From 50,000,000 Claims to One Analytical File

Louise Hadden, Mike Murphy, and Alan J. White
Abt Associates Inc., Cambridge MA

ABSTRACT

Researchers often confront the fact that the most common repositories of data are the least suitable for careful analysis. This is an outline of a small-area study performed by Abt Associates Inc. under one of its health care study contracts, during which a small mountain of data was turned into information that may be useful to our customer.

This paper illustrates the power of the SAS® System for processing a great quantity of disparate data, its statistical analytical capability to make use of this information, and its flexibility for producing quality tables and graphs for presentation. Some of the examples will be specific to the UNIX operating system. Base SAS, SAS/FSP®, SAS/GRAPH®, and SAS/STAT® are used in this application. For the purpose of confidentiality, the MSA-level statistics shown in each exhibit are fictitious.

BACKGROUND

Abt Associates Inc. is under contract to provide special analysis services to the Statistical Analysis Durable Medical Equipment Regional Carrier (SADMERC). The SADMERC is a support organization for Medicare's durable medical equipment regional carriers, or DMERCs. These four carriers report to HCFA (Health Care Financing Administration) and are responsible for processing what are known as DMEPOS claims: those for durable medical equipment, prosthetics, orthotics, and supplies covered by the Medicare programs. The SADMERC contract is held by Palmetto Government Benefits Administrators (PGBA).

The SADMERC and HCFA collected 56 million claims of Medicare beneficiaries' DMEPOS products last year. These records account for sales and rentals by 95,000 different suppliers to more than 5.5 million Medicare beneficiaries in all 50 states.

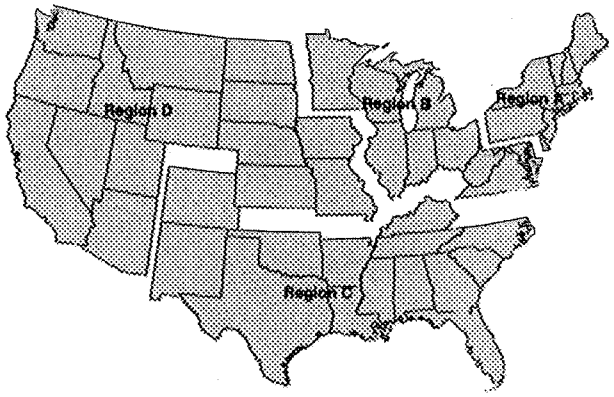
Abt is responsible for assisting the SADMERC in analyzing these claims histories to identify areas of over and under utilization, and other areas of fraud and abuse of the Medicare DMEPOS benefit. Specifically, we help in detecting trends in the cost and utilization of these items, and target enforcement efforts in locations of high abuse.

To accomplish this, we:

- develop consistent classification schemes for claims data
- identify business lines and geographic entities
- classify and summarize claims to those levels used in the small-area study
- include industry and census data where appropriate
- conduct series of univariate correlation and multivariate regression analyses
- provide reports, tables, and maps to convey our findings

Our UNIX computing environment at Abt is a chargeback system, so program efficiency when reading files of this size is of extreme importance to the contract.

Exhibit 1: PROC GMAP Displaying DMERC Regions



I. AGGREGATING CLAIMS

Our clients provide us with claims data from two sources:

- DMEPOS claims processed by the DME regional carriers (50 million claims)
- Another 6 million DMEPOS claims processed by regular Medicare carriers who were being phased out of processing responsibility for these claims

The two sets of claims arrive on 20 and 14 MVS 3480 cartridges respectively. Fortunately, when copied, each set fits easily on a single 8mm data cartridge. (A 3480 cartridge, without compression, holds about 200MB; an 8mm can hold up to 7GB.) Each claim should at least contain information about the provider(s), the identity and location of the Medicare beneficiary, the amounts of dollars and units submitted and allowed, and the nature of the item or service provided.

