

Using SAS to Process TIGER Line Files: A User's Viewpoint

Hendrik K. Meij, Population Research Institute
The Pennsylvania State University

Introduction

The ability to create detailed maps and spatial relationships has recently taken on new meaning with the introduction of the U.S. Census Tiger Line files. This capability is especially important in demographic research. This paper details the process of creating and processing on-line, detailed maps of Pennsylvania counties using SAS 6.07 tools available to a group of university demographers.

The Population Research Institute (PRI) is part of Penn State's Intercollege Research Programs. PRI includes thirty Research Associates, twenty-five graduate students, four postdoctoral candidates, and a professional computer staff with demographic interests in areas such as fertility, aging, and intergenerational relations. Recent research pursuits have focused on TIGER file processing and the establishment of in-house mapping capabilities. The need for these visualization techniques are myriad. The future direction of this type of demographic research focuses on two requirements:

- (1) The introduction of powerful desk top workstations provide the internal memory necessary to process TIGER Line files in a flexible, interactive, and versatile environment; and
- (2) The recent efforts of the U.S. Bureau of the Census and State Data Centers to provide a national, commonly referenced spatial coordinate system.

During Spring 1991, the author was tasked with evaluating the creation of in-house mapping with TIGER Line files. As a member of the PRI, three computing environments are available: individual microcomputers, the university sponsored IBM ES/3090 600S mainframe using CMS and MVS, and a UNIX based network, the PIRCNET operating in SunOs (4.1.1). Connectivity between the three environments is made possible through Penn State's high speed fibre-optic data backbone. The PIRCNET currently uses SAS* Version 6.07 and includes base SAS*, SAS/ASSIST*, SAS/CONNECT*, SAS/ETS*, SAS/IML*, SAS/INSIGHT*, and SAS/STAT*. A 400MB disk partition for SAS temporary work files is allocated to the primary SAS network server, a SPARCStation 1+ with 40 MB of RAM. The university mainframe uses SAS 6.06 and SAS 5.18 in MVS and CMS environments.

Software Selection

Very early in the software selection stage it became evident that there were two bifurcate paths to follow:

- (1) to pursue Geographic Information Systems (GIS) capable of handling the mapping tasks; or
- (2) to designate the tasks of mapping to SAS/Graph.

The latter option was primarily chosen because it takes full advantage of the network configuration at PRL. Contact with John Blodgett from the Urban Information Center (UIC), University of Missouri, St Louis, MO, resulted in information about and assistance with SAS source code to access the TIGER Line files. UIC has developed SAS source code to access the Census Bureau's TIGER Line files, and process the information into SAS data libraries. The SAS data libraries (SDL) are subsequently accessed to generate polygon boundary data sets suitable for input into PROC GMAP. The software available from the UIC was selected to create the capacity to generate maps from TIGER Line files. PRI Associates, graduate students, and staff require this data for the entire state of Pennsylvania at resolutions below county level, including census tracts and minor civil division geographic areas.

The SAS source code (macros) are publicly available via File Transfer Protocol (ftp) from UMMVSB.UMSYSTEM.EDU (128.206.100.100). For a detailed description on the functions of the individual macros, consult the papers by Blodgett in the proceedings of SUGI 15 (Nashville) and MWSUG '91 (Chicago). It is not the intent of this paper to describe the workings of the macros, but rather to highlight the author's approach. The SAS source code previously discussed can be retrieved from the following directories on UMMVSB.UMSYSTEM.EDU; \$6386.TIGRDEMO.SOURCE (contains the original source code on demo diskette), \$6386.TIGRJOBS.CNTL (contains various utility programs), and \$6386.PUBLIC.SASMACRO (contains the %TIGRLN and %TIGRLNK macros).

Creating the SAS Data Libraries

As a first step, three counties in PA (Centre, Huntingdon, and Mifflin) were processed. After successfully accessing and mapping these counties, a compiling plan was designed to process all 67 Pennsylvania counties. The major components of this plan included: (1) an MVS session, (2) a transport and restore session, (3) a multiple chaining session, and (4) a user application session. The next section describes each of these sessions.

The MVS Session

In accessing the Tiger Line files for Pennsylvania, it immediately became evident that the files would have to be sorted. The appropriate sort order is documented in %TIGERLN (SORT FIELDS=(6,10,CH,A,2,4,CH,A)). The JCL for sorting has been added to the front of the

%TIGERLN macro. After the files had been sorted, the %TIGERLN and %TIGRLNK macros were executed in series within a single MVS batch job. Upon successful completion of the MVS job, the newly created SAS TIGER SDLs were written to 3480 tape cartridge. The chaining of the macros greatly reduced logistical problems in processing, making the submission of these programs under nightly-batch modes possible (for reduced processing costs). The size of the SASLIB format library, as documented in %TIGRLN macro, was increased to accommodate Pennsylvania data. The following parameter was used: SPACE=(CYL,(60,60,6),RLSE).

