The NCI / NAACCR Cancer Reporting Zone Project

Zaria Tatalovich, Dave Stinchcomb, Recinda Sherman

NIH NATIONAL CANCER INSTITUTE

NAACCR
North American Association of Central Cancer Registries

October 23, 2019

Acknowledgements

- NCI
  - Zaria Tatalovich
  - Mandi Yu
  - Denise Lewis
  - Li Zhu
  - Rocky Feuer
- NAACCR
  - Recinda Sherman
- California
  - Scarlett Gomez
  - Debby Oh
  - Salma Shariff-Marco
- Louisiana
  - Lauren Maniscalco
  - Yong Yi
  - Tina Lefante
- Westat
  - Dave Stinchcomb
  - Matt Airola
  - Diane Ng

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

Background / motivation

- County is not a very satisfactory geographic unit to use for cancer reporting
  - Larger counties often have very heterogeneous populations
  - Data for smaller counties often suppressed due to small numbers

Los Angeles County, CA
Pop: over 10 million

Loving County, TX
Pop: 134

- Census tracts (or collections of census tracts) are a much better unit for analysis, but are generally unavailable because of identifiability issues
  - NCI and NAACCR have worked to make proxies for census tract available
    - Census tract poverty, SES, and urbanicity variables

Idea for this project

- We envision a common and stable set of zones for long-term cancer reporting
- Establish a minimum population size for release of cancer data (e.g. 20,000, 50,000)
- Work with registries to define zones for cancer reporting as groups of neighboring census tracts
- Zones would be formed having desirable properties such as homogeneity and compactness
- Since we historically release data at the county level, we assume that we would continue to do this

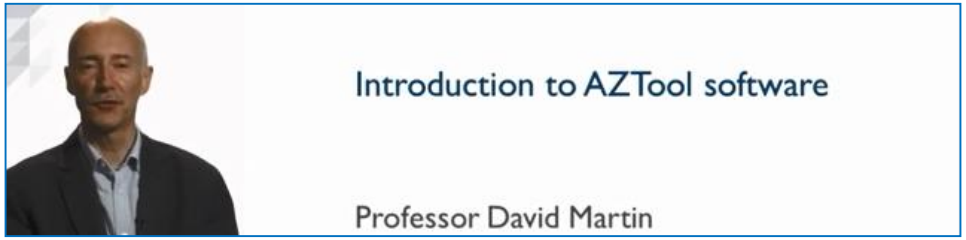
Goals

- Develop a set of cancer reporting zones that will:
  - Provide greater spatial resolution for large counties
  - Reduce suppression of data for small counties
  - Provide more meaningful data for communities & stakeholders
- Establish a common zone design method that can be applied to all states (with some flexibility)
- Work with individual states to apply the method:
  - Currently finalizing details for California and Louisiana zones with registry representatives
  - Invite other U.S. registries to participate (today)