Practice, Practice

There is no need yet to worry about sorting or otherwise processing these millions of records until the data are thoroughly investigated and brought up to a uniform standard that will enable proper analysis. The next step is to draw a sample from each source. We select a 1%

beneficiary terminal digit sample of 558,000 claims. This provides us with a representation of the incoming data that will not bankrupt our computer budget as we attempt to determine the problems we will encounter in subsequent processing.

We find the following about the important identifiers among these DMEPOS claims:

- The regional or regular carrier is shown on every record (reporting by each of the four DMERC regions is essential)
- The beneficiary is consistently identified by a health insurance claim (HIC) number and a zip code of residence on every record
- The provided service is coded on every record by means of a procedure code. There are 3500 possible codes, a bit much for modeling.
- A referring physician is noted by his/her unique physician identification number (UPIN) on 75% of all claims. The remainder are blank.
- The supplier of the service, however, is poorly identified. Claims processed by the regional carriers all consistently specify the seller by a National Supplier Clearinghouse (NSC) identification number. Claims processed by the regular carriers have a hodgepodge of tax numbers, provider IDs, and carrier numbers to define the supplying entity. Some refer to a supplier's national headquarters, others to a local branch store. Since the activity of sellers, and their relationships to physicians, is one of prime focus of our investigation, we will have to devise a scheme to define suppliers as consistently as our data sources allow.

Use of SAS FORMAT Libraries

Prior to receipt of these claims, we have already assembled much of the cross-referencing to be used on them in SAS FORMAT libraries. These formats will be called upon to standardize the classification schemes on each claim during the first pass of the 56 million claims file.

To enable small-area analyses, we want to equate zip code claims activity to larger geographies. As is the case at a lot of SAS sites, we have supplemented their built-in state/zip geographic functions by adding in our own formats that link zip codes to counties, metropolitan statistical areas (MSAs), and DMERC regions. MSAs (comprised of counties, except in the New England states) have been chosen as the foundation of our SADMERC small-area study, largely because it is more useful and understandable to the customer than counties or zips as a first level of investigation. Among the sources of supplemental geographic information that we use on the World Wide Web are:

- www.census.org - the U. S. Census Bureau
- www.ciesin.org - the Consortium for International Earth Science Information Network
- oseda.missouri.edu:80/uic - the University of Missouri - St. Louis Urban Information Center

We will also take the thousands of procedure codes that describe DMEPOS items, and map them to one of 50 product groupings provided by the customer. This is another simple task for a SAS format. There is no need to detail the drudgery of cleaning up the suppliers' identity at this point. Suffice to say that after much trial and error, we determined that an NSC ID can be accurately assigned using a best-available-match approach that links a supplier's zip code and one or more of its other identifiers to its known NSC ID. A series of SAS formats is built to accomplish this.

Standardize, Then Summarize Claims

Once the investigation and testing of the 1% sample is finished, the expensive part begins. The 56 million records are read from the two input cartridges in a single SAS data step. SAS formats assign the beneficiary's and supplier's region and MSA, the supplier's ID, and the product group. Other program statements define additional quantitative and indicator variables to be used in ensuing models. Retaining only the needed data items, the resulting clean claims file just fits onto a single 8mm cartridge. This tape to tape processing step runs on our RS/6000 AIX system for 24 hours of real time, 18 hours CPU time.

Turning this narrowed and consistent claims file into MSA-level analytical files involves counting identifiers of the principals involved (e.g., the number of suppliers per MSA), as well as calculating univariate statistics of analytical measures. In a normal-size task, this could be accomplished using PROCs MEANS or SQL. With so many more records to process than memory will allow, we use the manual method of PROC SORT / DATA STEP. This process must be repeated several times as we count and sum, across regions and MSAs, the intersection of every discrete supplier, beneficiary, and physician involved in a claim. In order to sort such a large file, we process subsets of the claims tape on separate disk packs (OPTIONS OBS=n; PROC SORT... OPTIONS FIRSTOBS=n+1 OBS=y; PROC SORT....), then SET all of the subsets with a BY statement.