About two dozen TIGER SDL libraries fit comfortably onto a 3480 cartridge. Only two Pennsylvania counties required disk space allocation increases to accommodate the data files: Allegheny County (Pittsburgh area) and Philadelphia County. The space parameter, SPACE=(CYL,(50,10)RLSE,,ROUND) was used for these two counties. All space allocation parameters refer to a 3380 minidisk (one cylinder equals approximately 0.72 MB).

The Transport and Restore Session

About a dozen TIGER SDL files were transported in a single MVS run. Sequential transport files were written using the SASV5XPT engine during a PROC XCOPY procedure with the EXPORT option. Part of the JCL is presented here:

```

...
// EXEC SAS606
//TIGER1 DD DSN=TIGER90.C42001,
// UNIT=TPCART,VOL=SER=XXXXXXX,LABEL=(1,SL)
//TIGER2 DD DSN=TIGER90.C42003,
// UNIT=TPCART,VOL=SER=XXXXXXX,LABEL=(2,SL)
...
/*OUTPUT VM DEST=PSUVM.HKM900,FORMS=XCPY
//TOVM1 DD SYSOUT=(N,,VM),
// DCB=(RECFM=F,LRECL=80,BLKSIZE=80),FREE=CLOSE
/*OUTPUT VM DEST=PSUVM.HKM900,FORMS=XCPY
//TOVM2 DD SYSOUT=(N,,VM),
// DCB=(RECFM=F,LRECL=80,BLKSIZE=80),FREE=CLOSE
...
//SYSIN DD *
OPTIONS NOCENTER;
PROC XCOPY IN=TIGER1 OUT=TOVM1 EXPORT;
PROC XCOPY IN=TIGER2 OUT=TOVM2 EXPORT;
...
RUN;
```

This program punches binary sequential transport files back to the designated CMS reader list defined in DEST=field. The files were then received (usually in batches of 3-6) on a 100MB temporary CMS disk. Using the ftp binary option, the files were ported to the PRI's UNIX-based LAN, PIRCNET. This method worked very efficiently. The slowest link was the file transfer

from the CMS reader to CMS temporary disk. A 60MB file might take 45 minutes on this link, yet, be transferred through the fibre-optic data backbone to the UNIX LAN in less than 10 minutes. The restore program simply pointed to the transported sequential files and performed a (reverse) PROC COPY using the SASV5XPT engine assigned in a LIBNAME statement.

The Multiple Chaining Session.

From the runs of the experimental counties, it was evident that the GREMOVES.SAS program was not as simple to apply as anticipated. In general, if one chains the polygons at census block levels, 'holes' are left when trying to aggregate the blocks in block groups. It was decided early on, by chaining three counties at all levels, that the desired levels for chaining included: blocks, block groups, tracts, mcds, and county. Unfortunately, the zip code parameter resulted in very poor chaining rates (25% success rates) and were dropped from the list. The chaining procedure started with chaining a county at the block level in the background (multi-tasking) mode. While this process was running, the other levels were chained in the foreground on the same machine. On average, all levels completed processing before the block level process ended. In a final step, the first page of each listing was echoed to the console (to check the Summary Table of % Successfully Chained versus Failures). To safeguard against accidental mislabeling of the files, file permission bits were adjusted to remove all write access to the newly created files. The files were then reduced in size by a factor of 40-65% using the UNIX compress utility. All data files were written to a 150MB disk partition. The rate of successful chaining was remarkable. In general, a typical county chained at 100% at the block level. The estimated number of not-so-perfect chained counties was around 30. Of these, the success rate remained within the 97-99% range. Never has the author seen a successful chain rate below the 97% mark. Unfortunately, the author did not have the foresight to have the results of the summary table appended to a separate file.

The only parameter changed during the chaining process was the MAXPTS= field in the CHAIN.SAS program. This parameter was increased from the default of 250 to 500. All programs executed flawlessly in SAS Version 6.07 under SunOS 4.1.1.

The final step in processing a county included a backup of all county related TIGER datasets (tiger1 through tiger7) to 150MB tapes. The UNIX tar (tape archival) utility was used to collect all county tiger files into a single archive file. The archive file was then compressed, and subsequently copied to tape. Only the polygon boundary files are currently stored on disk.