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

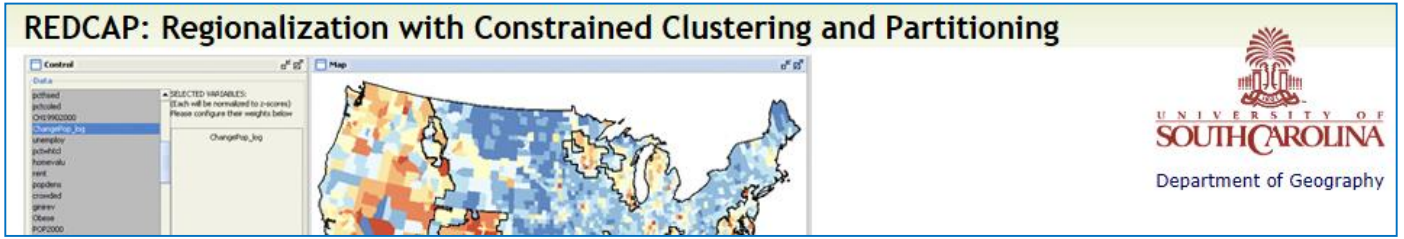
Evaluated three zone design tools

- AZTool

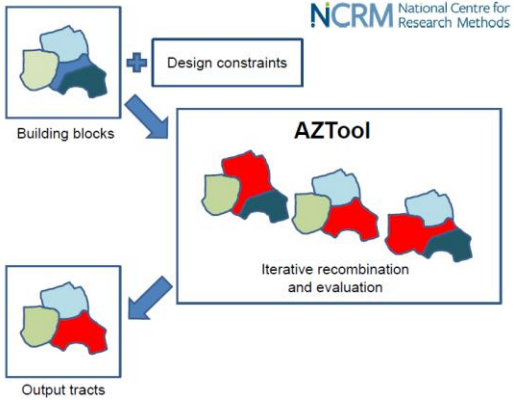


- GAT

- REDCAP



Comparison of methods



- AZTool
  - Random initial assignment
  - Iterative refinement to optimize the objective function
- GAT
  - Identify areas that do not meet the minimum population threshold
  - Pick a neighbor to merge:
    - Closest, smallest population, or most similar
- REDCAP
  - Statistical clustering with contiguity constraints
  - Partition the results to optimize the objective function

Tool comparison summary

- AZTool
  - Very flexible choice of objectives
  - Strong pedigree – used to define UK statistical reporting areas
  - User interface is fairly primitive

- GAT
  - Nicer user interface
  - Limited choice of objective functions
  - Simple assignment – does not seek the best aggregation
  - Some issues with both the R and SAS versions
- REDCAP
  - Does not meet basic needs: must specify desired number of zones and there is no compactness objective

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

Target population size

- What should the target population be?
  - Zones with smaller populations will have more geospatial resolution
  - Zones with larger populations will have fewer suppressed cells
- HIPAA minimum population size: 20,000
- If zones with 15 or fewer cancer cases are suppressed, how much suppression will there be?
  - By site; by site & sex; by site, sex, & race/ethnicity
- Ideally, the same population size for all sites
- Can reduce suppression by aggregating years
  - Case count estimates 1-year, 5-years, 10-years

Estimating population needed for 16 cases – crude rates

Site	Crude rate per 100,000 (percentile of SEER counties) 25th pctl	50th pctl
All Sites	483.5	566.2
Breast (female)	127.4	146.8
Lung and Bronchus	64.6	85.4
Prostate (male)	107.3	130.0
Colon and Rectum	42.9	53.9
Urinary Bladder	18.2	24.1
Melanoma of the Skin	18.5	26.0
Non-Hodgkin Lymphoma	18.0	22.2
Kidney and Renal Pelvis	16.6	20.8
Leukemias	13.4	16.6
Corpus and Uterus, NOS (female)	24.0	31.3
Oral Cavity and Pharynx	12.3	15.6
Pancreas	12.6	15.6
Thyroid	10.0	13.8
Liver and Intrahepatic Bile Duct	6.9	9.3
Myeloma	6.0	7.8
Stomach	5.5	7.3
Brain and Other Nervous System	5.5	7.2
Ovary (female)	9.8	13.0
Esophagus	4.0	5.6
Larynx	3.0	4.9
Cervix Uteri (female)	5.5	7.7
Hodgkin Lymphoma	1.7	2.5

