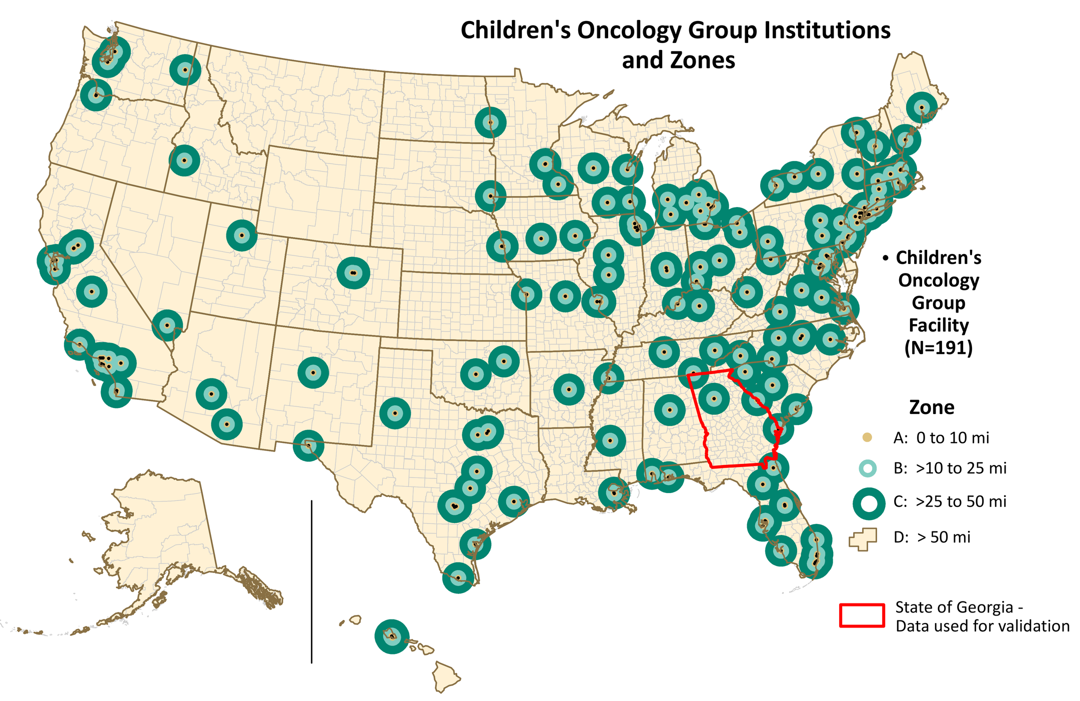
A Comparison of Methods to Change Spatial Scale

Elaine Hallisey, MA;[1] Eric Tai, MD;[2] Andrew Berens, MA;[1] Grete Wilt, MPH;[1] Lucy Peipins, PhD;[2] Brian Lewis, BS;[1] Shannon Graham, MA;[1] Barry Flanagan, PhD;[1] Natasha Buchanan Lunsford, PhD[2]
[1]Geospatial Research, Analysis and Services Program CDC/ATSDR/DTHHS [2]National Center for Chronic Disease Prevention and Public Health Promotion CDC/ONDIEH

Background

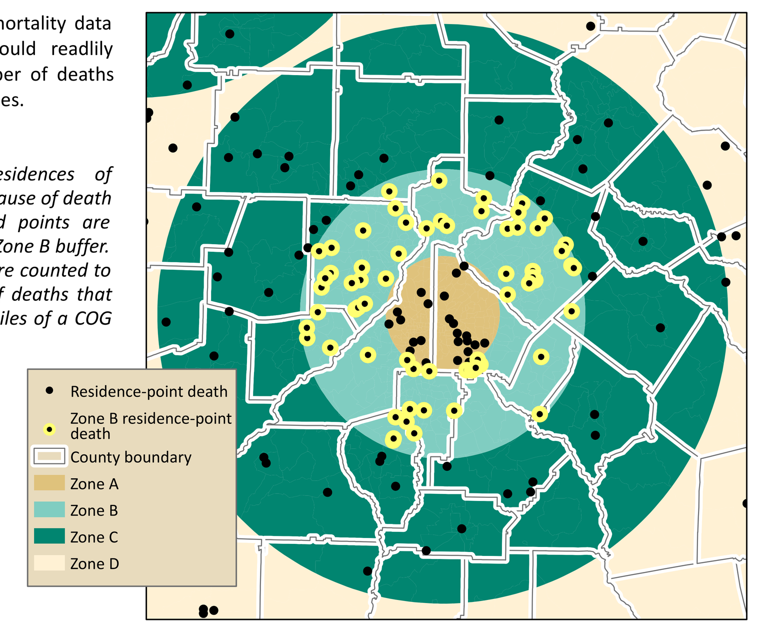
Transforming spatial data from one scale to another, referred to as change-of-support, is a challenge in geographic analysis. As part of a larger research project to understand the association between geographic barriers to pediatric cancer facilities and mortality rates among adolescents with cancer, we explored five methods to estimate adolescent cancer mortality rates for each of four zones surrounding Children's Oncology Group (COG) facilities: 1) Geographic Centroid Assignment, 2) Population-Weighted Centroid Assignment, 3) Simple Areal Weighting, 4) Combined Population and Areal Weighting, and 5) Geostatistical Areal Interpolation. Data sources for the primary study included U.S. Census 2000 and 2010 100% population counts at the tract level as well as 1999-2011 county-level cancer mortality data for adolescents, aged 15 through 19, from the National Center for Health Statistics (NCHS), compiled from individual state death certificates. To preserve confidentiality, the NCHS provides mortality data at the county level only. However, some states consider death certificates public record and share residence-level point data. We therefore obtained point-level mortality data from Georgia, a state that releases mortality data for research upon a substantiated request, to assess the accuracy of the methods.