When we run this job, tape to disk, it consumes a mere 9 1/2 hours clock time (4 hours CPU). Once finished, we have determined which of the 95,000 suppliers had claims in each region and MSA, to how many beneficiaries, in consort with how many physicians, in what amounts, and in which product groups. Two files are output: one for the 360 MSAs and one for the 18,000 combinations of MSAs and the 50 business lines.

Finishing Work

The value of this DMEPOS claims information is greatly enhanced when census and industry source data can be used as a baseline. This enables comparison of beneficiary usage against the population, supplier charges vs the industry, etc. One important source that we license is the *Bureau of Health Professions Area Resource File*, by Quality Resource Systems, Inc. It offers over 5,000 longitudinal data items covering census, mortality, and health care utilization in every US county. This data is recalculated at the MSA level, then merged to our own files. We then calculate the remaining statistics that will be used in the modeling process to come.

II. THE MODELING PROCESS

Anomalies between the actual and the predicted measures across MSAs, uncovered by regression analysis, helps us begin the process of uncovering suspected areas of abuse. It is not known what percentage of DMEPOS claims involve fraud or abuse, as only a fraction of the fraud and abuse that occurs is detected. Because we cannot directly identify fraud and abuse from claims data, we instead attempt to identify abuse-prone areas indirectly, by examining MSA-level variations in costs and utilization, using multivariate regression techniques.

Identify Aberrant Areas
We identify areas with high costs and utilization, given the characteristics of an area's Medicare population and provider community. Variables used to adjust for the influence of demographic, socioeconomic, and health-status factors include the gender, age and occupational distribution of the population, the percentage of the MSA's Medicare population with income below the poverty line, per-capita reimbursement for hospital insurance, and the percentage of the MSA's population with four or more years of college. To adjust for the impact of provider community characteristics, our models include measures of the number of general practice and internal medicine physicians and the number of short-term general hospital beds.

Relate Usage to Fraud Indicators
Potential "fraud and abuse indicators" include the percentage of new and inactive suppliers in a market, and the ratio of suppliers to Medicare beneficiaries in an area that are believed to be correlated with fraud and abuse. Holding constant MSA demographic, socioeconomic, and provider community characteristics, MSAs with higher values of these fraud and abuse indicators are expected to have higher levels of fraud and abuse. This type of information is useful in our attempts to develop a profile of the abusive supplier: part of Abt's efforts to develop an "early warning system," intended to identify potentially aberrant suppliers before they actually engage in abusive billing patterns.

Supplier entry and exit patterns and a high concentration of suppliers in a given market are potential indicators of abuse. MSAs that have a higher ratio of suppliers to Medicare beneficiaries are expected to also have higher underlying levels of fraud and abuse. According to microeconomic theory, the entry of new suppliers into a market occurs in response to profit opportunities, which may exist partly as a result of some type of abuse of the DMEPOS benefit. In addition, some suppliers that appear to be new to a market may be existing suppliers that have begun operating under a new name and Medicare billing number. The relationship between abuse and the proportion of inactive suppliers in a given market (i.e., those who have no allowed charges in the current period but had positive charges in the previous period) is less clear. A higher proportion of inactive suppliers in an area may reflect "hit and run" entry and exit by suppliers that enter an abuse-prone market and exit before their behavior attracts the attention of program integrity or medical review staff, perhaps subsequently reopening under a new name and billing number. If so, then MSAs with higher proportions of inactive suppliers would be also expected to have higher levels of fraud and abuse. A high proportion of inactive suppliers may also reflect the exit of suppliers from markets that have become unprofitable, perhaps due to program changes or other efforts to combat fraud and abuse, so that MSAs with higher proportions of inactive suppliers tend to have lower levels of fraud and abuse.

While our small-area studies do not directly identify suppliers engaged in fraudulent activity, these analyses may be helpful in identifying selected MSAs for more intensive pre-payment review or other types of preventive efforts, and serve as a complement to reports generated by the SADMERC, which performs statistical analyses of DMEPOS claims for the Medicare contractors responsible for processing those claims. Their reports focus on analyses at the level of the individual policy-related group, supplier, or prescribing physician, but do not analyze MSA-level differences in costs and utilization.