Reportable cancer sites – minimum population 20,000

	Crude rate per 100,000 (percentile of SEER counties)		Population* needed to have 16 cases in 1 year		Population* needed to have 16 cases in 5 years		Population* needed to have 16 cases in 10 years	
Site	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl
All Sites	483.5	566.2	3,309	2,826	662	565	331	283
Breast (female)	127.4	146.8	25,123	21,798	5,025	4,360	2,512	2,180
Lung and Bronchus	64.6	85.4	24,786	18,737	4,957	3,747	2,479	1,874
Prostate (male)	107.3	130.0	29,827	24,609	5,965	4,922	2,983	2,461
Colon and Rectum	42.9	53.9	37,297	29,701	7,459	5,940	3,730	2,970
Urinary Bladder	18.2	24.1	87,736	66,493	17,547	13,299	8,774	6,649
Melanoma of the Skin	18.5	26.0	86,398	61,604	17,280	12,321	8,640	6,160
Non-Hodgkin Lymphoma	18.0	22.2	88,965	71,974	17,793	14,395	8,896	7,197
Kidney and Renal Pelvis	16.6	20.8	96,403	76,773	19,281	15,355	9,640	7,677
Leukemias	13.4	16.6	119,592	96,230	23,918	19,246	11,959	9,623
Corpus and Uterus, NOS (female)	24.0	31.3	133,072	102,270	26,614	20,454	13,307	10,227
Oral Cavity and Pharynx	12.3	15.6	130,317	102,365	26,063	20,473	13,032	10,237
Pancreas	12.6	15.6	127,053	102,397	25,411	20,479	12,705	10,240
Thyroid	10.0	13.8	159,764	115,656	31,953	23,131	15,976	11,566
Liver and Intrahepatic Bile Duct	6.9	9.3	232,274	171,154	46,455	34,231	23,227	17,115
Myeloma	6.0	7.8	265,474	206,127	53,095	41,225	26,547	20,613
Stomach	5.5	7.3	292,359	220,164	58,472	44,033	29,236	22,016
Brain and Other Nervous System	5.5	7.2	290,332	223,676	58,066	44,735	29,033	22,368
Ovary (female)	9.8	13.0	327,214	245,583	65,443	49,117	32,721	24,558
Esophagus	4.0	5.6	395,260	283,551	79,052	56,710	39,526	28,355
Larynx	3.0	4.9	538,720	327,601	107,744	65,520	53,872	32,760
Cervix Uteri (female)	5.5	7.7	584,906	415,886	116,981	83,177	58,491	41,589
Hodgkin Lymphoma	1.7	2.5	936,620	642,309	187,324	128,462	93,662	64,231

* Populations have been doubled for sex-specific cancer sites to reflect approximate total population

Reportable cancer sites – minimum population 50,000

Site	Crude rate per 100,000 (percentile of SEER counties) 25th pctl	50th pctl	Population* needed to have 16 cases in 1 year 25th pctl	50th pctl	Population* needed to have 16 cases in 5 years 25th pctl	50th pctl	Population* needed to have 16 cases in 10 years 25th pctl	50th pctl
All Sites	483.5	566.2	3,309	2,826	662	565	331	283
Breast (female)	127.4	146.8	25,123	21,798	5,025	4,360	2,512	2,180
Lung and Bronchus	64.6	85.4	24,786	18,737	4,957	3,747	2,479	1,874
Prostate (male)	107.3	130.0	29,827	24,609	5,965	4,922	2,983	2,461
Colon and Rectum	42.9	53.9	37,297	29,701	7,459	5,940	3,730	2,970
Urinary Bladder	18.2	24.1	87,736	66,493	17,547	13,299	8,774	6,649
Melanoma of the Skin	18.5	26.0	86,398	61,604	17,280	12,321	8,640	6,160
Non-Hodgkin Lymphoma	18.0	22.2	88,965	71,974	17,793	14,395	8,896	7,197
Kidney and Renal Pelvis	16.6	20.8	96,403	76,773	19,281	15,355	9,640	7,677
Leukemias	13.4	16.6	119,592	96,230	23,918	19,246	11,959	9,623
Corpus and Uterus, NOS (female)	24.0	31.3	133,072	102,270	26,614	20,454	13,307	10,227
Oral Cavity and Pharynx	12.3	15.6	130,317	102,365	26,063	20,473	13,032	10,237
Pancreas	12.6	15.6	127,053	102,397	25,411	20,479	12,705	10,240
Thyroid	10.0	13.8	159,764	115,656	31,953	23,131	15,976	11,566
Liver and Intrahepatic Bile Duct	6.9	9.3	232,274	171,154	46,455	34,231	23,227	17,115
Myeloma	6.0	7.8	265,474	206,127	53,095	41,225	26,547	20,613
Stomach	5.5	7.3	292,359	220,164	58,472	44,033	29,236	22,016
Brain and Other Nervous System	5.5	7.2	290,332	223,676	58,066	44,735	29,033	22,368
Ovary (female)	9.8	13.0	327,214	245,583	65,443	49,117	32,721	24,558
Esophagus	4.0	5.6	395,260	283,551	79,052	56,710	39,526	28,355
Larynx	3.0	4.9	538,720	327,601	107,744	65,520	53,872	32,760
Cervix Uteri (female)	5.5	7.7	584,906	415,886	116,981	83,177	58,491	41,589
Hodgkin Lymphoma	1.7	2.5	936,620	642,309	187,324	128,462	93,662	64,231

