Utilizing the NAACCR Geocoder to Improve the Quality of County at Diagnosis

Justin George, M.P.H.
Alabama Statewide Cancer Registry (ASCR)

Background

Due to the nature of how County at Diagnosis is generally assigned, the ASCR has long suspected that the incidence rates for certain counties could be artificially higher or lower than what the true value should be. Each ZIP code has a default value for County at Diagnosis regardless of whether that ZIP code crosses county borders, and this could lead to the incidence rate for the default county being artificially inflated. Additionally, certain municipalities also cross county borders. In particular, Birmingham, the largest city in Alabama, is located in both Jefferson County and Shelby County, but Jefferson County is considered the default value for Birmingham. Jefferson County also contains 17 ZIP codes that are shared with its neighboring counties. Historically, Jefferson County has had one of the highest incidence rates for cancer in Alabama while Blount County, St. Clair County, and Shelby County had incidence rates significantly lower than the state average. The ASCR felt it was likely that misclassification of County at Diagnosis could explain a good portion of these differences.

Beginning in early 2013, the ASCR began to utilize the NAACCR geocoder to assign latitude and longitude as well as the appropriate census tract information for all ASCR cases. As part of an on-going process to improve the quality of ASCR data, the QA Manager for the ACSR began to notice a large number of advanced edits for Census Tract 2000, State, County, 2000-2009 (NPCR) and for Census Tract 2010, State, County, 2010-2019 (NPCR). Unlike ZIP codes, census tracts do not cross county borders. The investigation of these edits lead to the discovery that the census tract did not match the reported county of diagnosis. thankfully, the results from the NAACCR Geocoder also include the geocoded County at Diagnosis for every case. A list of all cases for which geocoded and reported County at Diagnosis did not agree was created, and changes were made to County at Diagnosis based on the quality of the geocode. All street-level matches were changed, but only ZIP code-level matches that involved ZIP codes that do not cross county borders were changed. Before the changes to County at Diagnosis were made, incidence rates were calculated for each county. Once changes were made the incidence rates were re-calculated to reflect the changes.

County at Diagnosis and ZIP Code

Alabama ZIP Code Assignment and Methodology

- In general, County at Diagnosis (NAACCR Item #90) is assigned based on the value(s) of Address at Diagnosis—Postal Code (NAACCR Item #100) and/or Address at Diagnosis—City (NAACCR Item #70).
- Each ZIP code has a default value for county.
- Through the website, the Alabama Statewide Cancer Registry (ASCR) provides each reporting facility with a ZIP code/County list that is updated annually. On this list no ZIP code appears in more than one county.
- 221 standard ZIP codes in Alabama cross county lines with 164 ZIP codes appearing in two counties, with 49 ZIP codes appearing in three counties, and with 8 ZIP codes appearing in four counties.
- Every county in Alabama has at least 2 ZIP codes that are shared with another county.
- The average county in Alabama has 7.6 shared ZIP codes with a median of 7.

Geocoding Results

- The ASCR utilized the NAACCR Geocoder for 234,086 cases from diagnosis years 2002-2011.
- 10,543 cases (4.5%) were returned with a geocoded county of diagnosis that differed from the reported county of diagnosis.
- 8,038 cases of the 10,543 misclassified cases (76.2% of the misclassified, 3.4% of the total) were street-level matches meaning the GIS Coordinate Quality for these matches was a value of '03' or better.
- All 8,038 cases with street-level matches were updated to reflect the geocoded County at Diagnosis.
- The remaining 2,505 misclassified cases (23.8% of the misclassified, 1.1% of the total) were ZIP Code-level matches (GIS Coordinate Quality of '09' or worse).
- Only 81 of these 2,505 cases were updated to the geocoded County at Diagnosis (more on this later).