One of the anticipated problems in chaining 1990 Tiger Line files is their incompatibility with 1980 geographic identification labels. Identification labels for the 1990 Minor Civil Division are unique within the state, while 1980 labels are only unique within the county. An attempt was made to generate a match file for 1990 and 1980 identifiers. The tiger3 SAS dataset contains the 1990 and 1980 labels. A short program was written which searched the tiger3 data files and appended the codes to an ascii file.

The User Application Session.

Following the complete processing of each county file, a mapping program was executed to view the results in the SAS Graphic Editor. The Graphic Editor's zoom feature was extensively used to check the exactness of merging lines. Overall, the plotted images had a very high quality with very few inconsistencies. In fact, it will require detailed local knowledge to find these inconsistencies. Enormous blow ups of city sections in several counties did not generate much criticism when viewed by local experts.

The user application section has not yet been fully developed. Two problems are still faced:

- (1) PROC GREMOVE (GREMOVES.SAS) is still puzzling in custom creating new geographic areas (for example, School Districts from aggregates of Minor Civil Division areas); and
- (2) An authoritative "match file" has not yet been created.

In addition, there was one missing county (due to a mixup in county files of posted Tiger Line files). Several utility macros were also applied and executed flawlessly. Included in this list are GENLABS.SAS (it calculates centroids for desired geographic areas and anchors a label to this centroid) and GOUTLINE.SAS (it generates different line characteristics for different polygon boundaries).

Figures 1-4 demonstrate some results from this project. Figures 1 and 2 display the different juxtapositional relationships of the census tracts versus the Minor Civil Division geographic areas. The census tracts' distribution is concentrated in large metropolitan areas (Pittsburgh, Philadelphia), and sparse in rural areas. The Minor Civil Divisions' distribution covers rural areas more uniformly than tracts, yet aggregates large cities such as Pittsburgh and Philadelphia into individual units. The results of GOUTLINE.SAS are displayed in Figure 2. The county borders have been emphasized for spatial referencing purposes for individuals not familiar with Pennsylvania.

Figure 3 displays the census blocks of a single county (Centre County is located in the center of Figure 2). Interstate highway I-80 cuts through the top quarter of this county. A detailed section of this county is shown in Figure 4. The images were generated using PROC GMAP and saved to SAS graphic catalogs. The templates were designed with PROC GREPLAY in full screen mode. Titles and other custom work were performed using the Graphic Editor, editing the image in the final template created with PROC GREPLAY.

Future Directions

The complex task of accessing large public data files, such as the TIGER Line files, was made much easier through the use of publicly available SAS software macros. These macros quickly generated polygon boundary files for input to the GMAP procedure. Since each polygon at any level of resolution is uniquely identifiable, spatial data referencing these labels can be easily

displayed. The anticipated use of these maps will require posting the boundary files in a publicly readable file system. The author's intent is to add two additional layers to these boundary files:

- (1) A layer of hydrological features (rivers and lakes) to provide delineation of watersheds; and
- (2) A layer of infrastructure (major roads and highways) to provide a transportation layer.

The approach taken in using the MVS environment will be replaced by mounting the TIGER Line files as file systems in the UNIX environment through the availability of CD-ROM media. This storage medium will greatly enhance TIGER Line processing flexibility. One speculative approach is to process the %TIGERLN and %TIGERLNK macros in sequence, and if successful, continue the same job with the CHAIN.SAS module at the desired geographic levels. The availability of the CD-ROM media through the university library system also implies access to geographic areas of particular interest outside Pennsylvania.

One major problem for users of these polygon boundary files will be correctly matching data at other levels (for example historic data) to the identification labels used in the 1990 TIGER Line files. A series of carefully compiled match files for common merge operations would fill a big void. Interest in geocoding and the linking of the results to the spatial grid system underlying the maps is anticipated.

NOTES

The author is greatly indebted to John Blodgett, Office of Computing, University of Missouri St Louis, for the help and advice provided. John Blodgett also maintains the SASPAC-L list service. To subscribe to this list send an electronic message to `LISTSERV@UMSLVMA.BITNET` containing the request: `SUBSCRIBE SASPAC-L username`.

For comments or questions regarding this paper contact the author via e-mail at `meij@darwin.psu.edu` or voice at (814) 863-7375.

*SAS is a registered trademark of SAS Institute, Inc., Cary, North Carolina, U.S.A.

All mentioned products are not to the exclusion of other equivalent products currently available.

Figure 1. Census Tracts for the State of PA.