Buffers, areas surrounding each COG, are combined so that all locations in the United States are assigned to one of four zones: Zone A) 0 to 10 miles in distance from a COG, Zone B) >10 to 25 miles from a COG, Zone C) >25 to 50 miles from a COG, or Zone D) More than 50 miles from a COG.

Ideally, if we had point-level mortality data for the entire nation, we could readlily determine the observed number of deaths for each of the four analysis zones.

The points indicate the residences of adolescents whose underlying cause of death is cancer. The yellow-encircled points are those deaths that fall within a Zone B buffer. The points within the buffer were counted to obtain the observed number of deaths that occurred between 10 and 25 miles of a COG (n=71).



Sources and Notes

Data Sources:
Children's Oncology Group https://childrensoncologygroup.org/index.php/locations; (December 2014).

U.S. Census Bureau; 2000 Census, Summary File 1 and 2010 Census, Summary File 1; generated using American FactFinder; http://factfinder2.census.gov; (December 2014).

Georgia Department of Public Health. Office of Health Indicators for Planning (OHIP). Georgia adolescent cancer mortality data. Received January 2015.

National Center for Health Statistics: Compressed Mortality File. NCHS ed. Hyattsville, Maryland 1999-2011.

Acknowledgements:
Special thanks to Gordon Freymann and Robert Attaway of GADPH/OHIP for their assistance.

Disclaimer:
The U.S. Census, the GADPH, and NCHS are only responsible for providing initial data. Analyses, interpretations, and conclusions are those of the authors.

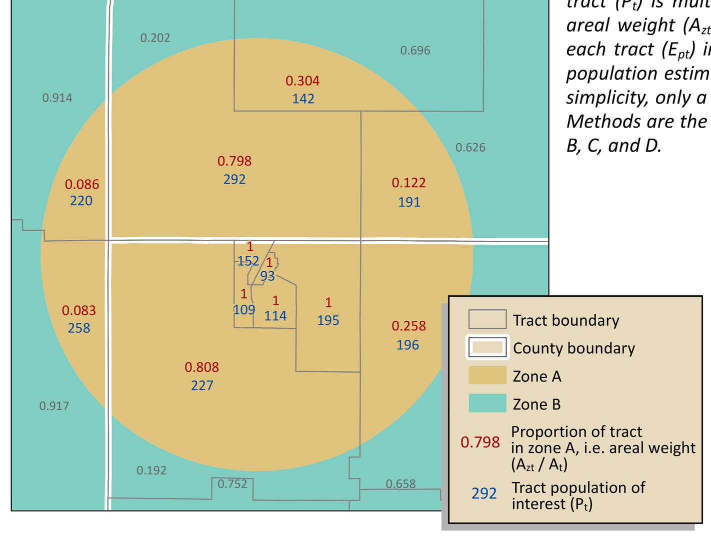
Please note, to preserve confidentiality, some of the mapped data on the poster have been randomly modified. Analyses are based on actual geospatial data.

Methods

The mortality rate for a geographic area is calculated as the number of deaths for a specified group (numerator) divided by the total population of that group (denominator).

Denominator (Population) Estimation

To approximate the study zone population for the denominator, we performed simple areal weighting using the Population Estimator tool, developed by CDC's Geospatial Research, Analysis, and Services Program (GRASP). The area of overlap of the census tract (source zone) with the study zone surrounding a COG (target zone) was divided by the area of the entire tract to obtain the proportion, or weight, of the tract area within the target zone. The population of interest for each source zone was then multiplied by the areal weight for that source zone. The resulting population proportions were summed to estimate a population total for the target zone for census years 2000 and 2010. We then calculated a weighted sum to estimate a total 13-year population for the denominator to match the 1999-2011 numerator's mortality data time range. This process was repeated for each of the four study zones.



