Detailed Procedures for Geo-Coding Health ABC Study

Source of Health ABC Address Files

- Pittsburgh Site
  - Diane Ives (reviewed IRB approval at Pittsburgh site)
  - Janet Fair (assisted with acquiring address file)
- Memphis Site
  - Suzanne Satterfield (reviewed IRB approval at Memphis site)
  - Margaret Caufield (assisted with acquiring address file)

What is Geocoding?

- Geocoding Health ABC by using participants' home addresses, researchers will be able to produce measures of environmental exposures. Geocoding identifies the longitude and latitude coordinates of a participant's place of residence. The X- and Y-coordinates can then be used to link a participant's geographical location with their US Census Bureau block polygon. The block polygon is the smallest geographical administrative unit available in the USA. In subsequent procedures, the polygon centroid of the block can be used to create measures of the "residential environment".

Geo-Coder

- Carlos Siordia PhD, MS
  - Has taken 12 hours of graduate-level hours in advance GIS techniques
  - Currently an Associate Editor for the *Journal of Maps: Social Science Section*
  - Has published multiple ArcGIS related manuscripts including the following:

Siordia C & Matthews SA. 2015. Extending the Boundaries of Place. Recapturing Space: New Middle-Range Theory In Spatial Demography. Volume 1, Spatial Demography Series. Dordrecht: Springer. Forthcoming.
Siordia C, Smith C, & Castañeda E. Geographically-Aware Multilevel Analysis on the Association between Atmospheric Temperature and the "Emergency and Transitional Shelter Population". Human Geographies, 8(2); 1843.
Siordia C. 2014. The Regional Geoboundarization of the Mexican-origin Population in the United States through the Saenzian Logic. Mexican Journal of Scientific Research, 3(1): 27-45.
Siordia C & Wunneburger DF. 2013. On the Quantity, Degree, and Location of Public Use Microdata Area (PUMA) Fragmentation. Human Geographies, 7(2): 5-13.
Siordia C. 2013. Benefits of Small Area Measurements: A Spatial Clustering Analysis on Medicare Beneficiaries. Human Geographies, 7(1): http://dx.doi.org/10.5719/hgeo.2013.71.53.
Siordia C. 2013. Sociospatial Inequality: Combining Multilevel and Spatial Analysis. Journal of Maps, 9(1):50-54.
Siordia C & Fox A. 2013. Public Use Microdata Area Fragmentation: Research and Policy Implications of Polygon Discontiguity. Spatial Demography, 1(1), 42-56.
Siordia C, Saenz J, & Tom SE. 2012. An Introduction to Macro-Level Spatial Nonstationarity: A Geographically Weighted Regression Analysis of Diabetes and Poverty. Human Geographies, 6.2:5-13.

Parallel Geo-Coder

- None available
- Compensating by documenting all procedures in detail
    - Notes should allow future validation, challenge, and/or improvements by 3rd party

Software and Geocoding Procedures

- Only attempting to geocode individuals with address inside Allegheny and Shelby County
- Raw addresses in ".xls" format (separate files from Memphis and Pittsburgh sites)
- Saved .xls files as a "Comma Separated Values Files" i.e., CSV (MS-DOS)
- CSV (MS-DOS) file imported to ArcMap 10.1
- X- and Y-coordinates identified in ArcMap 10.1
- X- and Y-coordinates used to identify tract of residence and link "tract measures"
- X- and Y-coordinates referenced to 2000 US Census blocks (smallest administrative unit available)
- This means HABC IDs were linked "all the way down to year-2000 block" (15 digits)
- Geocoded participants exported out of ArcGIS in ".dbf" format
- Then .dbf file opened in Excel
- The .xlsx file in previous staged imported to SAS 9.3
- Measures at the tract-level produced in ArcMap 10.1
- All statistical management and analysis conducted in SAS 9.3