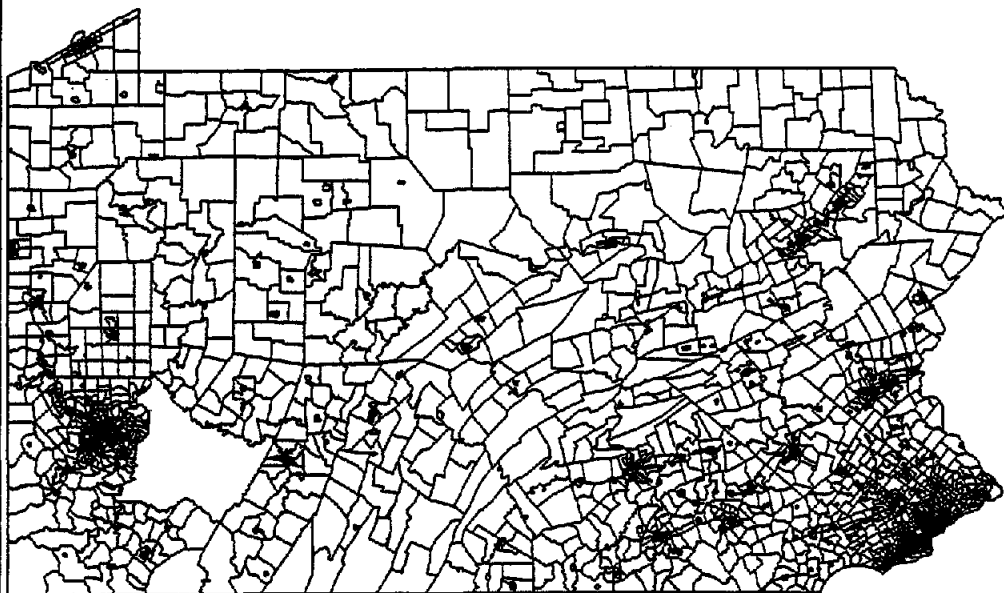


Figure 2. Census MCDs for the State of PA.