The population for those aged 15 through 19 for each tract (P_t) is multiplied by the proportion of the tract, or areal weight (A_{zt} / A_t), in the study zone. The output for each tract (E_{pt}) in the entire zone is summed to obtain a population estimate for the study zone. Note: For graphic simplicity, only a subset of zones are shown in the figures. Methods are the same for each of the four study zones, A, B, C, and D.

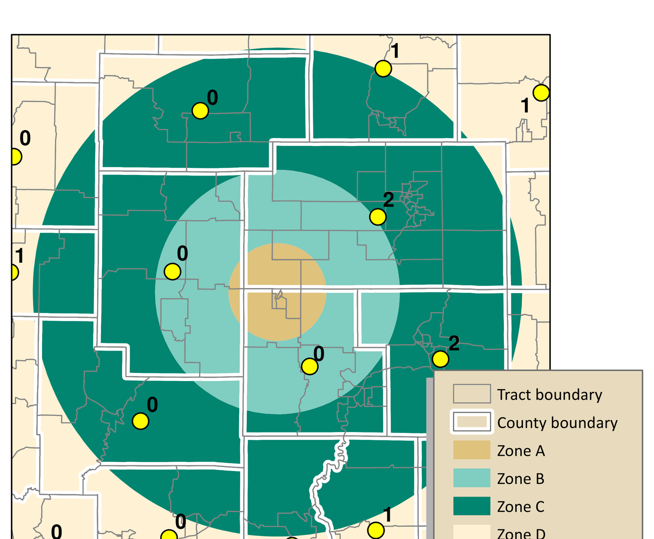
Areal weight A_{zt} / A_t	*	Tract population P_t	=	Estimate E_{pt}
0.083	*	258	=	21.41
0.086	*	220	=	18.92
0.122	*	191	=	23.30
0.258	*	196	=	50.57
0.304	*	142	=	43.17
0.798	*	292	=	233.02
0.808	*	227	=	183.42
1.000	*	93	=	93.00
1.000	*	109	=	109.00
1.000	*	114	=	114.00
1.000	*	152	=	152.00
1.000	*	195	=	195.00
		Total estimate:		1,236.80

Numerator (Deaths) Estimation

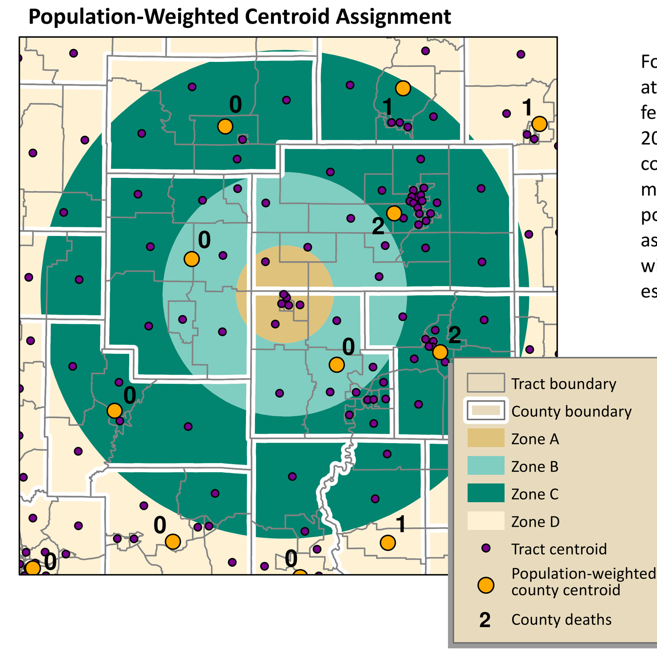
Method 1:
Geographic Centroid Assignment

For geographic centroid assignment, we attributed Georgia Department of Public Health (GADPH) mortality counts to each county's geographic center of gravity, or centroid. County deaths assigned to centroids that fall within a study zone were summed, by sex and year, to estimate the number of deaths for that zone

Each county centroid is attributed a county mortality count for the population of interest. Mortality counts for centroids falling within each study zone are summed to estimate mortality, as a whole number, by zone. In this hypothetical example, zones A and B are assigned zero deaths, despite the overlap of three counties on Zone A (two potential deaths) and five on Zone B (four potential deaths). Zone C is assigned four deaths, but has the possibility of more.