III. RESULTS

Our discussion of results is divided into two sections. First, we describe the observed MSA-level variation in allowed charges per Medicare beneficiary. Areas with aberrant cost and utilization patterns are likely to have greater levels of fraud and abuse, and may have a higher payoff to increased program integrity efforts. Second, we discuss the results of our multivariate regression analysis, focusing on the relationship between measures of supplier characteristics believed to be associated with fraud and abuse and allowed charges per beneficiary across MSAs.

MSA Variations in Costs and Utilization Patterns
We found large variations in allowed charges per Medicare beneficiary across MSAs. In the four DMERC regions, allowed charges per beneficiary were 42% higher in the highest region than for the second highest region. There was considerable inter-MSA variation in charges per beneficiary. Allowed charges per beneficiary in one MSA were $812, which was more than twice as high as for any other MSA.

High charges per Medicare beneficiary at the MSA level can result from either higher charges for the subset of the Medicare population that use DMEPOS, or high utilization rates (the percentage of an MSA's Medicare population that use DMEPOS). The high charges observed for the MSA with the nation's highest charges resulted from both of these factors. Its charges per user were more than double those of any other MSA. Nationwide, about 15% of the nation's Medicare population had allowed charges for DMEPOS items, but some MSAs had utilization rates of nearly 30%.

Relationship Between Abuse-Indicator Variables and MSA Cost/Utilization Patterns
Our multivariate regression models explore the relationship between cost/utilization patterns and supplier measures believed to be associated with abuse of the DMEPOS benefit. Table 1 reports the impact on MSA charges per beneficiary implied by the regression coefficients on our supplier measures believed to be related to abuse of the

DMEPOS benefit.

Table 1
Relationship Between 'Abuse Indicator' Variables and MSA Cost Patterns

Variable	Mean	Std Dev	Effect of 1 Std Dev Change on MSA Charges/Bene (%)
suppliers/1000 benes	2.28	0.614	+$22.70** (+16.6%)
% entering suppliers	.110	0.053	+ 7.44** (+5.5%)
% inactive suppliers	.156	0.076	+ 7.36** (+5.4%)
% charges denied	.225	0.040	- 5.46** (-4.0%)
% charges w/o NSC#	.012	0.016	+ 2.37 (+1.7%)

** Regression coefficient was statistically significant at the 5% level.

Overall, about 23% of the variation in MSA charges per beneficiary could be accounted for by differences in non-abuse related demographic and provider community characteristics. The explanatory power of the model increased to nearly 50% when 'abuse indicator' variables hypothesized to be correlated with fraud and abuse were added to the model.

Several 'abuse indicator' variables were consistently and significantly associated with MSA cost and utilization patterns. MSAs that had higher proportions of new suppliers, higher proportions of suppliers that exited from the MSA in 1994, or higher supplier-to-beneficiary ratios tended to have higher charges per beneficiary. These relationships held in models that examined charges across all policy related groups and for models that examined individual abuse-prone items. From Table 1, a one standard deviation increase in the percentage of suppliers entering an MSA was associated with a 5.5% increase in charges per Medicare beneficiary, holding other factors constant. A one standard deviation (0.614) increase in the number of suppliers per 1,000 Medicare beneficiaries was associated with a $22.70 (16.6%) increase in charges per Medicare beneficiary.

Our small-area studies have documented the existence of considerable MSA-level variation in DMEPOS charges and utilization patterns, both overall and within individual policy related groups. One MSA in particular had aberrant cost and utilization patterns for the enteral and nebulizer policy related groups.

