NAACCR Webinar Exercises

May 6, 2010

Kevin Henry

Francis Boscoe

EXERCISE 1

Google Earth and Geocoding Individual Cases

Part 1 Introduction to Google Earth

1 ☐ **Open Google Earth → Click Start → Programs → Google Earth.**

2 ☐ **Navigate/Fly to a few locations: Enter Kent County RI into the textbox on the 'fly to' menu (address textbox)**

→ **Click the magnifying glass.**

Next, fly to _zip code_02903 (just enter those 5 digits)

Next, fly to Rhode Island Hospital
- A bunch of possibilities should be offered, with different letters

Finally, fly to 593 Eddy Street
-Because you are already zoomed in, it correctly assumes you mean the one in Providence. Otherwise, you'd have to specify the city or zip.

Tilt/Rotate

3 ☐ **Navigating around the map (pan, zoom, tilt) . Practice using the navigation tools:**

Pan

Zoom

a. **<u>Pan</u>: Using the mouse click on the screen and move the mouse. Alternatively use the 'pan' tools direction arrows.**

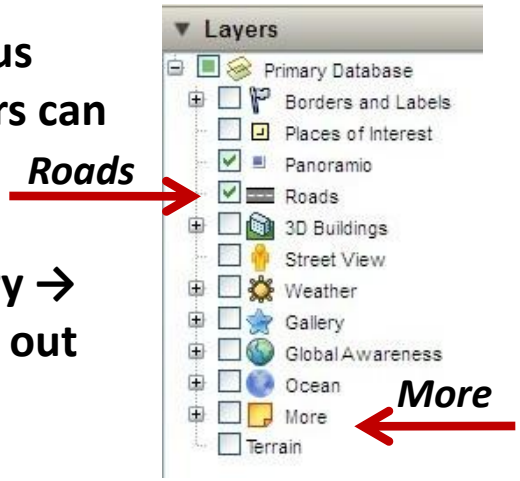
b. **<u>Zoom:</u> Click the negative sign (-) on the 'zoom' tool to zoom out and the positive sign (+) to zoom in**

c. **<u>Tilt/Rotate:</u> Slide N to rotate and click arrows to rotate**

4 ☐ **Google Earth Layers: The 'Layers' window display the various geographic layers that are included with Google Earth. Layers can be turned on/off by clicking the check box.**

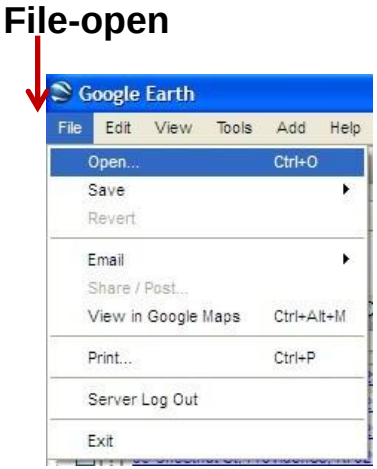
a. Try turning roads on and examine the map (zoom in)

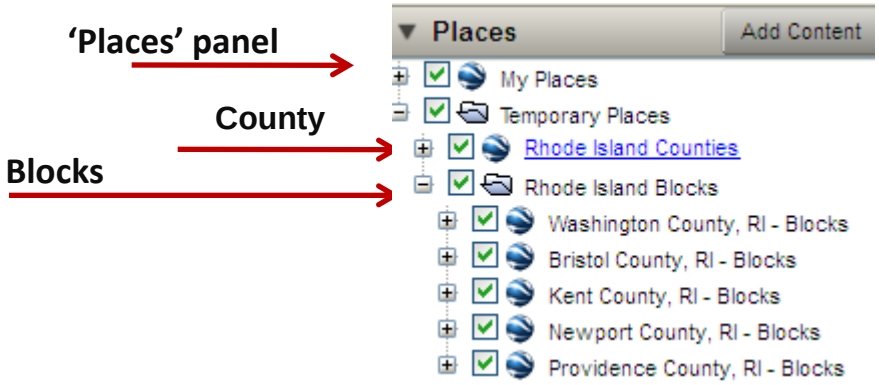
b. Open the hospital layer: Click the + on the 'More' category → Click the + on 'Place Categories' → Select Hospitals → Zoom out to see all of the Rhode Island hospitals



5 ☐ **Opening Geographic Files: Click File → Open→ Navigate to the exercise data folder (e.g C:\NAACCRWEBINAR\Exercisedata\→ Select the following files: RhodeIslandCounty, RhodeIslandBlocks (hold Ctrl key while selecting)**