Logic and Source of Public Shapefiles for Geocoding

- ***Logic***
  - Health ABC participants are being geographically referenced to their 2000 US Census block of residence
  - The only source that will allow for standardize procedures between Memphis and Pittsburgh sites is US Census shapefiles
  - Because the US Census Bureau does NOT provide single address-to-address shapefiles, HABC addresses are being geographically referenced to "potential address ranges"
  - Address ranges exist for "street features"
  - The file being used (i.e., ADDRFEAT.shp) allows the address ranges to move from the "street centerline" to the "left or right" boundary to ensure the address will geocode to the correct US Census block
  - Using ADDRFEAT.shp files increases the number of potential geocoding matches by allowing the use of a "US Address-Dual Ranges" address locator style

- ***Source***
  - 2014 TIGER/Line® Shapefiles: Relationship Files
    - https://www.census.gov/cgi-bin/geo/shapefiles2014/layers.cgi
  - Used the Census "Address Range-Feature Shapefiles" from Allegheny County (Pittsburgh site) and Shelby County (Memphis site)
  - 2010 US Census shapefiles for administrative geo-units
    - Tennessee= ftp://ftp2.census.gov/geo/pvs/tiger2010st/47_Tennessee/47/
    - Pennsylvania= ftp://ftp2.census.gov/geo/pvs/tiger2010st/42_Pennsylvania/42/

Health ABC Address Format

- One cell contains street number, direction, name, and type
  - e.g., 150 W Scenic Hwy

    Number=150

    Direction=W

    Name= Scenic

    Type= Hwy
- Street address stored (with all the above elements) stored in a single column
  - One space included between address elements
  - Correct spelling of elements reviewed before geocoding
  - Punctuation (e.g., ?) removed in address column reviewed before geocoding
  - Abbreviations (e.g., NW, SE, W, etc) for street elements standardized
- No secondary address information (e.g., apartment #) included in file
- Spelled out city name stored in a single column
- Abbreviated state name stored in a single column
- Five digit zip codes stored in a single column
- Four digit HABCID number stored in a single column

ArcGIS Procedures for Deriving Coordinates from Addresses

- Create "Address Locator" in ArcCatalog 10.1
- Open folder where CVS (MS-DOS) file with addresses is located
- Righ-click within folder, create "new" file, select Address Locator
- In "Create Address Locator" pop-up window enter:
    - Select "US Address – Dual Ranges" for Address Locator Style
    - Add county specific ADDRFEAT.shp
    - Select as follows:
        - ✓ Feature ID= FID
        - ✓ From Left= LFROMHN
        - ✓ To Left= LTOHN
        - ✓ From Right= RFROMHN
        - ✓ To Right= RTOHN
        - ✓ Street Name= FULLNAME
        - ✓ Left ZIP Code= ZIPL
        - ✓ Right ZIP Code= ZIPR
    - Request the creation of the address locator
    - Appendix A shows "results window" for Memphis address locator
    - Appendix B shows "results window" for Pittsburgh address locator
- Open created address locator in ArcMap
    - File→Add Data→Geocoding→Address Locator Manager
- Then "Geocode Addresses" from toolbar with CSV (MS-DOS) file for site
- X- and Y-coordinates were used to identify their US Census year-200 block of residence

Results from Geocoding Procedures

- **Pittsburgh**
  - At baseline, there were **1,527** participants in Pittsburgh site
  - Pittsburgh address file only contained 1,419 useable addresses
  - Siordia successfully geocoded **1,292** baseline Pittsburgh participants
  - Archived data links these 1,292 participants to their 15 digit year-200 block ID number

- **Tennessee**
  - At baseline, there were **1,548** participants in Memphis site
  - Memphis address file only contained 1,505 useable addresses
  - Siordia successfully geocoded **1,396** Memphis participants
  - Archived data links these 1,396 participants to their 15 digit year-200 block ID number

By: Carlos Siordia (2015)
Data: HabcGeoCs (n= 2,688)

Variable Names, Data Sources, and Descriptions

- **HABCID**
  - *Data Source*= From internal Health ABC files at Pittsburgh site
  - *Definition*= Unique participant identifier

- **STATE_2000**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census state
  - *Definition*= State FIP number in character format

- **COUNTY_2000**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census county
  - *Definition*= County FIP number in character format