Our multivariate regression analysis suggests that most of the observed difference between MSAs cannot be explained by differences in the composition of variances in the characteristics of the Medicare population across MSAs. A considerable amount of the observed variation in costs per beneficiary was associated with differences in the levels of the 'abuse indicator' variables. Our small-area analyses suggest a potentially important link between supplier entry patterns and abuse of the DMEPOS benefit. These analyses may be useful in targeting selected MSAs for more intensive pre-payment review or other program integrity efforts. They can help focus the limited resources available for program integrity activities to geographic areas where the largest payoffs, in terms of denied claims and program savings, are likely to be realized. The small-area analyses suggest 'abuse indicator' measures that are correlated with MSA cost and utilization patterns, and may be useful in developing a more proactive approach to program integrity activities.

IV. PREPARATION OF DELIVERABLES

Our customer is given a series of reports and maps that support and highlight our findings in this study. The state and county boundaries data sets supplied with SAS/GRAPH are used with PROC GMAP to produce all maps. The FSREPORT command, new in SAS/FSP Release 6.11, is used for final production of MSA tables. In the limited space available in this paper, we have included a few mapping examples in Exhibits 1-4. Our entry in the SUGI 22 Poster Section displays a more complete variety of tables and maps.

The regional map shown in Exhibit 1 was simply created from the states boundaries data set. The complete code to separate and ANNOTATE the regions appears as Exhibit 2. To produce the more complex MSA-level map shown in Exhibit 3, we begin with the SAS/GRAPH counties boundary file, and equate counties to MSAs via our own FORMAT library. The intervening county lines within MSAs are then erased using the GREMOVE procedure, then the result is run through the GPROJECT procedure to properly order the data points (see Exhibit 4). Our MSA map also attempts to show some restraint in the number of patterns used. A graphic with a dozen or more ranges in its legend makes no particular point to the audience and consumes too much of the printed area allotted to the map. It is best to arrange your data points in groups of no more than 5. After analyzing our data, we elected to present five groups in each map: the bottom three quartiles, the 75th to 90th percentile, and the top decile. The UNIVARIATE procedure is run on the analysis variable by MSA code, calculating Q1, median, etc., across MSAs.

Devices and Drivers
The system and device-dependent drivers used with SAS/GRAPH are as important as the data, and can be most frustrating. Abt uses SAS/GRAPH Release 6.11 on both Windows 3.1 (WIN32S) and our RS6000 AIX/UNIX environments. Our printers include various HP LaserJet 5si and HP1600C models, all with Postscript capability. The HPLJ3SI driver creates the map; the WINPRTC or WINPRTG driver replays and prints it. Finally, to coerce our maps into this WordPerfect for Windows 6.1 document, we constructed our own drivers, as detailed in the SAS Technical Note TS-252S: *Exporting SAS/GRAPH Output to Novell WordPerfect 6.1 and Novell Presentations 3.0 for Windows*. The bottom line is that you should read SAS manuals specific to your system, check recent tech support notes, run the GDEVICE and GTESTIT procedures, and experiment until you find the best combinations for your site.

REFERENCES

Brantley, Verna C. And Lintern, Helen L. (1996), "Helping the HELPLINE - SAS/AF to the Rescue," Proceedings of the Twenty-First Annual SAS Users Group International Conference, pp 1041-1046.

"Exporting SAS/GRAPH Output to Novell WordPerfect 6.1 and Novell Presentations 3.0 for Windows," (1996), SAS Technical Notes, TS-252S.