*** You should see the files added in the 'Places' panel.**



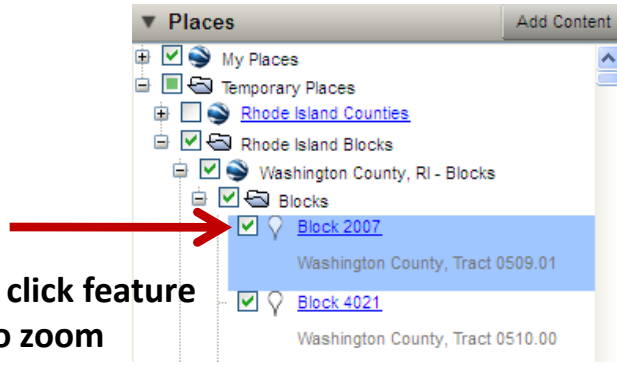


6 ☐ **Examining/Navigating geographic files in the 'Places' panel.**

a. Check boxes turn layers on/off

b. Plus (+) sign allows navigation of the geographic objects (e.g. individual tracts) within a feature layer.

c. From the 'Places' panel it is possible to zoom (fly) to selected features: → Click (+) sign for Rhode Island Blocks →Click (+) sign for Washington County → Click (+) sign for Blocks → select and double click Block '2007': Repeat step for Block '4021'

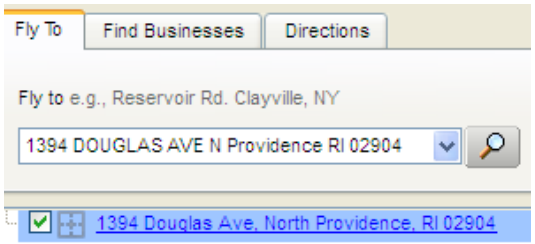


Double click feature to zoom

Part 2 Manual Geocoding

1 ☐ **Change the Google Earth coordinate system to decimal degrees: Select tools → Options → 3D View →Select 'Decimal Degrees' → click ok**

2 ☐ **Open Excel → Open file called 'manualgeocodes.xls' → Copy an address from the Excel spreadsheet and paste it into the textbox on the 'fly to' menu in Google Earth → Click Magnifying Glass to fly to location *first address will be** 1394 DOUGLAS AVE N Providence RI 02904



3 ☐ **Add a placemark- (Ctrl + Shift + p) or click pin icon**

Pin icon

4 ☐ **Edit new placemark: Mouseover new placemark**  **and move it over geocode marker**  **→ add placemark name (Study ID 3)**
→ Copy Latitude values (select and Ctrl C)→ paste into excel (Ctrl P)
→ Copy Longitude values (select and Ctrl C)→ paste into excel (Ctrl P)

5 ☐ **Find County/Census Tract/Block: Locate the census block area (boundary) where the address marker resides and find the census block marker (you might have to pan and zoom slightly). Mouseover the block marker → double click to get county-tract-block info. → Manually add county, tract, block info to excel spreadsheet .**

Note: Census tract codes have a 4-digit basic number (with leading zeros) and also may have a 2-digit suffix. Census tract numbers range from 0001.00 to 9999.98. In computer-readable files, the decimal point is implied. Leading zeros in a census tract number (for example, 011900) are shown only in computer-readable files. Thus, tract 119 is expressed as 011900 in the NAACCR layout (V11 Item # 130)

Census block marker

6 ☐ **Add census tract certainty codes and GIS Coordinate Quality codes based on the geocode result (see next page for codes)**

7 ☐ **Repeat steps 1-6 for the remaining addresses (4-20)**

Add these values to the address table

studyid	Addr at DX-No & Street	Addr at DX-City	Addr at DX State	Addr at DX--Postal Code	Latitude	Longitude	Census Tract Certainty	GIS Coordinate Quality	County (FIPS)	Census Tract	Census Block
1	449 W SHORE RD	Warwick	RI	02889	41.728206	-71.391185	1	03	003	021402	2002
2	176 N VIEW AVE	Cranston	RI	02920	41.779343	-71.465588	1	02	007	014300	2030
3	1394 DOUGLAS AVE	N Providence	RI	02904							
4	207 POCASSET AVE	Providence	RI	02909							
5	593 EDDY ST	Providence	RI	02903							
6				02886							
7	2031 SMITH ST	N Providence	RI	02911							
8	1524 ATWOOD AVE	Johnston	RI	02919							
9	844 ADMIRAL ST	Providence	RI	02904							
10	422 VALLEY ST	Providence	RI	02908							
11	39 STATE ST	Bristol	RI	02809							
12	65 EAST AVE	Pawtucket	RI	02860							
13	311 DORIC AVE	Cranston	RI	02910							
14	292 SEA VIEW DR	Warwick	RI	02886							
15	24 DELTA DR	Pawtucket	RI	02860							

