

An Advance in Geographic Information Systems using SAS® Software:

Address level Lat-Longing

Doug Surch: Audience Identification Inc.

For those of us who remember Big Bird, Cookie Monster and Oscar the Grouch, a way to receive and digest concepts in short bites is automatic and almost unconscious. In seven second bits, the alphabet was learned, we counted to 15 and learned that nice people come in many styles, shapes and colors. The adolescents now skip Sesame Street but have replaced it with MTV, again a series of images, flickering half thoughts, but with music added. As adults, we get more news stories with less time allotted to each to keep our quest for knowledge in small, concise bites. Even as a higher intelligent group, or data processors, we have divided the second in billionths, and made the nanosecond a household word.

Now for a scary reality check. The people that pay and use our talents are also participants in the Rat Race, possibly even the front runners. Getting their time is precious and few, thus we are required to play in this tyranny of time arena. Our objective then becomes the creative way to setup our messages so it can fit between the fax'd snippet and the car phone conversation, but have the retention far beyond both. If a picture is worth a thousand words, would a picture, richly loaded with information, but simple in appearance, compressing our SAS output provide the medium? I think it can and will continue to be a good bet.

In the marketing discipline, as well as many others, the identification of an audience, group or micro universe is required. If a segment can be identified, further planning, marketing, analysis and research is primed and ready. Everybody has a name and address file and some characteristics about their file. Mix this with a map containing streets, highways and states they already know and we are on our way to the memorable idea presentation.

Now let's take a time out and review our components so far. We have some people to define, a short time to shine and decided a complex but simple picture should do the job. RELAX! The idea and process to accomplish this can be done with SAS, some government information and a little creativity.

Most uses to date have involved thematic mapping, polygon definition or zip centroid usage. With the explosion and

availability of data, the ability to combine data and concepts to create a new level of mapping is present. Prior issues of SAS Communications have demonstrated how to draw maps and compute best routes from point to point. To reach our next level, a specific latitude and longitude, with detailed precision, is required. Once we have this precise location, the ability for distance calculations within yards, improved thematic mapping and territory definition is possible.

The process to do this requires the use of US Census Bureau Tiger Files and a clients' file. How to clean, compress and create workable versions of each is shown. After this is laid out, the plottable file is input into different mapping streams to create our "Complexly Simple" picture.

The 1990 Census Tiger/Line files are an extract of selected geographic and cartographic information from the Census Bureau data base. Their purpose is to provide users with the 1990 Census Boundaries and support the 1990 Census Data Products program. The file is in the form of 12 record types. Our application will only use 5. These are:

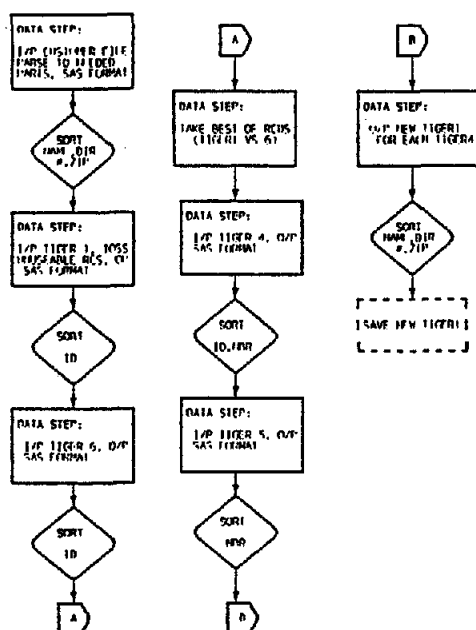
- a) Type 1 - Basic Data Records (individual feature segments)
- b) Type 2 - Shape Coordinate Points (feature shapes)
- c) Type 4 - Index to Alternate Feature Name
- d) Type 5 - Feature Name list
- e) Type 6 - Additional Address Range & Zip Code information