ACKNOWLEDGMENTS

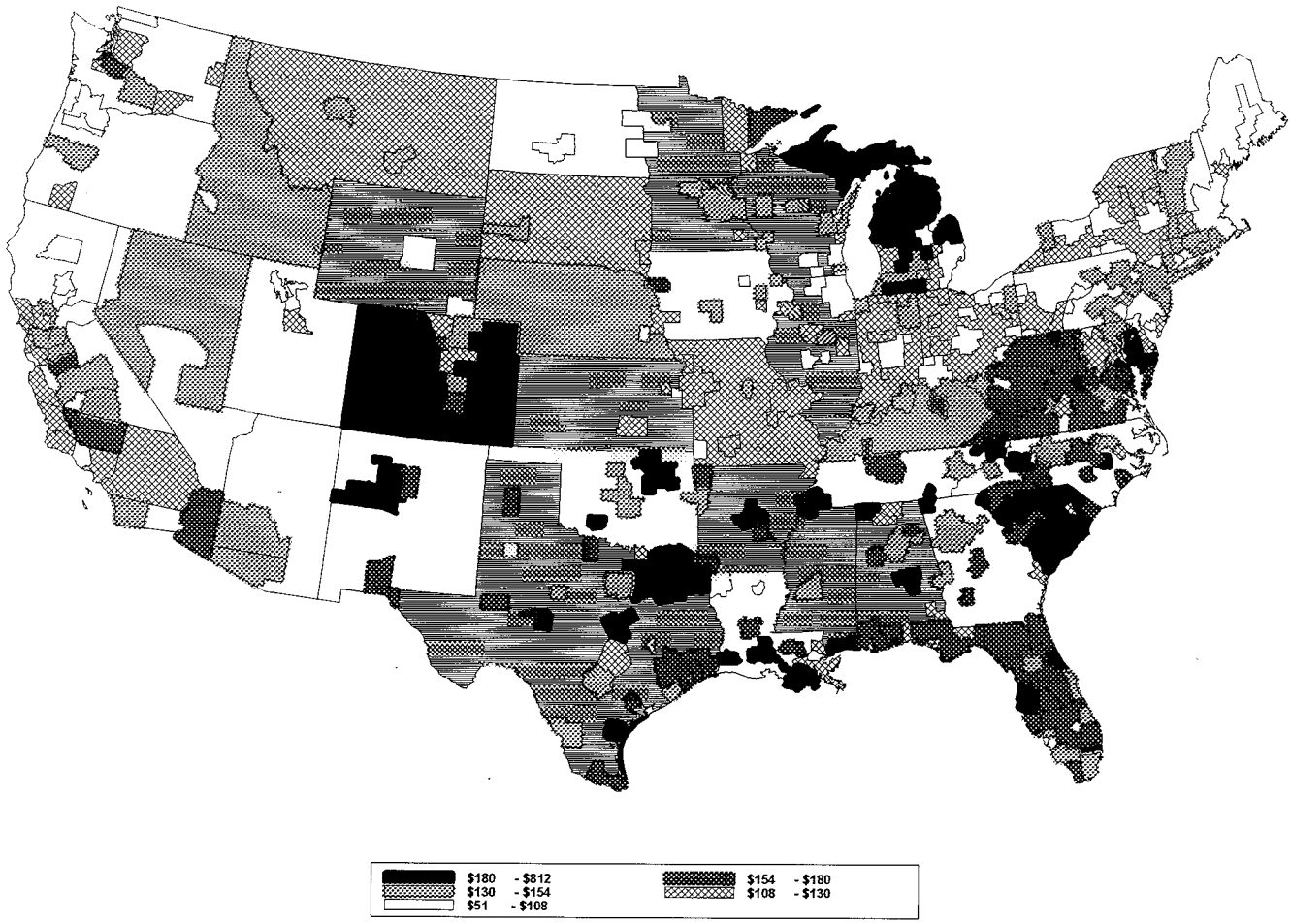
The authors wish to thank our project directors: Mike Edge at Palmetto GBA, and Leo Reardon at Abt Associates. A special thanks to our former colleagues, Ron Boheim and Bill Marder, who showed the way.