GIS Coordinate Quality

Description

Code indicating the basis of assignment of latitude and longitude coordinates for an individual record from an address. This data item is helpful in identifying cases that were assigned coordinates based on incomplete information, post office boxes, or rural routes. Most of the time, this information is provided by a geocoding vendor service. Alternatively, a central registry staff manually assigns the code. This item is not coded by the hospital. Codes are hierarchical, with lower numbers having priority.

Rationale

Spatial analysis of cancer data often requires identifying data records with a high degree of locational precision. Researchers can use this code as a basis for selecting records with a degree of precision that is appropriate to the study.

- **Codes**
  - **00 Coordinates derived from local government-maintained address points, which are based on property parcel locations, not interpolation over a street segment's address range**
  - **01 Coordinates assigned by Global Positioning System (GPS)**
  - **02 Coordinates are exact match of house number and street, and based on property parcel location**
  - **03 Coordinates are exact match of house number and street, interpolated over the matching street segment's address range**
  - **04 Coordinates are street intersections**
  - **05 Coordinates are at midpoint of street segments (missing or invalid building number)**
  - **06 Coordinates are address ZIP code+4 centroid**
  - **07 Coordinates are address ZIP code+2 centroid**
  - **08 Coordinates were obtained manually by looking up a location on a paper or electronic map**
  - **09 Coordinates are address 5-digit ZIP code centroid**
  - **10 Coordinates are point ZIP code of Post Office Box or Rural Route**
  - **11 Coordinates are centroids of address city (where address ZIP code is unknown or invalid, and there are multiple ZIP codes for the city)**
  - **12 Coordinates are centroid of county**
  - **98 Latitude and longitude are assigned, but coordinate quality is unknown**
  - **99 Latitude and longitude are not assigned, but geocoding was attempted; unable to assign coordinates based on available information**
  - **Blank GIS Coordinate Quality not coded**

CENSUS TR CERTAINTY 2000

Alternate Name	Item #	Length	Source of Standard	Column #
	365	1	NAACCR	101-101

Description

Code indicating basis of assignment of census tract for an individual record. Helpful in identifying cases tracted from incomplete information or P.O. Box. Most of the time, this information is provided by a geocoding vendor service. Alternatively, a central registry staff manually assigns the code. This item is not coded by the hospital. Codes are hierarchical, with lower numbers having priority.

Note: Codes 1-5 and 9 are usually assigned by a geocoding vendor, while code 6 is usually assigned through a special effort by the central registry.

Codes

1	Census tract based on complete and valid street address of residence
2	Census tract based on residence ZIP + 4
3	Census tract based on residence ZIP + 2
4	Census tract based on residence ZIP code only
5	Census tract based on ZIP code of P.O. Box
6	Census tract/BNA based on residence city where city has only one census tract, or based on residence ZIP code where ZIP code has only one census tract
9	Unable to assign census tract or bloc numbering based on available information
Blank	Not applicable (e.g., census tracting not attempted); Census Tract Certainty information for 2000 not coded

EXERCISE 2
Batch Geocoding

1 ☐ Open SAS 9.1 or 9.2 → Open the sas program 'Exercisedata.sas ' found in the exercise folder (e.g. C:\ NAACCRWEBINAR\EXERCISES) → submit (run) the entire file
This program creates the data for the exercises. N=100 'fictitious' Rhode Island breast cancer cases (Note: case data based on business addresses)

2 ☐ Open the SAS program 'exercise2.sas' ('C:\ NAACCRWEBINAR \EXERCISES\')

3 ☐ Step 3 of the SAS code will format the address file which will be uploaded to the University of Southern California's (USC) geocoding service: → Run Step 3 of the SAS code only <u>(highlight only step 3 and then run)</u>

→ A file called 'TOGEOCODE.CSV' will be found in the directory 'C:\ NAACCRWEBINAR\ EXERCISEDATA' .