Tiger/Line File :					
COUNTY	RECORDS WITH DIRECTION INDICATOR	RECORDS WITH NAME	RECORDS WITH TYPE INDICATOR	RECORDS WITH ZIPCODE	RECORDS WITH ADDRESS NUMBER
Allegheny County, PA	3.5%	84.1%	76.8%	54.1%	54.1%
Westmoreland County, PA	2.0%	69.1%	46.0%	8.5%	8.5%
Allen County, IN	10.0%	84.4%	70.9%	49.1%	49.1%
Hillsborough County, FL	23.4%	78.9%	70.8%	41.5%	41.5%
Polk County, FL	10.5%	73.9%	61.6%	21.1%	20.2%
Kendall County, IL	3.9%	84.3%	51.6%	5.5%	5.3%
LaSalle County, IL	15.6%	63.5%	46.2%	0.0%	0.0%
Dubuque County, IA	6.2%	68.0%	53.9%	21.9%	21.9%
Jackson County, IA	4.4%	32.8%	15.5%	0.0%	0.0%

The average file size for a state is 500 megabytes (range from 4 to 1800 megabytes). The average file size for a county is 7.5 megabytes (range from less than 1 to 125 megabytes). The entire file is 25,000 megabytes. The file is grouped at the county level. Above is a chart reflecting the various conditions of selected counties. From this chart, we can see the need to improve our input data.

To improve processing, a single file would be the easiest with which to work with. The logic to combine these files and create a smaller file is explained below, with a diagram following. To both files, we need to make certain that upper case will be compared to upper case. From the Tiger1 file, only those records with names and addresses can be used. We also need to determine right or left side of the street in order to do calculations. The Tiger6 contains data to enhance the Tiger1. After parsing the Tiger1, the data in Tiger6 is used to replace data in Tiger1 if it is more relevant and revealing. The Tiger4 contains alternate names for the file. As this may provide a match alternative to the client file, as many names for the same street is desired, each contained on their own record. Finally, a sort to setup matching is required. Each file has a descending parameter for the street direction and zip code. This is imperative to our matching logic, explained after the flow chart below.

BUILD A BETTER TIGER



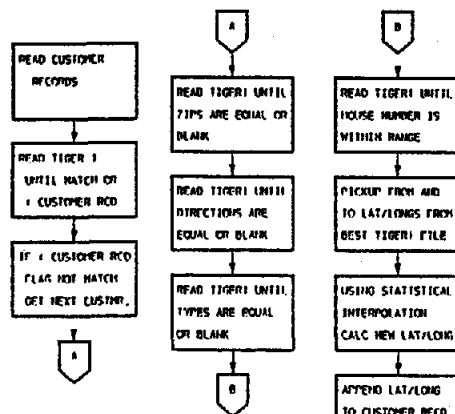
Our next step is to clean and parse our client file. We need to break apart the address into street direction, street name and street type to allow matching to Tiger1. Existing software exist to perform this function and also standardize and correct the spelling of street name, city and state. Although it may seem straight forward, the combinations are a nightmare. Software is available on all platforms to perform this function. Our choice was to contract this out to a firm that handles this as part of their regular business.

Time out again, let's take a breather. We now have the Census files combined and enhanced into a single source which contains the best of the various other records. We also have the clients file, broken apart, corrected and parsed down to our required components. Now lets jump back in and calculate the latitude and longitude.

The tricky part of matching these files relates to our original concern. The Tiger files were not setup for calculations using street name and addresses. Although we have cleaned up the file, getting close at times is the best we can do. A benefit for this is that the files are county specific. When matching, we try for a zip code and name match. If unable to match zip, we accept a missing zip as a match. After a name is matched, the best direction and type is used. Our earlier sort by descending direction and type allows us to try non missing fields first, defaulting to accepting missing as a match. The Tiger from and to latitude are then used along with the client from and to addresses, mixed with the left or right side to extrapolate the exact coordinate. The Tiger files have 6 digit precision, adequate for our calculations. Those not matching can be fed into another program, identical to this but with looser criteria, such as a partial name matching and street number projection. This was not done, but testing showed some credible improvements and matching could be found. The problem was time and cost verse the added benefit. Shown below is a diagram of the data step logic.