Method 2:
Population-Weighted Centroid Assignment



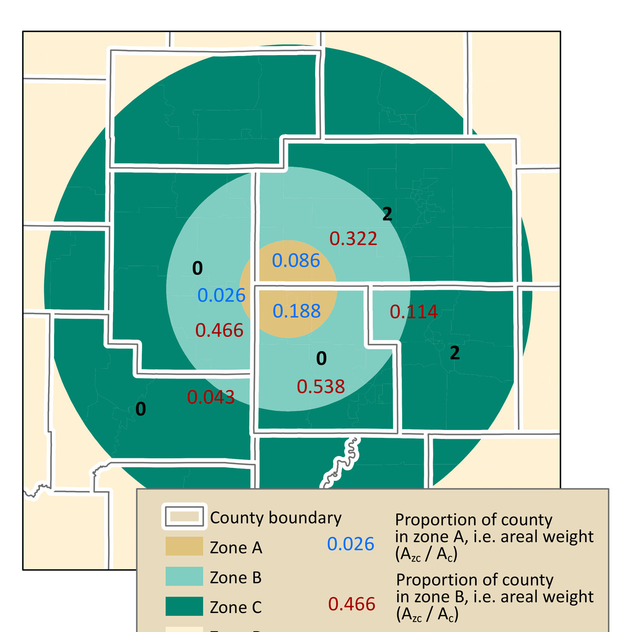
For population-weighted centroid assignment, we attributed census tract populations of males and females aged 15 through 19, for years 2000 and 2010, to tract centroids. For each of Georgia's 159 counties, we used the tract centroids to calculate mean centers, weighted by the tract-level population of interest, for each year. County deaths assigned to population-weighted centroids that fall within a zone were summed, by sex and year, to estimate the number of deaths for that zone.

Each tract centroid is attributed the population of interest. County centroids are placed using the mean center of tract centroids weighted by the tract population. Mortality counts for centroids falling within each zone are summed to estimate mortality by zone. Results for Zones A and B in this example, zero deaths for both, are the same as those for geographic centroid assignment. Zone C is assigned five deaths because the centroid in the northeast, with a value of "1," is now positioned within zone C.

Method 3:
Simple Areal Weighting

Simple areal weighting, is the same technique used for the denominator estimates, as described above. In this case, the area of overlap of the county (source zone) with the study zone surrounding a COG (target zone) was divided by the area of the entire county to obtain the proportion, or areal weight, of the county area within the study zone. The number of deaths for each county was then multiplied by the corresponding areal weight for that source zone. The resulting mortality count estimates were summed to estimate number of deaths for each of the four study zones, A, B, C, and D.

Areal weight A_{zc} / A_c	*	Deaths in county M_c	=	Mortality count estimate
0.026	*	0	=	0.00
0.086	*	2	=	0.17
0.188	*	0	=	0.00
		Zone A Estimate:		0.17
0.466	*	0	=	0.00
0.322	*	2	=	0.64
0.114	*	2	=	0.23
0.538	*	0	=	0.00
0.043	*	0	=	0.00
		Zone B Estimate:		0.87



Methods (con't)

Method 4:
Combined Population and Areal Weighting

We estimated the numerator for each zone using a conceptually dasymetric population-weighted interpolation combined with areal weighting. Because we had county-level counts only, we took advantage of the county/tract hierarchy and assigned each tract a population-weighted mortality estimate as follows:

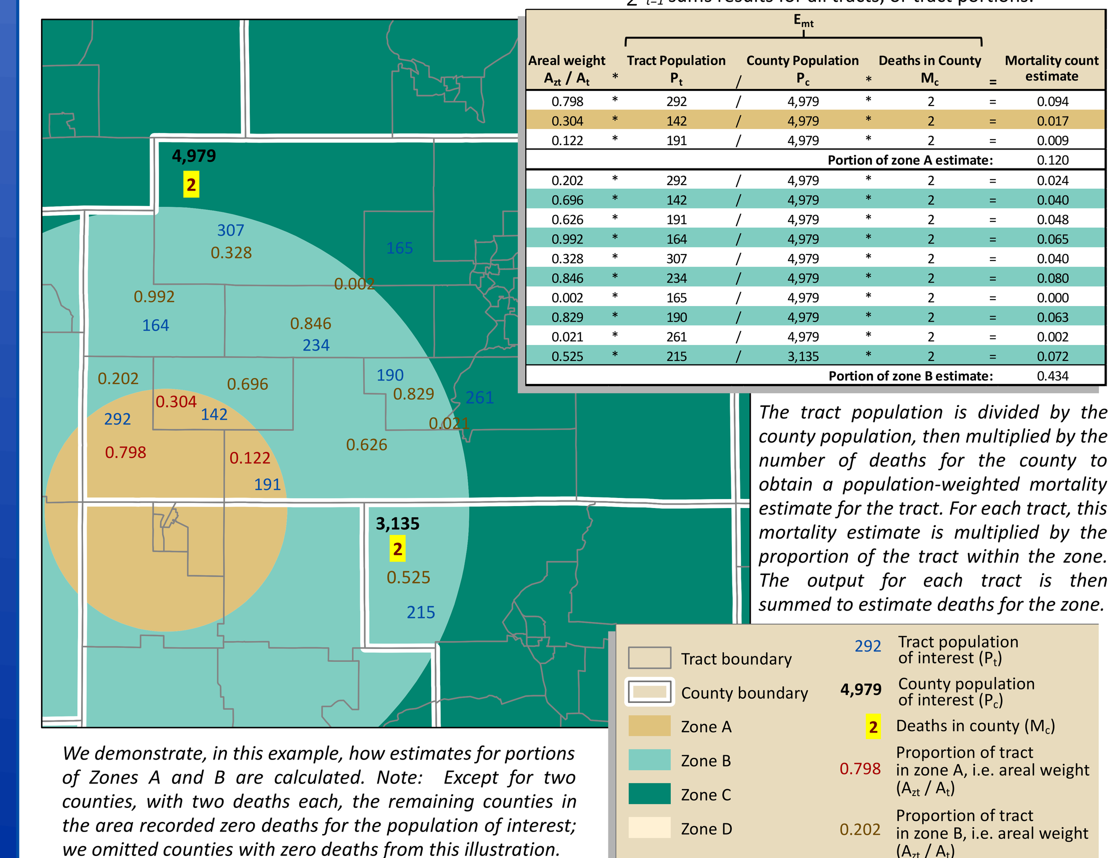
$$E_{mt} = (P_t / P_c) * M_c$$

Where:
E_{mt} is the population-weighted mortality estimate for the tract;
P_t is the tract population;
P_c is the county population; and
M_c is the number of deaths in the county.

The output of the formula was then multiplied by the proportion of the tract that falls within the zone. We summed the resulting proportions, by sex and year, to estimate the number of deaths for the zone. Expressed in its entirety, the target zone mortality is estimated as:

$$M_z = \sum_{t=1}^{n} ((A_{zt} / A_t) * E_{mt}$$

Where:
M_z is the study zone mortality count estimate;
A_{zt} is the geographic area of the overlap of the tract and study zone;
A_t is the geographic area of the entire tract;
E_{mt} is the population-weighted mortality estimate for the tract; and
$\sum_{t=1}^{n}$ sums results for all tracts, or tract portions.



Areal weight A_{zt} / A_t	*	Tract Population P_t	/	County Population P_c	*	Deaths in County M_c	=	Mortality count estimate
0.798	*	292	/	4,979	*	2	=	0.094
0.304	*	142	/	4,979	*	2	=	0.017
0.122	*	191	/	4,979	*	2	=	0.009
						Portion of zone A estimate:		0.120
0.202	*	292	/	4,979	*	2	=	0.024
0.696	*	142	/	4,979	*	2	=	0.040
0.626	*	191	/	4,979	*	2	=	0.048
0.992	*	164	/	4,979	*	2	=	0.065
0.328	*	307	/	4,979	*	2	=	0.040
0.846	*	234	/	4,979	*	2	=	0.080
0.002	*	165	/	4,979	*	2	=	0.000
0.829	*	190	/	4,979	*	2	=	0.063
0.021	*	261	/	4,979	*	2	=	0.002
0.525	*	215	/	3,135	*	2	=	0.072
						Portion of zone B estimate:		0.434

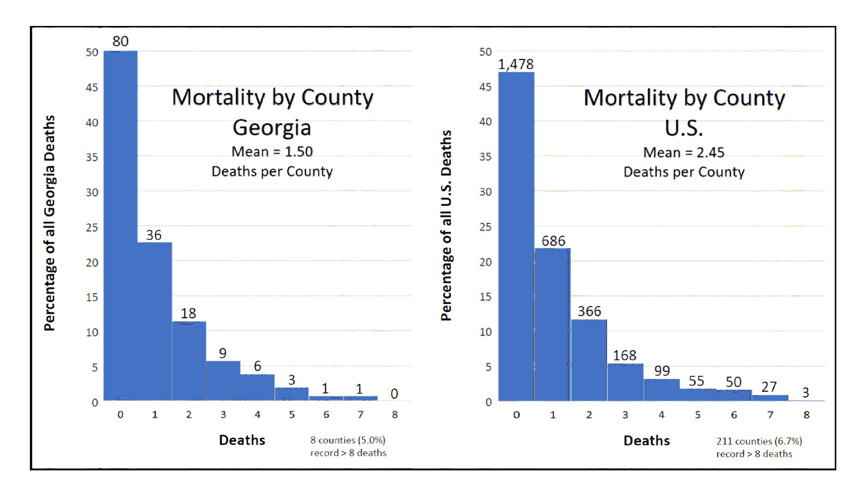
The tract population is divided by the county population, then multiplied by the number of deaths for the county to obtain a population-weighted mortality estimate for the tract. For each tract, this mortality estimate is multiplied by the proportion of the tract within the zone. The output for each tract is then summed to estimate deaths for the zone.