(Steps 4-5 use USC Geocoder)

4 ☐ Upload and verify data

a. Log on to USC GEOCODER Website
https://webgis.usc.edu/Services/Geocode/Default.aspx

b. At top of page select databases → upload databases

c. <u>UPLOAD NEW DATABASE</u> → Select type comma separated values (CSV) → Click browse → SELECT FILE 'togeocode.csv' → Select text separator=comma → select text qualifier=double quote → click upload *(DATA IS NOW UPLOADED)*

d. <u>VALIDATE DATABASES</u>: "DATABASES YOU OWN"→ Click validate →[STEP 1] 1)Click validate Database → Click proceed to step 2 → Click 2)Validate Table → Click proceed to step 3 → Review table (VERIFICATION IS COMPLETE)

5 ☐ a. GEOCODE DATA: → At the top of page select 'services'- → Geocoding

b. Select 'batch database geocoding'

c. Select your database (togeocode.csv) → Click start-step 1 → Click next-step 2

d. Select your database and table → Click next-step 3

e. Identify address data fields

id=studyid
streetaddress=dxstreet
city=dxcity
state=dxstate
zip=dxzip

f. Autogenerate the remaining fields

g. For output fields (optional) → Select censusvalues → autogenerate fields ;

h. Click next step 4 ;

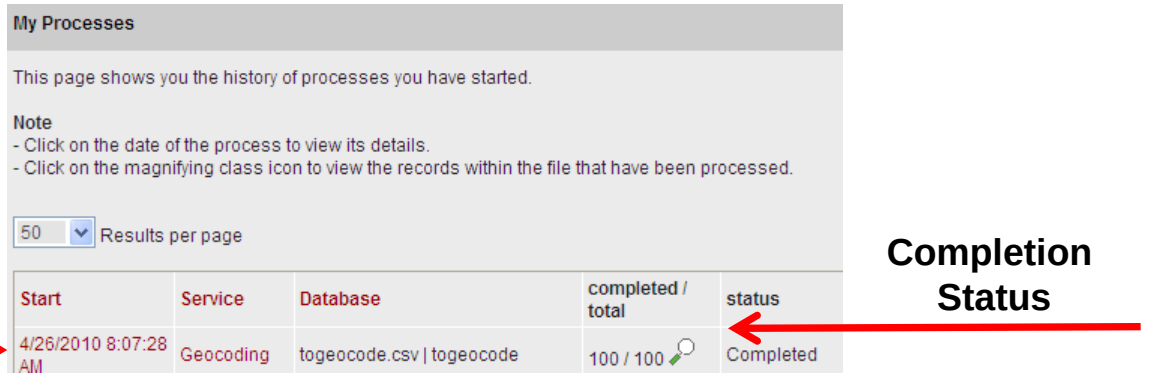
i. Click start process (bottom of page);

j. Click view process status *(you may have to refresh webpage to see updates)*

k. Under the field 'start' click the 'data & time' text after processing is complete.

l. Click download → Name the file geocodefromusc → Save the file at C:\NAACCRWEBINAR\EXERCISEDATA\

Click date/time to download results



Completion Status

(Steps 6-9 use SAS)

6 ☐ → Select and Run step 6 of SAS code only
Imports and formats geocoded records returned from USC Geocoder & code census tract certainty and geocoding quality (output file= geocodefromUSC3**)**

7 ☐ → Select and Run step 7 of SAS code only
Summarizes geocode match types & quality (output file= geocodefromUSC4)

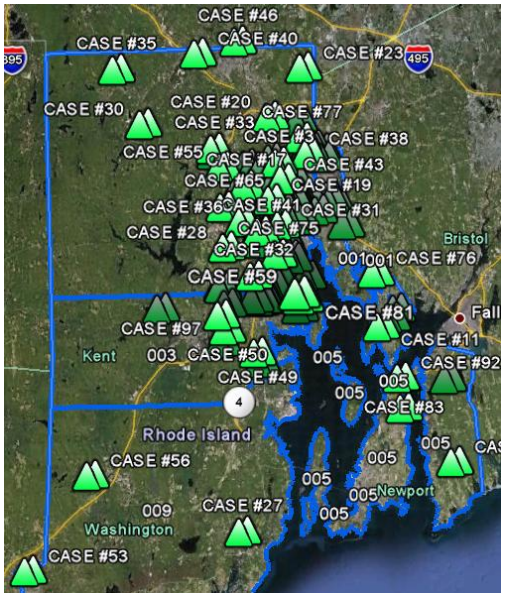
8 ☐ → Select and Run step 8 of SAS code
Joins geocoded cases to the breast cancer data based on the id-key created in step 3 (output file=Rigeocases)