APPENDING THE LAT/LONG

(LOGIC WITHIN DATA STEP)



The match rate varies across areas. Addresses that contain rural routes, military addresses and post office boxes will not match. Those that contain street numbers within the name are also matched less than expected due to combinations of spelling, such as Hwy 1, State Hwy 1, H-1 and other combination. Below is the output of frequencies done on varying counties, some urban, some metropolitan and some rural reflecting the matching rate. As can be seen, if you hit a updated county, then matching is great, but for some areas, it may not be viable to continue.

LAT/LONG APPEND RESULTS					
COUNTY	TYPE 1 TIGER RECORDS	USEABLE TIGER RECORDS	CLIENT RECORDS INPUT	NUMBER RECORDS MATCHED	MATCH RATE
Allegheny County, PA	88,132	74,225	4,350	3,463	79.06%
Westmoreland County, PA	41,529	31,929	0	0	0.00%
Allen County, IN	23,850	18,963	8,407	5,063	79.02%
Hillsborough County, FL	62,068	52,777	0	0	0.00%
Polk County, FL	46,515	37,531	2,747	1,780	64.80%
Kendall County, IL	6,595	4,479	378	153	40.48%
LaSalle County, IL	25,399	17,521	688	292	43.71%
Dubuque County, IA	12,637	8,924	729	539	73.94%
Jackson County, IA	10,310	5,844	127	28	22.05%
TOTAL	317,035	252,193	15,438	11,318	73.32%

The PROC GMAP was not used due to our requirements for a radius, which could not be programmed with GMAP. The DGSI was not used because of our intention to setup a batch type macro and interface within the program as little as possible. The PROC PLOT was used instead. The process is described below and diagramed following this description.

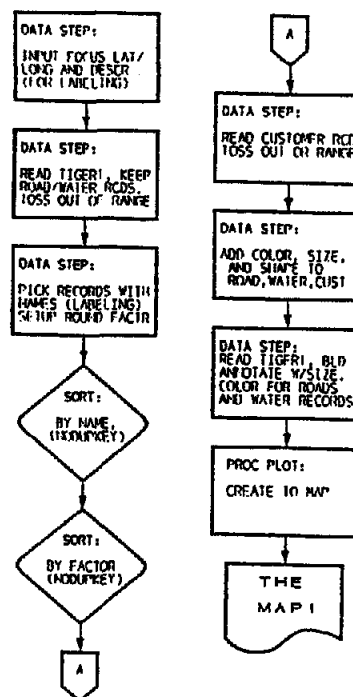
In order to parse data outside the range, the following formula was used. This allows us to discard records with both from and to latitude/longitudes outside our mapping area.

```

DISTANCE = 3949.99 * ARCOS (
  SIN(LAT1) * SIN(LAT2) +
  COS(LAT1) * COS(LAT2) *
  COS(LONG1-LONG2)
);
  
```

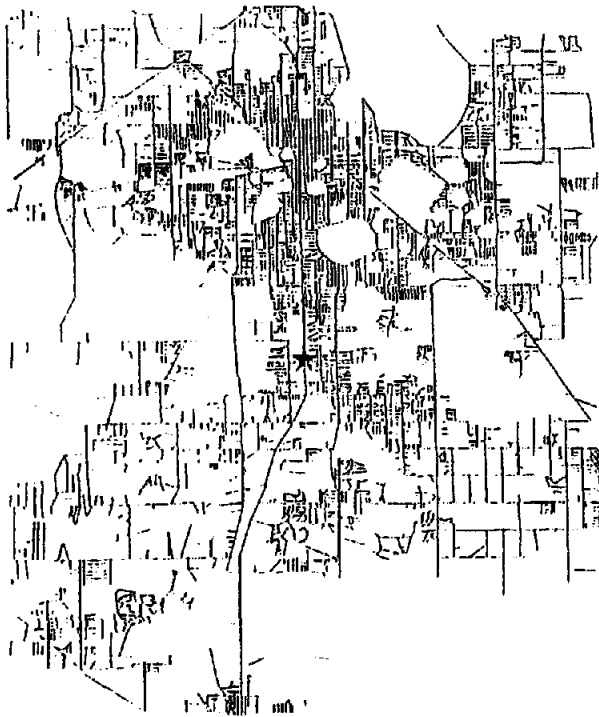
The program flow for the map is shown below. Special processing imbedded in the flow helps to create street labels and special features extracted from the Tiger files. The street labels cannot use every label. Description overlap makes the map unreadable when trying to get street labels, river labels and special features on the page without covering the client mapped responses. Documented in the Tiger files are special features. Although they were seldom present in our tested counties, logic was incorporated to intercept them, flag their point and description and setup for the map.