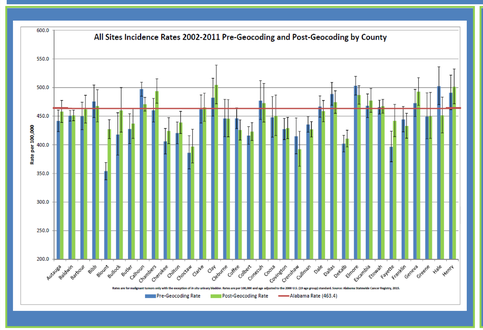
Street-Level Matches

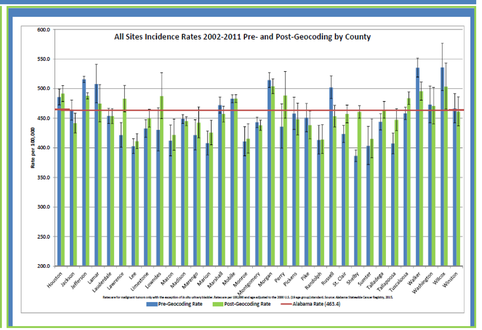
- 7,720 of 8,038 (96.0%) of street level matches with differences between geocoded and reported County at Diagnosis involved addresses with ZIP codes that were shared between at least two counties.
- Addresses in 25 ZIP codes that cross county borders were responsible for over 53% of all street level matches with differences between geocoded and reported County at Diagnosis.
- The top 5 ZIP codes with the most corrections (and 7 of the top 12 ZIP codes) included Jefferson County and a neighboring county. In fact, Jefferson County was involved in over a third of all County at Diagnosis corrections based on street level matches.
- Only 318 (4.0% of street level matches, 0.1% of the total) of the cases with street-level matches and with incorrect values for County at Diagnosis were due to miscoding by the abstractor.

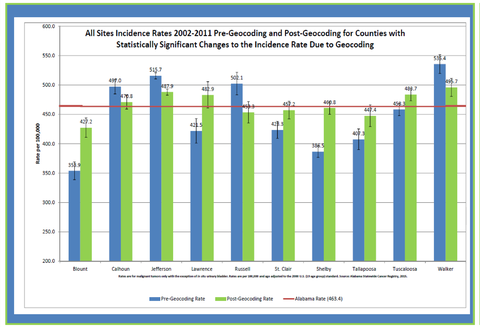
ZIP Code-Level Matches

- 2,424 of 2505 (96.8%) of ZIP code-level matches with differences between geocoded and reported County at Diagnosis involved addresses with ZIP codes that were shared between at least two counties.
- Due to the poor quality of these matches, changes were not made to the reported county of diagnosis as this would have artificially inflated the rates for the default county for each of these ZIP codes.
- Only 81 (3.2% of ZIP code-level matches, <0.1% of the total) of the cases with ZIP-code level matches and with incorrect values for County at Diagnosis were due to miscoding by the abstractor. These values for County at Diagnosis were corrected to the geocoded value for these cases.

Incidence Rates for Reported County and Geocoded County at Diagnosis







All Sites Cancer Incidence Rates 2002-2011 Pre- and Post-Geocoding for Counties with Statistically Significant Changes to the Incidence Rate Due to Geocoding

County Name	Pre-Geocoding Rate (CI)	Post-Geocoding Rate (CI)	% Change
Blount County	353.9 (339.0-369.2)	427.2 (410.9-440.0)	20.7%
Calhoun County	497.0 (484.9-509.3)	470.8 (459.0-482.8)	-5.3%
Jefferson County	515.7 (510.4-521.0)	487.9 (482.8-493.2)	-5.4%
Lawrence County	421.5 (401.1-442.8)	482.9 (461.0-505.6)	14.6%
Russell County	502.1 (483.3-521.6)	453.3 (435.4-471.8)	-9.7%
St. Clair County	423.3 (409.0-438.0)	457.2 (442.4-472.4)	8.0%
Shelby County	386.5 (376.9-396.4)	460.8 (450.2-471.5)	19.2%
Tallapoosa County	407.3 (390.0-425.2)	447.4 (429.2-466.1)	9.8%
Tuscaloosa County	458.3 (448.0-468.7)	483.7 (473.2-494.5)	5.5%
Walker County	535.4 (519.7-551.6)	495.7 (480.6-511.2)	-7.4%

Rates are for malignant cases only with the exception of in situ urinary bladder. Rates are per 100,000 and age adjusted to the 2000 U.S. (19 age groups) standard; Confidence Intervals are 95%.
The Alabama comparison rate for this time period is 463.4 (461.5-465.3).
Source: Alabama Statewide Cancer Registry, 2015.