9 ☐ → Select and Run step 9 of SAS code
Exports geocoded cases (Rigeocases) from SAS to KML format for visualization in Google Earth (output file= Rigeocases.kml)

(Step 10 use Google Earth)

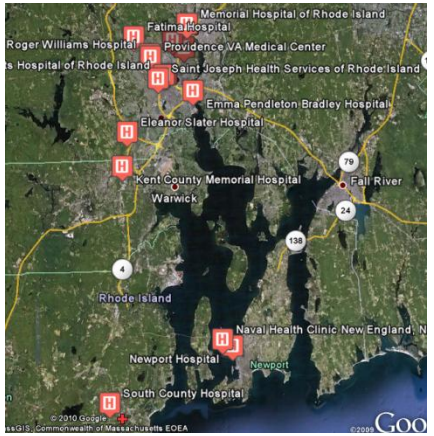
10 ☐ **Open RIgeocases.kml file in Google Earth**

→ Click File → Open→ Navigate to the 'exercisedata' folder (e.g C:\NAACCRWEBINAR\Exercisedata\→ Select 'RIgeocases.kml'
*** You should see the files added in the 'Places' panel;**



EXERCISE 3
Creating and Editing Geographic Data in Google Earth

(Steps 1-4 use Google Earth)



1 ☐ **Open *RIHOSPITALS.kml* file in Google Earth:**
→ Click File → Open→ Navigate to the 'exercisedata' folder C:\NAACCRWEBINAR\EXERCISEDATA\ → Select 'RIHOSPITALS.kml'
*** You should see the 'RIHOSPITALS.kml' file added in the 'Places' panel;**

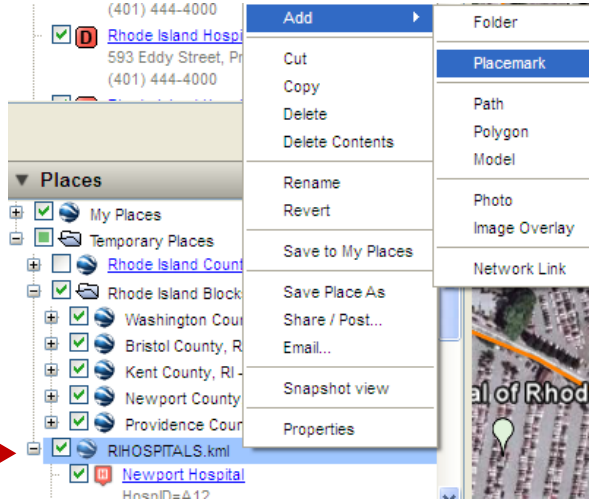
2 ☐ **You will add one missing hospital to the RIHOSPITALS.KML layer:**

a.Type the following hospital address into the Google Earth address window: Rhode Island Hospital,
593 Eddy Street,
Providence, RI 02895

b. Select the 'RIHOSPITALS' layer in the 'Places panel' → right mouse click and select 'ADD PLACEMARK' → ADD NAME: 'Rhode Island Hospital' → click OK;