* Populations have been doubled for sex-specific cancer sites to reflect approximate total population

Reporting by site and sex

		Crude rate per 100,000 (percentile of SEER counties)		Population needed to have 16 cases in 1 year		Population needed to have 16 cases in 5 years		Population needed to have 16 cases in 10 years	
Site	Sex	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl
All Sites	Male	499.3	598.1	3,205	2,675	641	535	320	268
All Sites	Female	457.4	536.1	3,498	2,985	700	597	350	298
Breast (female)	Female	127.4	146.8	12,561	10,899	2,512	2,180	1,256	1,090
Prostate (male)	Male	107.3	130.0	14,914	12,304	2,983	2,461	1,491	1,230
Lung and Bronchus	Male	68.5	96.9	23,343	16,518	4,669	3,304	2,334	1,652
Lung and Bronchus	Female	56.4	73.1	28,366	21,884	5,673	4,377	2,837	2,188
Colon and Rectum	Male	44.5	57.6	35,919	27,775	7,184	5,555	3,592	2,778
Colon and Rectum	Female	38.4	49.0	41,688	32,621	8,338	6,524	4,169	3,262
Urinary Bladder	Male	28.0	37.2	57,245	42,971	11,449	8,594	5,724	4,297
Corpus and Uterus, NOS (female)	Female	24.0	31.3	66,536	51,135	13,307	10,227	6,654	5,113
Melanoma of the Skin	Male	22.3	29.7	71,753	53,815	14,351	10,763	7,175	5,381
Kidney and Renal Pelvis	Male	20.4	25.8	78,546	62,108	15,709	12,422	7,855	6,211
Non-Hodgkin Lymphoma	Male	19.5	24.7	82,060	64,764	16,412	12,953	8,206	6,476
Oral Cavity and Pharynx	Male	17.2	21.9	93,172	72,904	18,634	14,581	9,317	7,290
Leukemias	Male	15.1	19.2	105,932	83,161	21,186	16,632	10,593	8,316
Thyroid	Female	14.8	20.3	108,056	78,899	21,611	15,780	10,806	7,890
Non-Hodgkin Lymphoma	Female	14.8	19.3	108,086	83,023	21,617	16,605	10,809	8,302
Melanoma of the Skin	Female	13.6	20.5	117,221	77,999	23,444	15,600	11,722	7,800
Pancreas	Male	12.2	15.8	130,727	101,460	26,145	20,292	13,073	10,146
Kidney and Renal Pelvis	Female	11.3	14.6	142,087	109,282	28,417	21,856	14,209	10,928
Pancreas	Female	11.0	14.8	145,946	108,466	29,189	21,693	14,595	10,847
Leukemias	Female	10.1	13.4	159,087	119,645	31,817	23,929	15,909	11,964
Ovary (female)	Female	9.8	13.0	163,607	122,791	32,721	24,558	16,361	12,279
Liver and Intrahepatic Bile Duct	Male	9.7	13.4	165,236	119,732	33,047	23,946	16,524	11,973
Urinary Bladder	Female	6.9	10.5	230,837	152,215	46,167	30,443	23,084	15,221
Stomach	Male	6.5	9.2	245,363	174,277	49,073	34,855	24,536	17,428
Myeloma	Male	6.4	9.0	248,451	178,503	49,690	35,701	24,845	17,850
Esophagus	Male	6.3	9.1	252,385	175,337	50,477	35,067	25,238	17,534
Brain and Other Nervous System	Male	5.6	8.1	283,732	196,943	56,746	39,389	28,373	19,694
Oral Cavity and Pharynx	Female	5.6	8.2	288,118	194,114	57,624	38,823	28,812	19,411
Cervix Uteri (female)	Female	5.5	7.7	292,453	207,943	58,491	41,589	29,245	20,794
Larynx	Male	4.7	7.6	341,928	209,437	68,386	41,887	34,193	20,944
Myeloma	Female	4.6	6.5	349,099	246,570	69,820	49,314	34,910	24,657
Thyroid	Male	4.4	6.9	365,162	231,110	73,032	46,222	36,516	23,111
Brain and Other Nervous System	Female	4.3	6.0	374,471	267,065	74,894	53,413	37,447	26,707
Stomach	Female	3.2	5.0	497,229	320,073	99,446	64,015	49,723	32,007
Liver and Intrahepatic Bile Duct	Female	3.2	4.9	503,159	329,534	100,632	65,907	50,316	32,953
Testis (male)	Male	3.1	4.8	524,560	330,074	104,912	66,015	52,456	33,007
Hodgkin Lymphoma	Male	1.3	2.7	1,231,364	587,984	246,273	117,597	123,136	58,798
Esophagus	Female	0.0	1.8		872,812		174,562		87,281
Hodgkin Lymphoma	Female	0.0	2.2		743,143		148,629		74,314