Geocoded County vs. Reported County at Diagnosis and Incidence Rates

- The incidence rates for the majority of Alabama counties (45 of 67, 67.2%) were not significantly affected.
- Twelve Alabama counties (17.9%) had new incidence rates that were outside of the 95% confidence interval of the original rates but have overlapping 95% confidence intervals for the old and new incidence rates. The changes for these counties are considered borderline significant.
- Ten Alabama counties (14.9%) had new incidence rates that were statistically significantly different than the previous incidence rates. Six of these counties include Jefferson County and its bordering counties.
- The three counties with the smallest percentage changes to the incidence rates (Baldwin County, Lauderdale County, and Mobile County) each only have 2 ZIP codes that share county borders.
- The counties without significant changes to the incidence rate had an average of 7 shared ZIP codes per county.
- The counties with borderline significant changes to the incidence rate had an average of 8 shared ZIP codes per county.
- The counties with statistically significant changes to the incidence rate had an average of 9.6 shared ZIP codes per county.

Geocoding and the ASCR

History: Prior to 2007, no ASCR data had been geocoded. In 2008, the ASCR partnered with the College of Community Health Sciences at the University of Alabama (Tuscaloosa campus) to geocode ASCR data. Unfortunately, the geocoding results received from this partnership were unusable as no coordinate quality measure was provided. Beginning in 2009, the ASCR began geocoding data in-house by utilizing ARCGIS software with TIGER/Line files from the 2000 Census. This process took a long amount of time and was exceedingly tedious. As such, typically only the five most recent years for each data submission were geocoded. Because the TIGER/Line files typically had much better information for urban counties than rural counties, the quality of geocodes was noticeably better in urban counties. In fact, for the more heavily rural counties the GIS Coordinate Quality was typically '09' or worse. The ASCR continued to use this process until the NAACCR Geocoder was released.

NAACCR Geocoder Implementation: Once the NAACCR Geocoder was released, the ASCR began to test the service. A comparison of previously geocoded cases that had a coordinate quality of street-level match was made. The results from the NAACCR Geocoder were found to be consistent, and in some cases superior, to the existing geocodes. Additionally, a test file containing cases with a valid address but a GIS Coordinate Quality of '09' or worse was submitted to the NAACCR Geocoder. As a result, the NAACCR Geocoder was able to produce a street level match for many of the cases that the ASCR was previously unable to geocode to that level. In particular, the NAACCR Geocoder did a much better job of obtaining street-level matches in rural counties. Utilizing the NAACCR Geocoder was much easier and more efficient than the previous method, particularly because census tract data could be obtained concurrently with the geocode information. For these reasons, a decision was made to utilize the NAACCR Geocoder for all ASCR cases. In order to have a consistent system for geocoding for all ASCR data, the geocode information for all previously geocoded cases in the ASCR was replaced with the results from the NAACCR Geocoder.

ASCR Current Process: In September of each year, the ASCR utilizes the NAACCR Geocoder to geocode and to assign census values for all non-geocoded cases. Cases are grouped by year of diagnosis with 1996-1997 cases being assigned to the 1990 Census, with 1998-2007 cases being assigned to the 2000 Census, and with cases from 2008 and forward being assigned to the 2010 Census. This process is then repeated in November of each year prior to the Call for Data Submission. Additionally, the ASCR will resubmit previously geocoded cases with a valid address but a GIS Coordinate Quality of '09' or worse in the hopes of obtaining a better quality match.