The addition of Rhode Island Hospital to the RIHOSPITALS layer is complete

'RIHospitals' layer

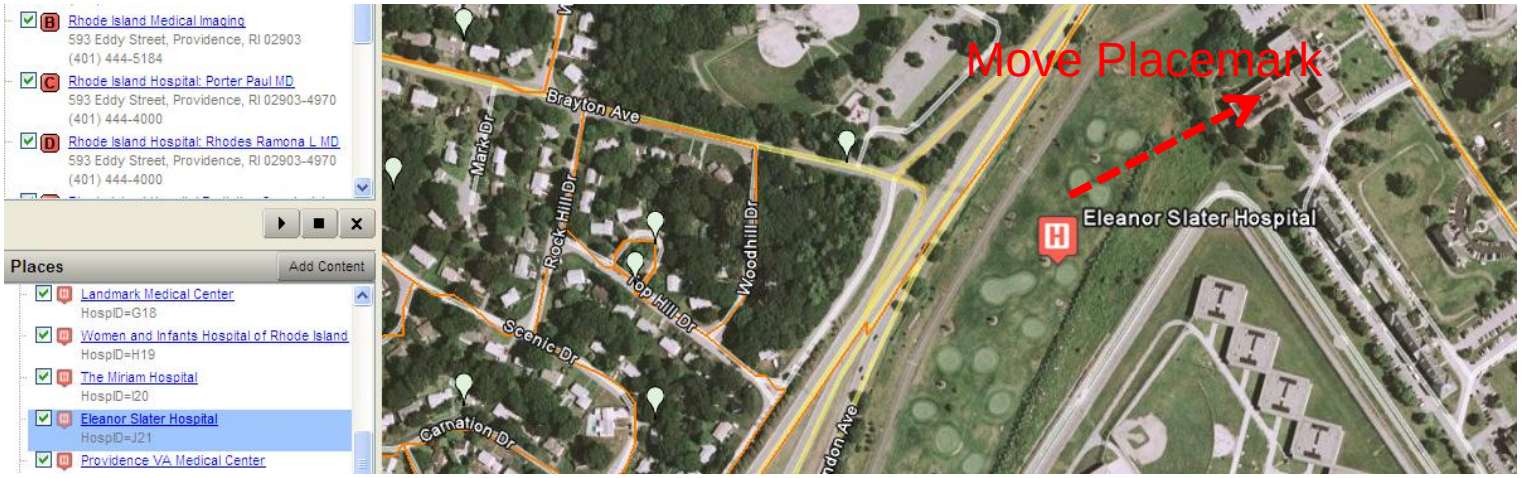


(Steps 3-4 use Google Earth)

3 ☐ **Move the placemark for 'Eleanor Slater Hospital' to the correct location:**

Locate 'Eleanor Slater Hospital' by locating this hospital in Google Earth 'Places panel' →Select Eleanor Slater Hospital → double mouse click → Google Earth will fly to 'Eleanor Slater Hospital'

→ Move the cursor over the top of the 'Eleanor Slater Hospital' placemark (red cross) → right mouse click and select properties → Move the placemark northeast about 700 feet from its current location (building looks like a cross) → click ok



4 ☐ **Save Edits. → Select RIHOSPITALS in the 'Places panel' →right mouse click → select 'Save Places As'- 'EDITED_HOSPITALS' → 'Save as Type' KML**

(Steps 5-6 use SAS)

5 ☐ Import the RIHOSPITALS_EDITS.KML file into SAS

→**Select and Run step 7 of SAS code**

6 ☐ **→ Open the SAS file 'RIEDITED_HOSPITALS' and note Rhode Island Hospital and the changed coordinates for 'Eleanor Slater Hospital' ;**

EXERCISE 4
Calculating Distances Between Locations
(Example: Census Tract Centroids To Hospitals)

(Steps 1-8 use SAS)

1 ☐ **Open the SAS program 'exercise4': (C:\ NAACCRWEBINAR \EXERCISES\)**

2 ☐ **Merge (Join) census tract centroids to geocoded RI breast cancer case file (geocases)**
→Select and Run step 2 of SAS code (output file= GEOCASES_DIST)

3 ☐ **Get only unique census tracts (reduces process time for distance combinations)**
→Select and Run step 3 of SAS code (output file= UNIQUETRACTS)

4 ☐ **Get unique combinations of census tracts & hospitals for distance calculations.**
-For this exercise we are testing all n*n combinations, however to reduce the number of combinations it is possible to select n nearest neighbors first using great circle distance calculations. In this example: 16 hospitals * 74 tracts =1184 combinations.
→Select and Run step 4 of SAS code (output file= UNIQUETRACTSCOMBO)

5 ☐ **Send records to google.maps** (from census tract centroid (latcentroid, longcentroid) to hospital (hosplat, hosplong);

→Select and Run step 5 of SAS code (output file=distance) *processing time is around 7-10 minutes

Note that in this exercise we are not sending any case records to google (only census tract centroid and hospital locations for each unique comparison [this is public information]. Cases records should only be sent to vendors that have security and encryption. Data sent to google.com is not encrypted. Within the next few months NAACCR will offer a secure service to registries to calculate driving distances and time for individual cases as part of the Komen project.

6 ☐ Select closest hospital from each census tract centroid

→Select and Run step 6 of SAS code (output file= Distance4)

7 ☐ Merge closest hospital to case data based on census tract variable [stcountytract]

→Select and Run step 7 of SAS code (output file= geocases_FINAL)

8 ☐ Summary statistics comparing great circle distance and driving distance measures

→Select and Run step 8 of SAS code

Review the statistics in the output:

What is the correlation coefficient between driving distance and GCD?