We demonstrate, in this example, how estimates for portions of Zones A and B are calculated. Note: Except for two counties, with two deaths each, the remaining counties in the area recorded zero deaths for the population of interest; we omitted counties with zero deaths from this illustration.

Method 5:
Geostatistical Areal Interpolation

To determine how geostatistical methods of interpolation compared to the cartographic methods described above, Georgia mortality counts were interpolated from county level data using the areal interpolation function of the Geostatistical Wizard in ArcMap 10.3.1. We used event areal interpolation, specifically over-dispersed Poisson, based on mortality count data for adolescent males and females separately. Using visual variography, we fitted a stable kriging interpolation model to a plot of empirical covariance versus distance, creating a continuous surface depicting the probability of event occurrence in the study area. During variography we used a lattice spacing of 1,000 meters, a lag size of 5,000 meters, and 18 lags. The continuous probability surface was then used to predict the mortality event counts for the COG zones, providing a numerator to determine a mortality rate for each zone based on the previously calculated denominator population.

Results



Adolescent cancer mortality counts from the GADPH were appropriate for testing the methods. The distribution of county mortality counts for Georgia mirror those of the U.S. Likewise, patterns of zone values are roughly similar for the state and the nation.

Comparisons between observed 1999-2011 Georgia adolescent cancer mortality and estimated mortality, by method and zone.