Geocoding Timeline

Abstract

Utilizing the NAACCR Geocoder to Improve the Quality of County at Diagnosis

Justin T. George, MPH
Alabama Statewide Cancer Registry (ASCR)

The quality of data is crucial for any data analysis, and misclassification of variables can easily lead to erroneous conclusions. In the state of Alabama, there exist numerous ZIP codes that extend across the borders of multiple counties, and there are a few municipalities that extend across county borders. Because each ZIP code is associated with a default value for county at diagnosis, misclassification of the county at diagnosis variable can occur. When this happens, certain counties will appear to have incidence rates that are artificially higher or lower than what the true rates should be.

The NAACCR Geocoder is a powerful tool that can be used to supply cancer registries with many geographical variables for each case such as latitude, longitude, GIS coordinate quality, census block, census tract, census tract certainty, geocoded county, etc. By comparing the geocoded county result from the NAACCR Geocoder to the reported county at diagnosis in the ASCR database, I am able to indentify cases with potentially misclassified county at diagnosis and update the county at diagnosis field.

From 2002 through 2011, approximately 4.5% of ASCR cases had an incorrect county at diagnosis value based on the NAACCR Geocoder. Every county in Alabama had at least five miscoded cases. However, once adjustments were made and incidence rates recalculated, 45 of the 67 counties saw no significant change to their incidence rates. Twelve counties had a new incidence rate that was outside of the 95% confidence interval of the original rate, but was not significantly different as the 95% confidence interval of the new rate overlapped the original 95% confidence interval. Ten Alabama counties had new incidence rates that were either significantly higher or significantly lower than the original incidence rates. The majority of counties with significantly different rates were clustered around Jefferson County which contains Birmingham, the largest city in Alabama. More than 96% of all misclassifications of county at diagnosis involved an address containing a ZIP code shared by more than one county.

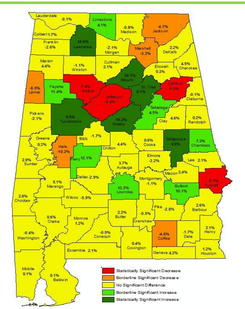
The ASCR will continue to utilize the NAACCR Geocoder to obtain geographic information and to improve the quality of our data.

GIS Coordinate Quality

Code	Value	Description
00	Address Point	Coordinates are derived from local government-maintained address points, which are based on property parcel locations, not interpolation over a street segment's address range.
01	GPS	Coordinates are assigned by Global Positioning System (GPS).
02	Parcel	Coordinates are match of house number and street, and based on property parcel location.
03	Street Segment Interpolation	Coordinates are match of house number and street, interpolated over the matching street segment's address range.
04	Street Intersection	Coordinates are street intersections.
05	Street Centroid	Coordinates are at mid-point of street segment (missing or invalid building number)
06	Address ZIP Plus4 Centroid	Coordinates are address ZIP code+4 centroid.
07	Address ZIP Plus2 Centroid	Coordinates are address ZIP code+2 centroid.
08	Manual Lookup	Coordinates are obtained manually by looking up a location on a paper or electronic map.
09	Address ZIP Centroid	Coordinates are address 5-digit ZIP code centroid.
10	P.O. Box ZIP Centroid	Coordinates are point ZIP code of Post Office Box or Rural Route.
11	City Centroid	Coordinates are centroid of address city (when address ZIP code is unknown or invalid, coordinates are centroid of the city).
12	County Centroid	Coordinates are centroid of county.
98	Unknown Quality	Latitude and longitude are assigned, but coordinate quality is unknown.
99	Unmatchable	Latitude and longitude are not assigned, but geocoding was attempted; registry is unable to assign coordinates based on available information.

Codes 00-03 are considered street-level (high-quality) matches.