Reporting by site and race/ethnicity

		Crude rate per 100,000 (percentile of SEER counties)		Population* needed to have 16 cases in 1 year		Population* needed to have 16 cases in 5 years		Population* needed to have 16 cases in 10 years	
Site	RaceEth	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl	25th pctl	50th pctl
All Sites	Black	214.1	363.1	7,472	4,407	1,494	881	747	441
All Sites	Hispanic	82.6	143.9	19,367	11,122	3,873	2,224	1,937	1,112
Prostate (male)	Black	0.0	123.1		26,003		5,201		2,600
Breast (female)	Black	0.0	95.1		33,646		6,729		3,365
Lung and Bronchus	Black	0.0	45.2		35,369		7,074		3,537
Colon and Rectum	Black	0.0	34.3		46,714		9,343		4,671
Breast (female)	Hispanic	0.0	34.5		92,688		18,538		9,269
Kidney and Renal Pelvis	Black	0.0	11.0		144,864		28,973		14,486
Non-Hodgkin Lymphoma	Black	0.0	8.0		201,225		40,245		20,122
Pancreas	Black	0.0	7.7		208,674		41,735		20,867
Prostate (male)	Hispanic	0.0	14.6		218,994		43,799		21,899
Myeloma	Black	0.0	7.1		224,819		44,964		22,482
Colon and Rectum	Hispanic	0.0	7.0		229,883		45,977		22,988
Corpus and Uterus, NOS (female)	Black	0.0	12.4		258,580		51,716		25,858
Lung and Bronchus	Hispanic	0.0	5.7		280,984		56,197		28,098
Leukemias	Black	0.0	5.3		304,283		60,857		30,428
Oral Cavity and Pharynx	Black	0.0	4.6		346,316		69,263		34,632
Urinary Bladder	Black	0.0	3.8		424,127		84,825		42,413
Stomach	Black	0.0	3.4		474,368		94,874		47,437
Liver and Intrahepatic Bile Duct	Black	0.0	3.3		481,916		96,383		48,192
Melanoma of the Skin	Hispanic	0.0	0.0						
Melanoma of the Skin	Black	0.0	0.0						
Urinary Bladder	Hispanic	0.0	0.0						
Non-Hodgkin Lymphoma	Hispanic	0.0	0.0						
Kidney and Renal Pelvis	Hispanic	0.0	0.0						
Corpus and Uterus, NOS (female)	Hispanic	0.0	0.0						
Thyroid	Hispanic	0.0	0.0						
Thyroid	Black	0.0	0.0						