DOING THE MAPPING



Third quarter, home team slight advantage, let's take another time out to recap. The cleaned Tiger and client files were matched together. The output was an unmatched file and a matched file with latitude and longitude appended. Our next step is to create a map, using these data elements. Selecting a focus point, or retail outlet for example, we want to keep all customer points and Tiger data within a given radius. This subset of data will allow us to create our memorable map.

Example 1



Example 1:

```
Style = Centx1, Size = 1.65
Position = 8
Centered: Focus Longitude coordinate
          Minimum Latitude coordinate
```

Example 2:

Road Labeling:	Road Widths:
Style = Centx1	Default = 0.5
Size = 1.18	A1 = 5.75
Position = 5	A2 = 3.25

- 1) Outlet mapped, no annotation.
- 2) Road and water labeling added, Road and Water width set by type of road.
- 3) Responses mapped, showing two types of notation.
- 4) Location of choice spots marked, showing font use variations.

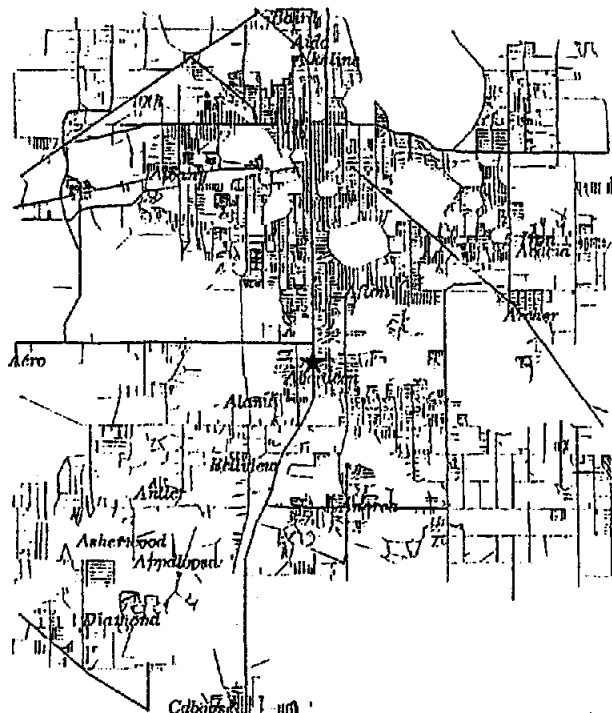
Example 1:

Using Tiger Records:

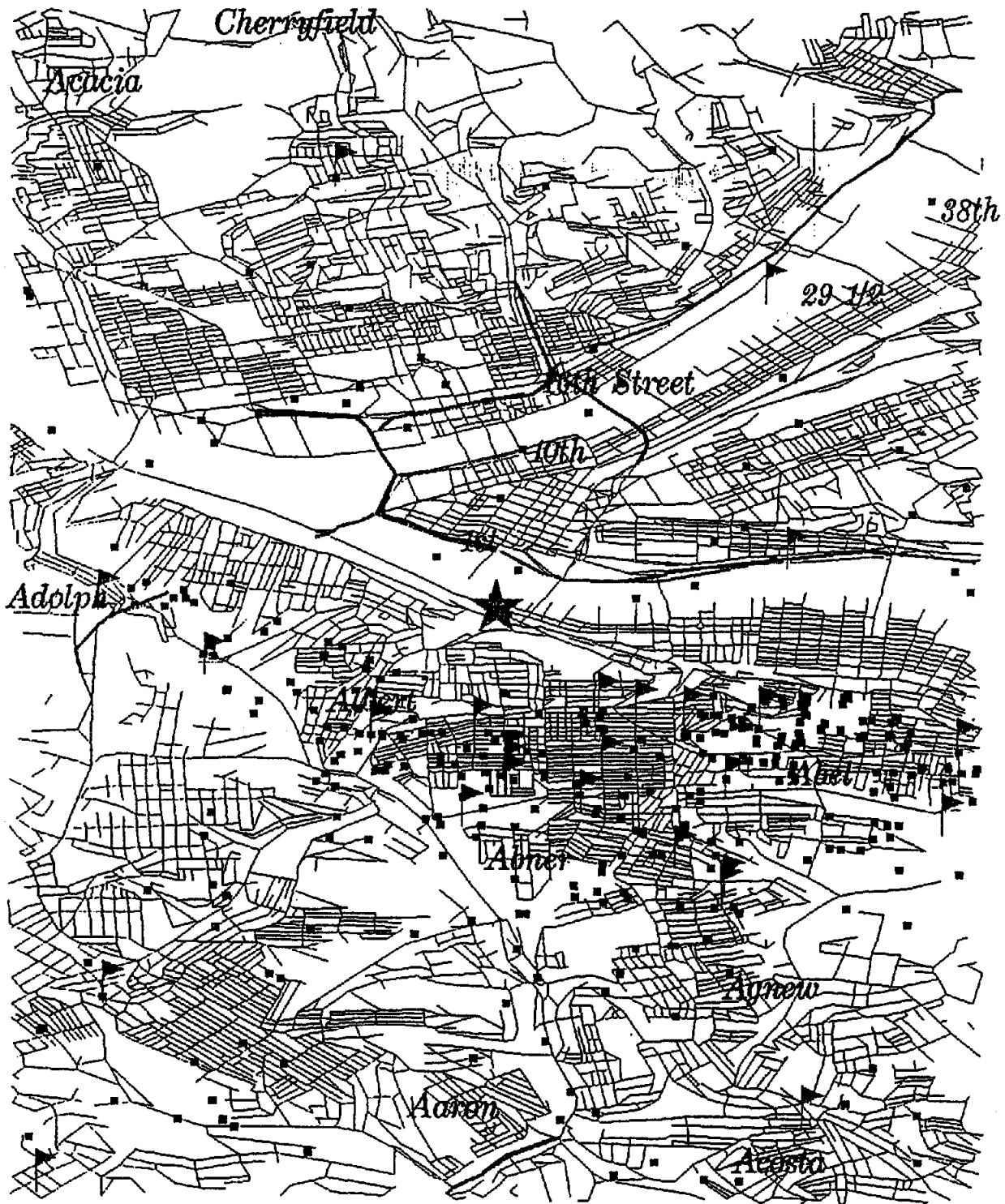
- A1: Primary roads, Interstates
- A2: Secondary roads, US Highways
- A3: Connecting and county roads
- A4: Neighborhood and city roads

Focus Point (Star):
Style = Marker
Text = V
Size = 2.2

Example #2



Example 3



FIPS Code: 42003 (Allegheny County, Pa.)