Percentage Change for All Sites Combined Incidence Rates by County after Geocoding, 2002-2011



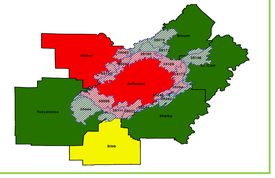
Source: Alabama Statewide Cancer Registry, 2015.
A change is considered borderline significant if the new rate is outside of the 95% confidence interval of the original rate, but the 95% confidence intervals for both rates overlap.

Geocoding and Jefferson County

Jefferson County Geocoding Findings

- Jefferson County is the most populous county in Alabama, and the majority of Birmingham, the largest city in Alabama, is contained in Jefferson County.
- Jefferson County shares 17 ZIP codes with its surrounding counties.
- As of 2013, an estimated 217,720 people live in these shared ZIP codes.
- After geocoding 2002-2011 cases, 2,486 cases previously assigned to Jefferson County were reassigned to other counties. Also, 308 cases were reassigned to Jefferson County for a net loss of 2,178 cases from Jefferson County over this time period.
- In total, 2,794 of 8,038 cases (34.8%) for which the county at diagnosis was reassigned due to geocoding involved Jefferson County.
- Due to geocoding, the all sites cancer incidence rate for Jefferson County statistically significantly declined by 5.4% from 515.7 (510.4-521.0) to 487.9 (482.8-493.2).
- Most of the cases reassigned from Jefferson County went to Shelby County which is the only reason for the statistically significant increase in the all sites cancer incidence rate for Shelby County.
- Cases reassigned from Jefferson County were also the primary reason for the statistically significant increases in the all sites cancer incidence rates for Blount County, St. Clair County, and Tuscaloosa County.

Jefferson County and Shared ZIP Codes



For More Information

Contact

Justin T. George
Director of Cancer Epidemiology
Cancer Prevention and Control Division
201 Monroe Street, Suite 1400 H Montgomery, AL 36104
P O Box 303017 Montgomery, AL 36130-3017
Phone: 334-206-3962 Fax: 334-206-3757
Email: justin.george@adph.state.al.us
www.adph.org/ascr

*The **ASCR** was established in 1995 by the Alabama Department of Public Health in response to state law (Act 95-275) that made cancer a reportable condition. Data collection began in January 1, 1996.*

*The **ASCR** is a member of the North American Association of Central Cancer Registries (NAACCR) which sets standards for completeness, timeliness, and data quality. Registries that meet the highest standards receive NAACCR Gold Certification. Alabama has achieved the highest NAACCR standards and received the Gold Certification since data year 2004.*

*The **ASCR** is supported by the Cooperative Agreement Number 1U58/DP003854 from the Centers for Disease Control and Prevention (CDC).*

Capturing Cancer Data In Alabama

www.adph.org/ASCR

Conclusions

- Although the incidence rates for the majority of Alabama counties were not affected by the use of geocoded County at Diagnosis in place of reported County at Diagnosis, certain Alabama counties were significantly affected, particularly counties bordering Jefferson County.
- The counties with significantly altered incidence rates as a group contained more ZIP codes that share county borders on average than the counties that did not have significantly altered incidence rates.
- The overwhelming majority of all cases (96.2%) for which geocoded County at Diagnosis and reported County at Diagnosis did not match involved addresses with ZIP codes that cross county borders.
- The reported County at Diagnosis matched the geocoded County at Diagnosis 95.5% of the time.
- Less than 0.2% of all reported cases had an incorrect value for County at Diagnosis solely due to miscoding by the abstractor.
- The NAACCR Geocoder is a powerful and easy-to-use program that allows registries to geocode data and to improve the quality of the geographic data in their databases (i.e. correcting the value for County at Diagnosis). The ASCR will continue to utilize this program for these purposes.
- The advanced edits for Census Tract 2000, State, County, 2000-2009 (NPCR) and for Census Tract 2010, State, County, 2010-2019 (NPCR) should be altered to include a condition in which census tracts with a certainty based on less than a street-level match are ignored.

ALABAMA PUBLIC HEALTH