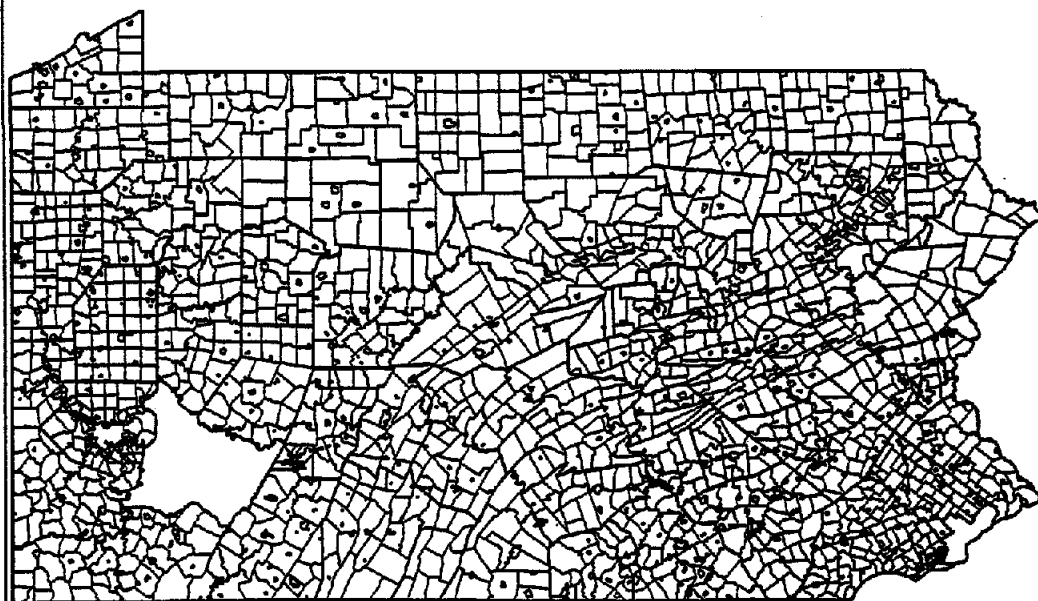
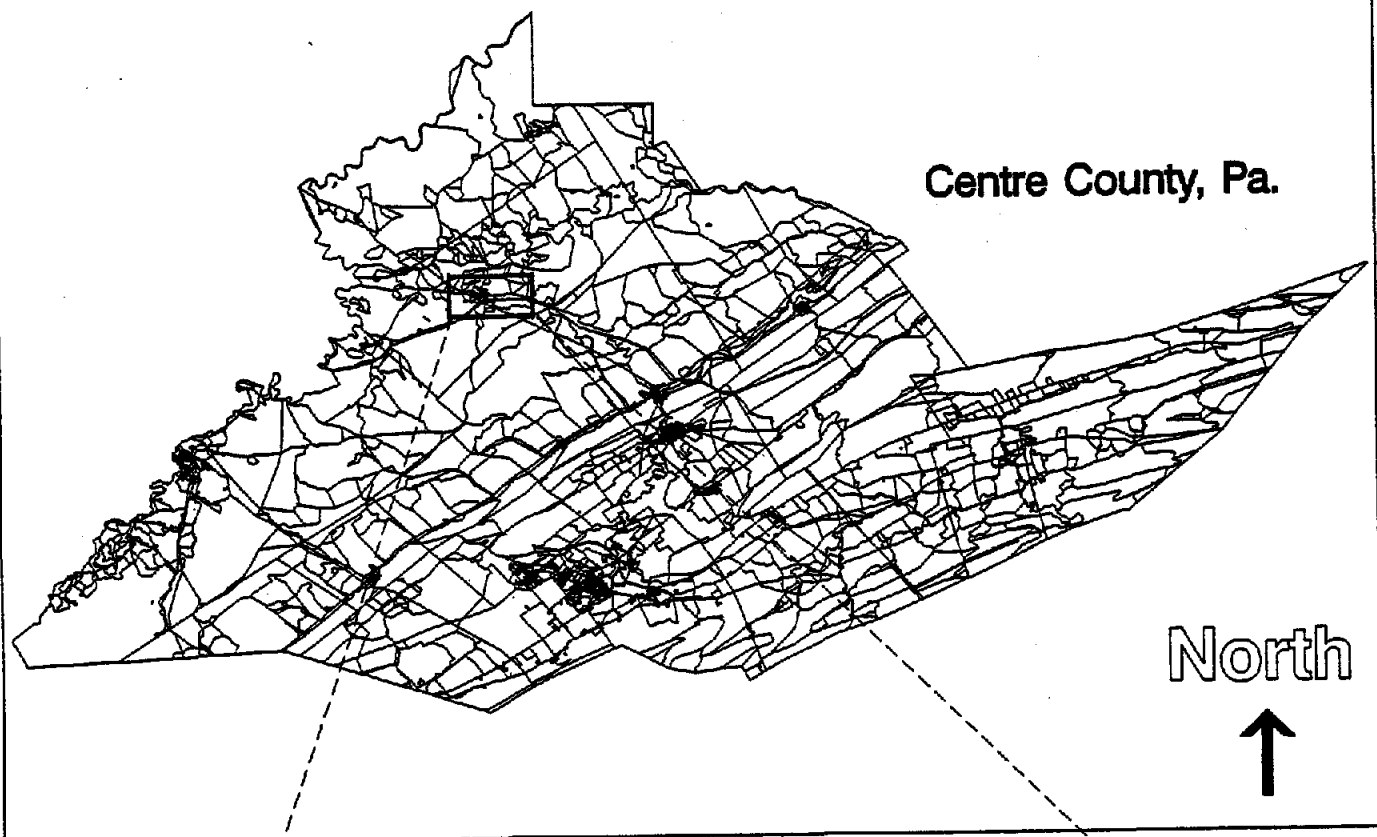


Figure 3. Census Blocks, Major Highways, and Hydrologic Features.



Snowshoe, Pa.

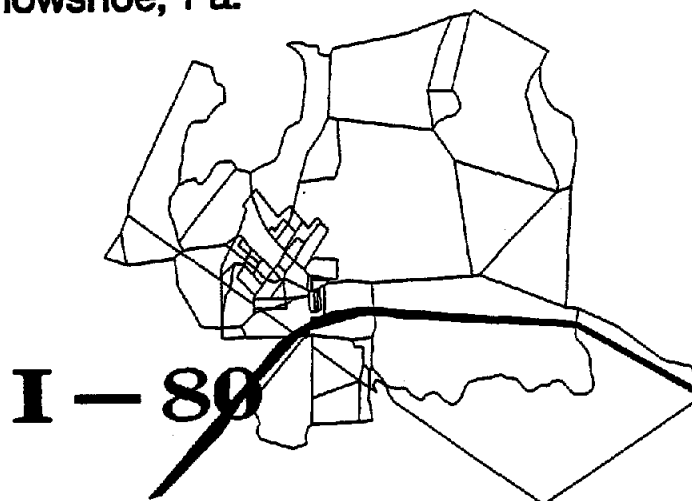


Figure 4. Detailed Area of Above.