AUTHOR CONTACT

Mike Murphy
Abt Associates Inc
HSRE
55 Wheeler St
Cambridge MA 02138
mmurphy@world.std.com

```
*************************************************************************;
* EXHIBIT 2  (code for regional map Exhibit 1)
* Read SAS/GRAPH states coordinates file
* Assign DMERC regions using state-region format
* Output choro map of regions in con US
* 12/96, MM
*************************************************************************;
LIBNAME MAPS 'X:\SAS611A\MAPS';
LIBNAME DB  'C:\DATA\SUGI91';

%LET DEV   =WINPRTG;            /* HPLJS3, WINPRTG, WINPRTC, WIN*/
%LET PATTERN=GRAYCC;            /* GRAY22  OR  CYAN              */
%LET DISPLAY=DISPLAY;           /* DISPLAY OR  NODISPLAY         */

GOPTIONS RESET=ALL BORDER FTEXT=ZAPF DEVICE=&DEV  CBACK=WHITE
      ROTATE=LANDSCAPE &DISPLAY;
PATTERN1 VALUE=SOLID  COLOR=&PATTERN ;
PATTERN2 VALUE=SOLID  COLOR=&PATTERN ;
PATTERN3 VALUE=SOLID  COLOR=&PATTERN ;
PATTERN4 VALUE=SOLID  COLOR=&PATTERN ;

%INC 'G:\USERS\HSRE\SUGI22\ST2DMC.SAS';

*************************************************************************;
* Read SAS/GRAPH states coordinates file (keep Con US only)
* Equate state to DMERC region
* Output a response (region) file
* Output a state coord file and project it
*************************************************************************;
DATA COORD
   RESPONSE (KEEP=STATE DMERC);
  LENGTH DMERC $ 1;
  SET MAPS.STATES
    (WHERE=(STATE < 57 AND STATE NE 2 AND STATE NE 15));
  BY STATE;
  DMERC = PUT(STATE, ST2DMC.); /* STATE TO DMERC REGION */
  OUTPUT COORD;
  IF FIRST.STATE THEN OUTPUT RESPONSE;
  RUN;

PROC GPROJECT DATA=COORD OUT=PCOORD;
  ID STATE;
  RUN;

*************************************************************************;
* Separate the DMERC regions
*************************************************************************;
DATA PCOORD;
  SET PCOORD;
  SELECT (DMERC);
    WHEN ('A')  DO;
              X = X + .010;
              Y = Y + .020;
              END;
    WHEN ('B')  Y = Y + .010;
    WHEN ('C')  DO;
              X = X - .015;
              Y = Y - .030;
              END;
    WHEN ('D')  X = X - .020;
    OTHERWISE;
    END;
  RUN;

*************************************************************************;
* Create an annotate dataset to label the regions
* Compute the lat and lon centroids, label them, and project it
*************************************************************************;
PROC SUMMARY DATA=COORD NWAY;
  CLASS  DMERC;
  VAR   X Y;
  OUTPUT OUT=SUMDS (DROP=_TYPE_ _FREQ_)
      MIN  =MINX MINY
      RANGE=RANX RANY;
  RUN;

DATA ANNO;
  RETAIN FUNCTION 'LABEL' POSITION '3' XSYS '2' YSYS '2'
      SIZE 1.5  WHEN 'A' STATE 100;
  LENGTH TEXT $ 8;
  SET SUMDS;
  X = MINX + ( RANX / 2 );
  Y = MINY + ( RANY / 2 );
  TEXT = 'REGION '||DMERC;
  DROP MINX MINY RANX RANY;
  RUN;

PROC GPROJECT DATA=ANNO  OUT=PANNO;
  ID STATE;
  RUN;

*************************************************************************;
* Separate the DMERC regions
*************************************************************************;
DATA PANNO;
  SET PANNO;
  SELECT (DMERC);
    WHEN ('A')  DO;
              X = X + .010;
              Y = Y + .020;
              END;
    WHEN ('B')  Y = Y + .030;
    WHEN ('C')  DO;
              X = X - .015;
              Y = Y - .030;
              END;
    WHEN ('D')  DO;
              X = X - .050;
              Y = Y + .050;
              END;
    OTHERWISE;
    END;
  RUN;

*************************************************************************;
* Define titles and footnotes - generate da map
*************************************************************************;
  TITLE1          F=ZAPFB  H=1.2 J=CENTER COLOR=BLACK
 "EXHIBIT 1:  PROC GMAP DISPLAYING DMERC REGIONS" ;
  FOOTNOTE1 J=L C=BLACK  H=.8  'ABT ASSOCIATES INC'
        J=R C=BLACK  H=.8  "&SYSDATE";

PROC GMAP MAP=PCOORD DATA=RESPONSE ANNO=PANNO ALL;
  CHORO DMERC /
      DISCRETE NOLEGEND COUTLINE=BLACK CEMPTY  =BLACK;
  ID   STATE;
  RUN;
```