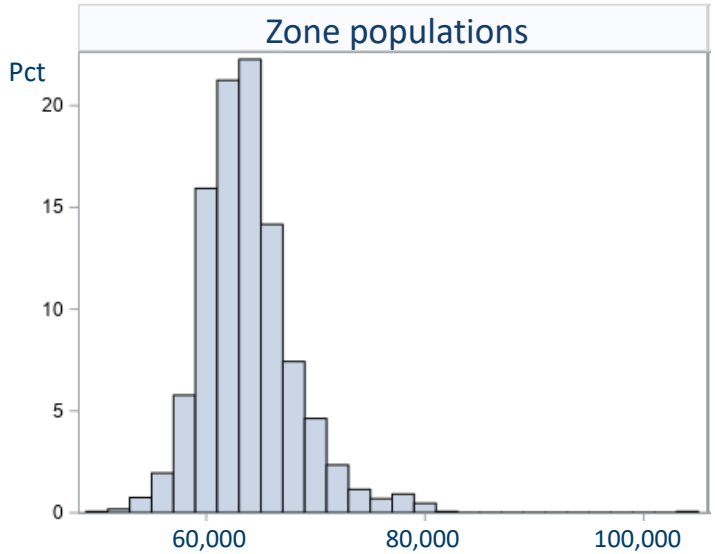
Zone design parameters

- Minimum and target population = 50,000
- Homogeneity objectives
  - Urbanicity
  - Percent below poverty
  - Percent minority
- Compactness objective
- Even weights among objectives

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

Simplest approach – a single step

- Aggregate tracts across the state specifying a minimum population of 50,000 in a single step
- Resulting zones have populations between 50,000 and 85,000



The differencing problem

- Differencing: a known problem in statistical disclosure control:
  - If tables are published for two sets of areas, users can compare the tables and produce new statistics for the areas formed by differencing, which may have populations below confidentiality thresholds.