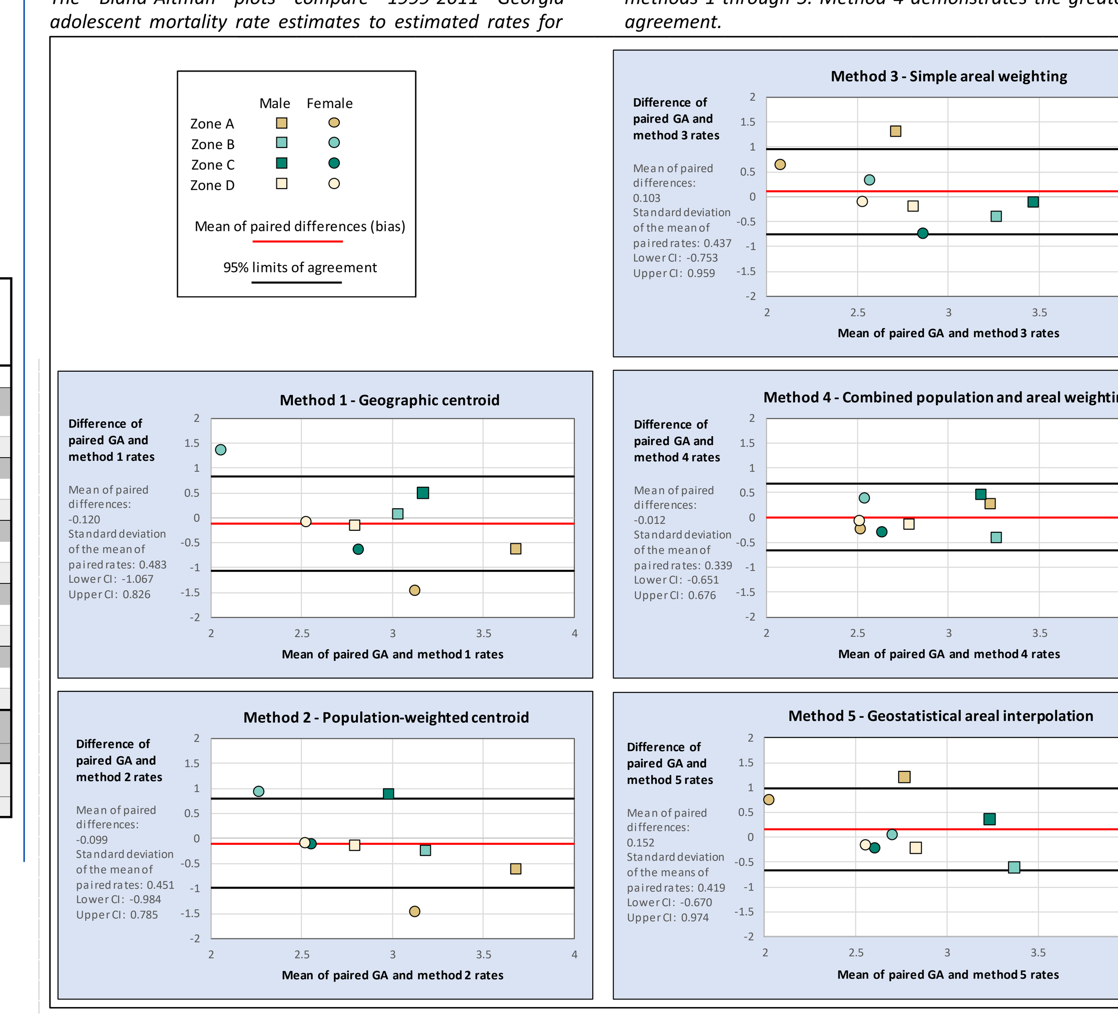
Zone	Category	State of Georgia (observed) counts and estimated rates		Method 1 GADPH geographic centroid estimates		Method 2 GADPH population-weighted centroid estimates		Method 3 GADPH simple areal weighting estimates		Method 4 GADPH combined population and areal weighting estimates		Method 5 GADPH geostatistical areal interpolation estimates	
		Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Georgia total	Deaths	134	104	134	104	134	104	134	104	134	104	133.38 (0.62)	102.22 (1.78)
	Population	4,346,302	4,087,631	4,346,302	4,087,631	4,346,302	4,087,631	4,346,302	4,087,631	4,346,302	4,087,631	4,346,302	4,087,631
	Rate	3.083	2.544	3.08	2.54	3.08	2.54	3.08	2.54	3.08	2.54	3.07 (0.01)	2.50 (0.04)
Zone A 0 to 10 miles	Deaths	22	15	26 (4)	24 (9)	26 (4)	24 (9)	13.41 (8.59)	10.88 (4.12)	20.13 (1.87)	16.38 (1.38)	14.15 (7.85)	10.22 (4.78)
	Population	652,675	624,721	652,675	624,721	652,675	624,721	652,675	624,721	652,675	624,721	652,675	624,721
	Rate	3.37	2.40	3.98 (-0.61)	3.84 (-1.44)	3.98 (-0.61)	3.84 (-1.44)	2.06 (1.32)	1.74 (0.66)	3.08 (0.29)	2.62 (-0.22)	2.17 (1.20)	1.64 (0.76)
Zone B > 10 to 25 miles	Deaths	39	32	38 (1)	16 (16)	42 (3)	21 (11)	44.03 (5.03)	28.01 (3.99)	43.95 (4.95)	27.36 (4.64)	46.76 (7.76)	31.13 (0.87)
	Population	1,271,909	1,171,041	1,271,909	1,171,041	1,271,909	1,171,041	1,271,909	1,171,041	1,271,909	1,171,041	1,271,909	1,171,041
	Rate	3.07	2.73	2.99 (0.08)	1.37 (1.37)	3.3 (-0.24)	1.79 (0.94)	3.46 (-0.40)	2.39 (0.34)	3.46 (-0.39)	2.34 (0.40)	3.68 (-0.61)	2.66 (0.07)
Zone C > 25 to 50 miles	Deaths	35	24	30 (5)	30 (6)	26 (9)	25 (1)	36.05 (1.05)	30.98 (6.98)	30.16 (4.84)	26.56 (2.56)	31.33 (3.67)	25.99 (1.99)
	Population	1,024,270	961,549	1,024,270	961,549	1,024,270	961,549	1,024,270	961,549	1,024,270	961,549	1,024,270	961,549
	Rate	3.42	2.50	2.93 (0.49)	3.12 (-0.62)	2.54 (0.88)	2.6 (-0.10)	3.52 (-0.10)	3.22 (-0.73)	2.94 (0.47)	2.76 (-0.27)	3.06 (0.36)	2.70 (-0.20)
Zone D > 50 miles	Deaths	38	33	40 (2)	34 (1)	40 (2)	34 (1)	40.51 (2.51)	34.13 (1.13)	39.75 (1.75)	33.69 (0.69)	41.14 (3.14)	34.88 (1.88)
	Population	1,397,448	1,330,320	1,397,448	1,330,320	1,397,448	1,330,320	1,397,448	1,330,320	1,397,448	1,330,320	1,397,448	1,330,320
	Rate	2.72	2.48	2.86 (-0.14)	2.56 (-0.08)	2.86 (-0.14)	2.56 (-0.08)	2.9 (-0.18)	2.57 (-0.09)	2.84 (-0.13)	2.53 (-0.05)	2.94 (-0.22)	2.62 (-0.14)
Mean **absolute** arithmetic difference from paired Georgia death counts				5.50		5.00		4.18		2.84		3.43	
and standard deviation				5.04		4.04		2.67		1.72		2.62	
Mean of arithmetic differences from paired Georgia death rates				-0.12		-0.10		0.10		0.01		0.15	
and standard deviation*				0.83		0.77		0.65		0.33		0.42	