- **TRACT_2000**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census tract
  - *Definition*= Tract FIP number in character format

- **TRACT2000**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census tracts in ArcGIS format
  - *Definition*= US Census tract ID (11 digits) in numeric format

- **BLOCK_2000**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census block
  - *Definition*= Block FIP number in character format

- **BLOCK2000ID**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census block ID
  - *Definition*= Complete block ID (15 digits) in character format

- **ARESQMI**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 census tracts
  - *Definition*= Area (in square miles) of the 2000 US Census tract

- **HOUUNITS**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Data Location*= American FactFinder (factfinder.census.gov)
  - *Data Vintage*= 2000 census of the complete population
  - *SF1 Variable*= Number of "housing units" in tract
  - *Definition*= Total number of housing units in tract

- **NETRESDENS**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Definition*= "Net Residential Density" calculated by taking HOUUNITS (tract's total housing units) and dividing it by ARESQMI (tract's area in square miles)

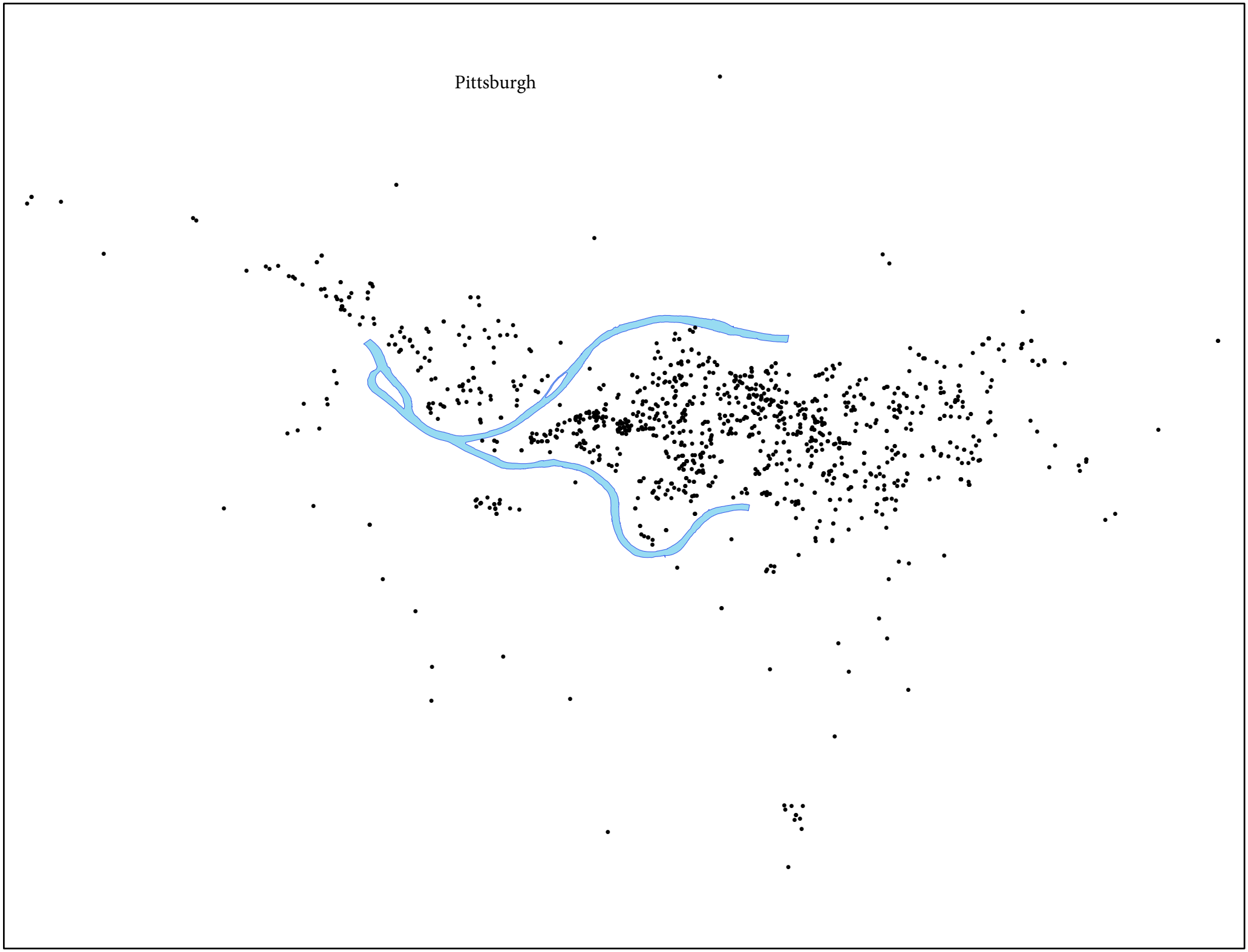
- **TOTPOP**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Data Location*= American FactFinder (factfinder.census.gov)
  - *Data Vintage*= 2000 census of the complete population
  - *SF1 Variable*= "Total population" in tract
  - *Definition*= Number of individuals living in 2000 tract