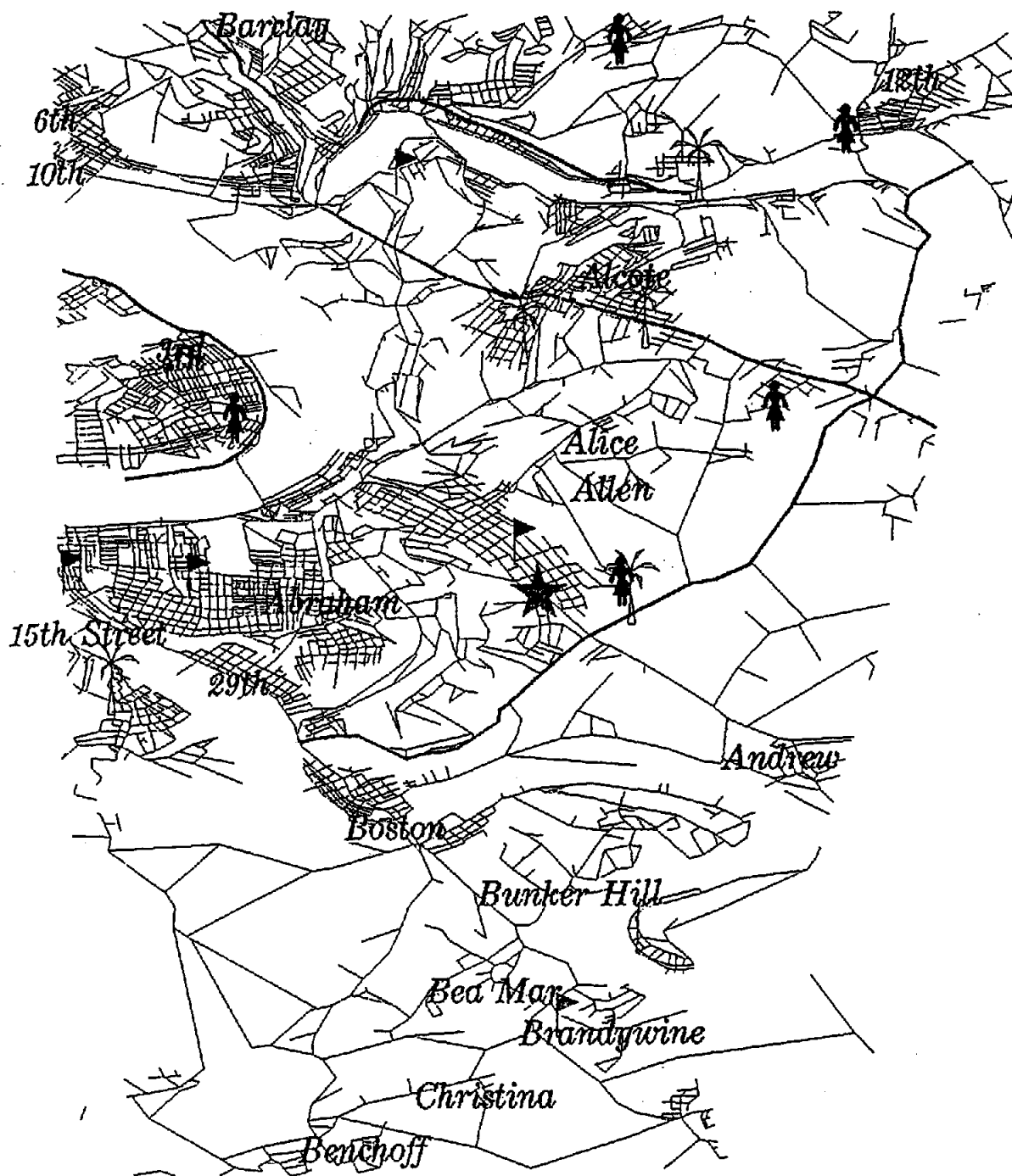
Star indicates client outlet location
 ADDRESS: 13 Carson Street
 McKeesport, PA 15132

Longitude: 80:000900
 Latitude: 40:431100

Squares indicate type A customer, Flags indicate type B customer

Favorite Spots

Example 4



↑ GOLF COURSES ★ NIGHT CLUBS † SWIMMING POOLS

SAS is a registered trademark or trademark of SAS Institute INC.
in the USA and other countries. * indicates USA registration.

Douglas Surch
AUDIENCE IDENTIFICATION INC.
645 Blackhawk Drive
Westmont IL 60559
(708) 654-8414