Rate per 100,000 per year;
1999-2011 person-year population estimated using simple areal weighting with 2000 and 2010 tract-level source zones, then weighted by year;
Methods 3 through 5 use death estimates expressed as real numbers, in contrast to the centroid methods which use integer counts;
Value in parentheses is the arithmetic difference between the estimate and its paired state of Georgia value;
We use absolute arithmetic differences for death counts because if we used actual differences, the mean difference would equal 0;
*Standard deviations of the mean arithmetic difference from paired Georgia deaths rates are distinct from those shown in the Bland-Altman plots, which are standard deviations of the mean of the paired rates.

Zone	Denominator - Census Tract Source Zones				Numerator - County Source Zones			
	% Degree of Hierarchy		% Degree of Fit		% Degree of Hierarchy		% Degree of Fit	
	Georgia	U.S.	Georgia	U.S.	Georgia	U.S.	Georgia	U.S.
A	76.4	84.8	87.0	92.0	0.0	4.3	18.2	24.3
B	62.2	61.6	81.1	81.8	3.4	0.9	41.4	36.0
C	51.0	57.8	75.6	78.9	5.3	3.8	53.3	49.3
D	81.8	78.2	91.7	91.7	60.5	53.9	81.4	80.1
Overall	81.7	83.7	96.6	97.0	52.2	45.1	88.7	87.2

Degree of hierarchy (nesting) and degree of fit (overlap) between source and target study zones. The higher the percentage, the better the estimate.

The Bland-Altman plots compare 1999-2011 Georgia adolescent mortality rate estimates to estimated rates for methods 1 through 5. Method 4 demonstrates the greatest agreement.



Key Findings

Among the five numerator estimation methods tested, Method 4, Combined Population and Areal Weighting returned the best results. Method 4 had the lowest mean absolute value difference between the estimated death counts and the observed Georgia counts, and generated the only strongly positive correlation (r=0.63) with the estimated Georgia rates. However, correlation tests, which support the selection of Method 4 as the optimal method, are inadequate to completely assess the accuracy of an estimation method. A more definitive measure of method performance is that of agreement. To visualize agreement, we used Bland-Altman plots which display the means of each pair of estimates - the Georgia rates compared to each of the five methods - against the difference between the estimates. Method 4 again produced the best results, with each of the eight data points falling within small 95% limits of agreement.

Combined Population and Areal Weighting incorporated ancillary census tract data to weight deaths by the study populations, the intent being to reduce the error associated with assuming an evenly distributed population across counties. In Method 4, unlike centroid methods or Simple Areal Weighting, error is also distributed across the study zones by allocating "mortality" in proportion to population. Although it is more processing-intensive than the other methods, the processing can be automated. Further, Method 4 is conceptually simple, whereas Geostatistical Areal Interpolation, which produced moderate results, requires expert knowledge of geostatistical methods.

Adolescent cancer mortality counts from the GADPH were appropriate for testing the methods explored. The distribution of county mortality counts for Georgia mirror those of the U.S. Likewise, measures of hierarchy and fit, as well as patterns of zone values are roughly similar for the state and the nation. In terms of area, however, medium-sized Georgia has some of the smallest counties in the country (N=159) and therefore may not be representative of other U.S. states. The mean number of mortalities per county is 1.50 vs. 2.45 for the U.S. as a whole. It may be that smaller counties return better results than larger counties for the four tested methods. However, as Method 4 distributes error across target zones, we would still expect to observe improved estimation over the centroid methods in regions of the country with larger counties. With Georgia's smaller counties, improvements over the other methods in this study should be seen as conservative.

Conclusion

This research demonstrates that Combined Population and Areal Weighting, compared to other areal interpolation methods examined here, returns the most accurate estimates of mortality in transforming small counts by county to aggregated counts for large, non-standard enumeration zones. This methodology should be of interest to practitioners and researchers limited to analysis of relatively large enumeration units, such as NCHS county-level mortality data, due to data confidentiality concerns.

Centers for Disease Control and Prevention
Agency for Toxic Substances and Disease Registry

Contact: Elaine J. Hallisey (ehallisey@cdc.gov) Geospatial Research, Analysis, and Services Program, Division of Toxicology and Human Health Sciences, Agency for Toxic Substances and Disease Registry | Project 04355 | 6/26/2017