set annually by the steering committee and voted on by the member ship at the fourth quarter meeting. Dues will be payable by the first quarter general meetings. Dues will be used to finance mailings, telephone charges, newsletter and other expenses as approved by the group of officers.
                    Keep proposed (current) rate ($10.00 per year individual, $50.00 per year corporate)
            Increase   Decrease

Please mail ballot to:

Washington DC SAS Users Group
P.O. Box 44670
Washington, DC 20026-4670

PROGRAMMER/ANALYST

The Human Resources Research Organization (HumRRO) is seeking an individual for mid-level SAS programming in a Windows environment. The successful candidate will work under the direction of senior researchers with diverse types of data (e.g., survey data as well as large MIS databases) in statistically based research projects involving human resources issues and will create and manipulate databases and generate statistical output. BA in social science with at least one course in statistics is required. A minimum of three years work experience in a research organization using SAS BASE and STAT products is also required. The successful candidate will have work experience outside of school in data handling and basic statistical techniques. Experience with Access, and Visual Basic and mainframe SAS is a plus. Location convenient to METRO, with competitive salaries and an excellent benefits package. Please send resume and salary requirements to: HumRRO, Attn: Personnel Office, Suite 400, 66 Canal Center Plaza, Alexandria, VA 22314.   EOE

RESEARCH ASSISTANT

The Human Resources Research Organization (HumRRO) is seeking an individual to perform research assistant duties. Duties will include, but are not limited to, gathering data, programming in SAS, and conducting literature reviews. The successful candidate will also conduct statistical analyses under the direction of senior research staff and will provide word processing, filing, faxing and xeroxing support. BA/BS in social sciences, preferably Psychology, is required. The successful candidate will have a working knowledge of SAS (BASE and STAT) and MS Word and/or Wordperfect. Previous work experience or internship in a research organization is highly desirable. Attention to detail is a must; knowledge of spreadsheets (e.g., EXCEL, LOTUS, or Quattropro) is a plus. Location convenient to METRO, with competitive salaries and an excellent benefits package. Please send resume and salary requirements to: HumRRO, Attn: Personnel Office, Suite 400, 66 Canal Center Plaza, Alexandria, VA 223l4   EOE

Membership Application/Renewal

Date:_____

Name:_____

Company Name:_____

Mailing Address:_____

City,_____   State, _____   Zip _____

Phone # _____   Fax # _____

E-Mail:_____

Preferred mode of contact (check one):

                    Fax   e-mail

SAS Version(s) _____

SAS Product(s): _____

Operating System(s) _____

Check one:

                    Individual membership   Corporate sponsorship
                     for 1 year is $10.00   for 1 year is $50.00

Please indicate a contact person for Corporate Sponsorships!

_____

                    Check one:
                      New membership   Renewal membership
                    DCSUG Membership Directory

DCSUG is considering publishing a membership directory in 1997, containing members name, work affiliation (name of institution), work telephone number, e-mail address, and any areas of expertise that you would be willing to identify. This directory would be provided to DCSUG members only and would not be given to outside sources. If you would like to be included in such a directory please mark the boxes below, as appropriate.

                    I would like to be included in the membership directory DCSUG is considering publishing. Include my name, work affiliation, and work telephone number.
                    Include my e-mail address as shown on this form.
                    Include the following areas of expertise (up to 2 products, platforms).

                    Mail to:
                Washington DC SAS Users Group
                      P.O. Box 44670
                  Washington, DC 20026-4670
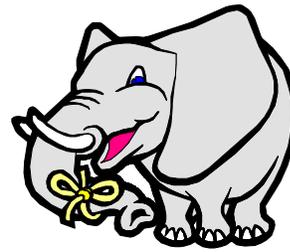
1999 Meetings

Plan ahead and mark your calendars now. The DCSUG will tentatively meet on the following dates in 1999: March 9, June 8, September 14, December 7. Meeting times and locations to be announced. If at any time you have a question about an upcoming meeting or other DCSUG information, call one of the members of the Steering Committee.

Remember to Join DCSUG!!!

Don't forget to renew your DCSUG membership for 1999. Dues are used to defray the costs of producing and mailing the newsletter and provide refreshments at our meetings. You'll make sure that you keep receiving the newsletter and keep DCSUG going. If you have not joined DCSUG, now is a great time to do so. Joining DCSUG is as easy as completing the membership form included in this newsletter. Individual memberships are only $10; corporate memberships are $50.

You're Invited
to visit the DCSUG home page

The URL for the DCSUG home page on the Web:
http://www.ita.doc.gov/industry/otea/dcsug

Washington DC SAS Users Group
P.O. Box 44670
Washington, DC 20026-4670