Exhibit 3: PROC GMAP Displaying Charges Per Beneficiary by MSA




```
*************************************************************************;
* EXHIBIT 4  (Partial code for Exhibit 3)
* Read SAS/GRAPH county coord file, and SADMERC analytical file
* Assign MSA codes from state/county
* Calculate percentiles of desired analysis var for pattern ranges
* 12/96, Louise
*************************************************************************;
LIBNAME MAP 'M:\SAS611A\MAPS';
LIBNAME LIB 'C:\SADMERC\PAPER';
LIBNAME LIBRARY 'C:\SADMERC\PAPER';

DATA COUNTY (DROP=STC) STATE (DROP=STC);
   SET MAP.COUNTY (WHERE=(1 LE STATE LE 56 AND NOT (STATE IN (2, 15))));
   LENGTH MSA_CODE $4 STC $5;
   STC = PUT(PUT(STATE,Z2.) || PUT(COUNTY,Z3.), $FIP2ARF.);
   MSA_CODE = PUT(STC, $ARF2MSA.);
   IF SUBSTR(MSA_CODE,1,2) NE '99' THEN OUTPUT COUNTY;
   ELSE IF SUBSTR(MSA_CODE,1,2) EQ '99' THEN OUTPUT STATE;

PROC SORT DATA=COUNTY;
   BY MSA_CODE STATE COUNTY;
PROC GREMOVE DATA=COUNTY OUT=COUNTY;
   BY MSA_CODE;
   ID STATE COUNTY;

PROC SORT DATA=STATE;
   BY MSA_CODE STATE COUNTY;
PROC GREMOVE DATA=STATE OUT=STATE;
   BY MSA_CODE;
   ID STATE COUNTY;

DATA MAP;
   SET STATE COUNTY;
   BY MSA_CODE;
   LENGTH DMERC $1;
   DMERC = PUT(MSA_CODE, $MSA2DMC.);
PROC GPROJECT DATA=MAP OUT=LIB.MSAMAP3Y;
   ID MSA_CODE;

 /*  DETERMINE PATTERN RANGES OF ANALYSIS VAR  */
LIBNAME DD 'G:\USERS\HSRE\SUGI21';
LIBNAME DAT 'G:\USERS\HSRE\SADMERC\DATA';

%MACRO MAKAFILE (ANALVAR,SHORT);

PROC UNIVARIATE DATA=DAT.MSA95 (KEEP=&ANALVAR) NOPRINT;
   VAR &ANALVAR;
   OUTPUT OUT=UNIV Q1=UQ1 MEDIAN=UMEDIAN Q3=UQ3 P90=UP90 MIN=UMIN
MAX=UMAX;

DATA DD.U_&SHORT;
   LENGTH ULEVEL $ 1;
   SET DAT.MSA95;
   IF _N_=1 THEN SET UNIV;
   IF (UMIN LE &ANALVAR LT UQ1)              THEN ULEVEL='5';
   ELSE IF (UQ1 LE &ANALVAR LT UMEDIAN)      THEN ULEVEL='4';
   ELSE IF (UMEDIAN LE &ANALVAR LT UQ3)      THEN ULEVEL='3';
   ELSE IF (UQ3 LE &ANALVAR LT UP90)         THEN ULEVEL='2';
   ELSE IF (&ANALVAR GE UP90)                THEN ULEVEL='1';
   MSA_CODE=BENE_MSA;
   LABEL ULEVEL="CATEGORIZED NATIONAL &ANALVAR."
      MSA_CODE='MSA';

PROC SORT;
   BY MSA_CODE;

%MEND;
%MAKAFILE(A_E_PC,A_E_PC);
```