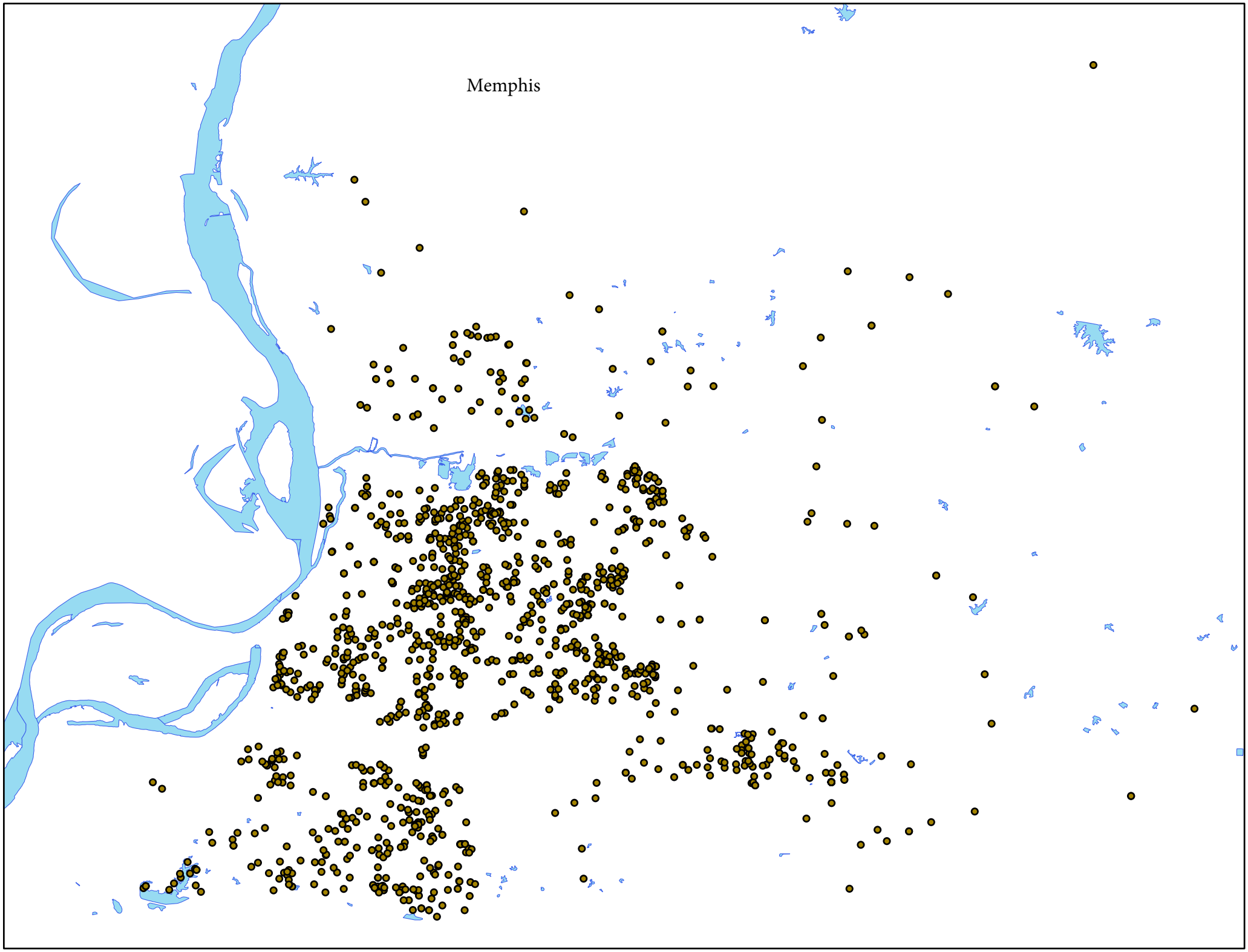
- **PERINPOV**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Data Location*= American FactFinder (factfinder.census.gov)
  - *Data Vintage*= 2000 census of the complete population
  - *SF1 Variable*= Percent of people "in-poverty" in 2000 tract
  - *Definition*= Percent of people living below 2000 federal poverty threshold (universe is all those with a poverty score)

- **PER65UP**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Data Location*= American FactFinder (factfinder.census.gov)
  - *Data Vintage*= 2000 census of the complete population
  - *SF1 Variable*= Percent of people "age 65 and up" in tract
  - *Definition*= Percent of people aged 65 and up in tract

- **PERNLW**
  - *Data Source*= Summary File 1 (100% sample file)
  - *Data Location*= American FactFinder (factfinder.census.gov)
  - *Data Vintage*= 2000 census of the complete population
  - *SF1 Variable*= Percent of "Non-Latino-Whites" in tract
  - *Definition*= Percent of Non-Latino-Whites in tract

- **STCONEK**
  - *Data Source*= TIGER® Lines shapefiles from US Census Bureau
  - *Data Vintage*= 2000 US Census "all edges" in ArcGIS format
  - *Definition*= Number of street intersections (i.e., points where $\geq$ 2 street lines connect) within the 2000 US Census tract


Pittsburgh


Memphis

The CONTENTS Procedure

Data Set Name	EMILY.GEOCODE	**Observations**	2688
Member Type	DATA	**Variables**	15
Engine	V9	**Indexes**	0
Created	Tuesday, April 12, 2016 12:52:26 PM	**Observation Length**	112
Last Modified	Tuesday, April 12, 2016 12:52:26 PM	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	12288
Number of Data Set Pages	25
First Data Page	1
Max Obs per Page	109
Obs in First Data Page	80
Number of Data Set Repairs	0
Filename	C:\Users\TGlasser\Desktop\US Census\geocode.sas7bdat
Release Created	9.0301M2
Host Created	W32_7PRO

The CONTENTS Procedure

Variables in Creation Order						
#	**Variable**	**Type**	**Len**	**Format**	**Informat**	**Label**
1	habcid	Num	8			Health ABC Id
2	State_2000	Char	2	$2.	$2.	State FIP number in character format
3	County_2000	Char	3	$3.	$3.	County FIP number in character format
4	Tract_2000	Char	6	$6.	$6.	Tract FIP number in character format
5	Block_2000	Char	4	$4.	$4.	Block FIP number in character format
6	Block2000id	Char	15	$15.	$15.	Complete block ID (15 digits) in character format
7	Tract2000	Num	8			US Census tract ID (11 digits) in numeric format
8	AreSqMi	Num	8			Area (in square miles) of the 2000 US Census tract
9	NetResDens	Num	8			Net Residential Density = HOUUNITS divided by ARESQMI
10	TotPop	Num	8			Number of individuals living in 2000 tract
11	PerInPov	Num	8			Percent of people living below 2000 federal poverty line (universe is all those with a poverty score)
12	Per65up	Num	8			Percent of people aged 65 and up in 2000 tract
13	PerNlw	Num	8			Percent of Non-Latino-Whites in 2000 tract
14	HouUnits	Num	8			Total number of housing units in 2000 tract
15	StConek	Num	8			Number of street intersections (where>=2 streets connect) within the 2000 tract