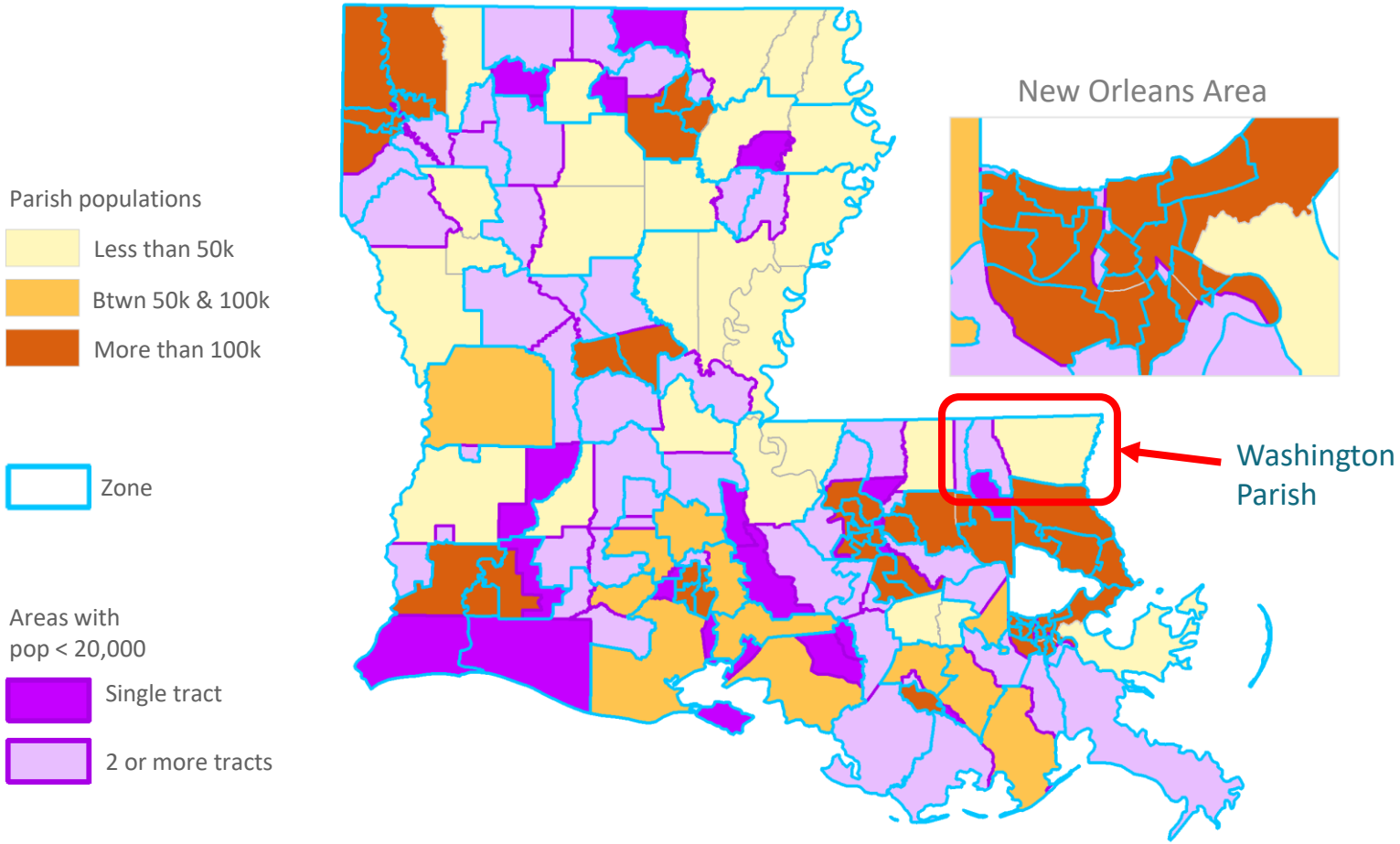
Reference: Duke-Williams & Rees, 1998

- Could the new zone data be compared with county data in this way?

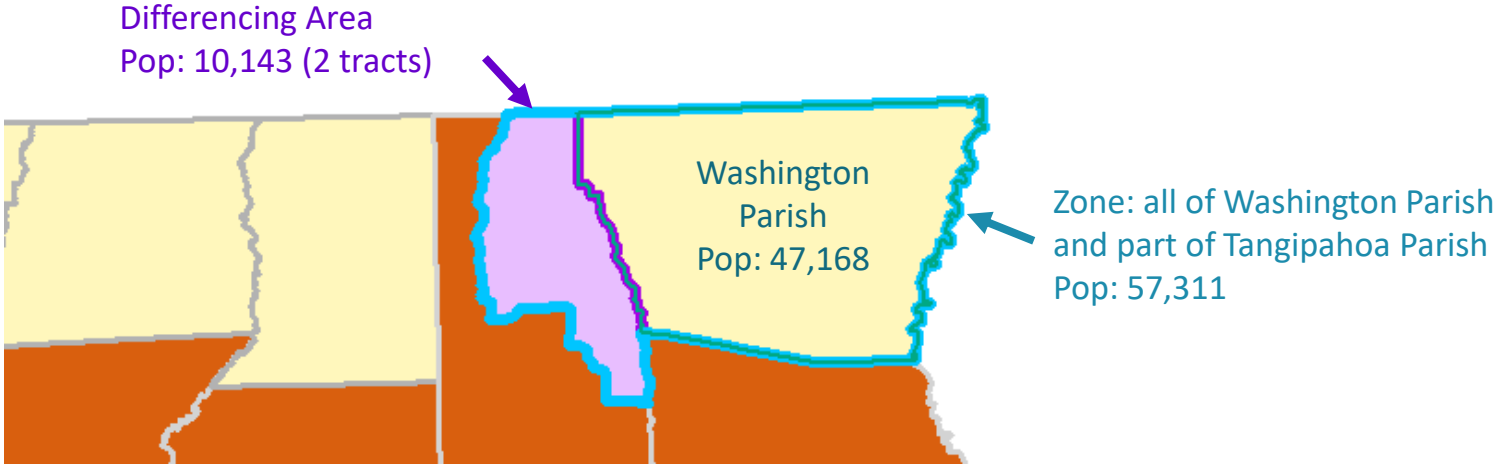
- Note: Census is moving toward differential privacy methods to keep confidentiality risk within quantifiable limits

Reference: Garfinkel et al, 2018

Potential differencing issues – Louisiana



Differencing example – Washington Parish



Hypothetical* 5-year cancer incidence data:

Area	Incidence Rate	Case Count	Population
Zone: Tangipahoa.Washington_1	69.8	20	57,311
Washington Parish	72.1	17	47,168
(differencing area)		3	10,143

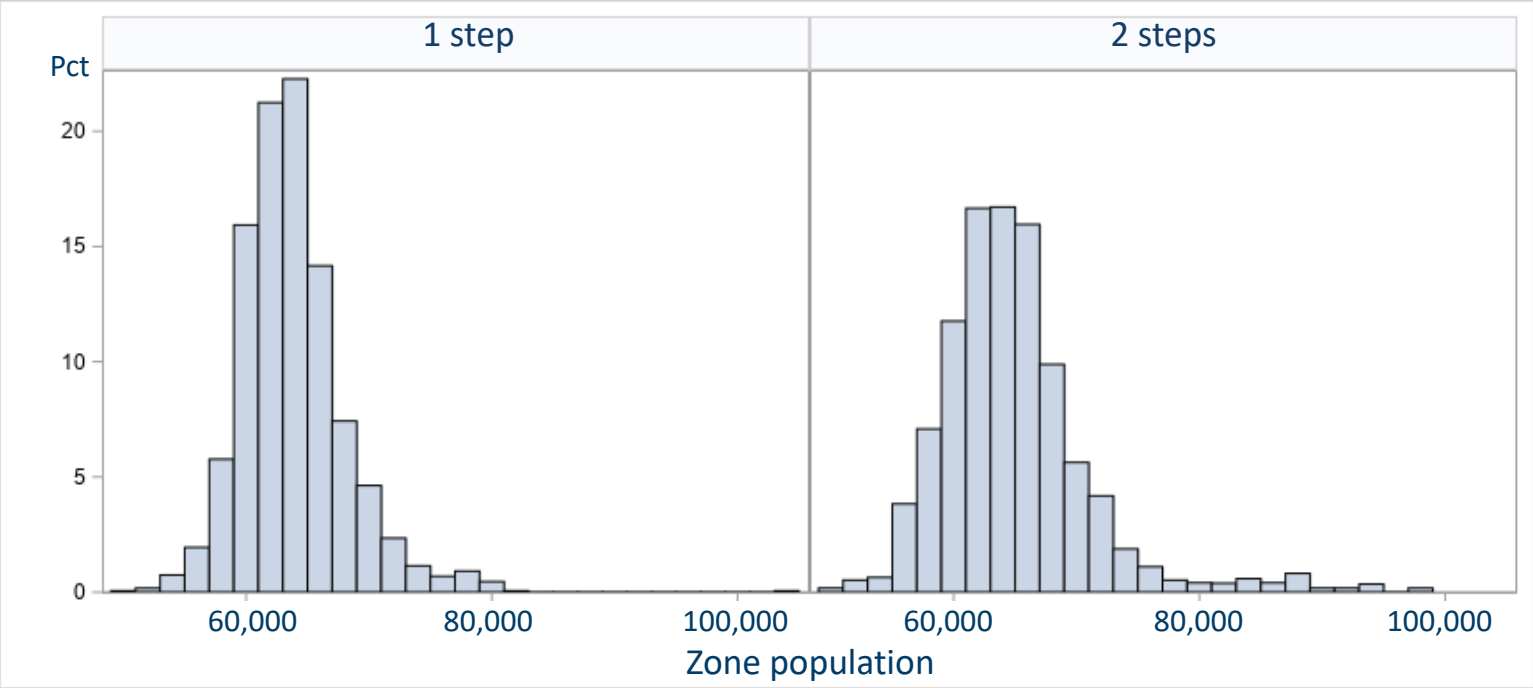
* Populations are real but incidence rates and case counts are made up

Solution: a 2-step process

- To protect against differencing, we've set up a 2-step process
- With the minimum population set to 50,000:
  - Step A: Aggregate census tracts in the large counties (populations over 100,000)
    - Zones cannot cross county boundaries
  - Step B: Aggregate:
    - the small and medium counties (populations less than 100,000)
    - with zones from Step A (with at least 50,000 people)
- Differencing areas between zones and counties will have at least 50,000

Zone populations: 1-step versus 2-step process

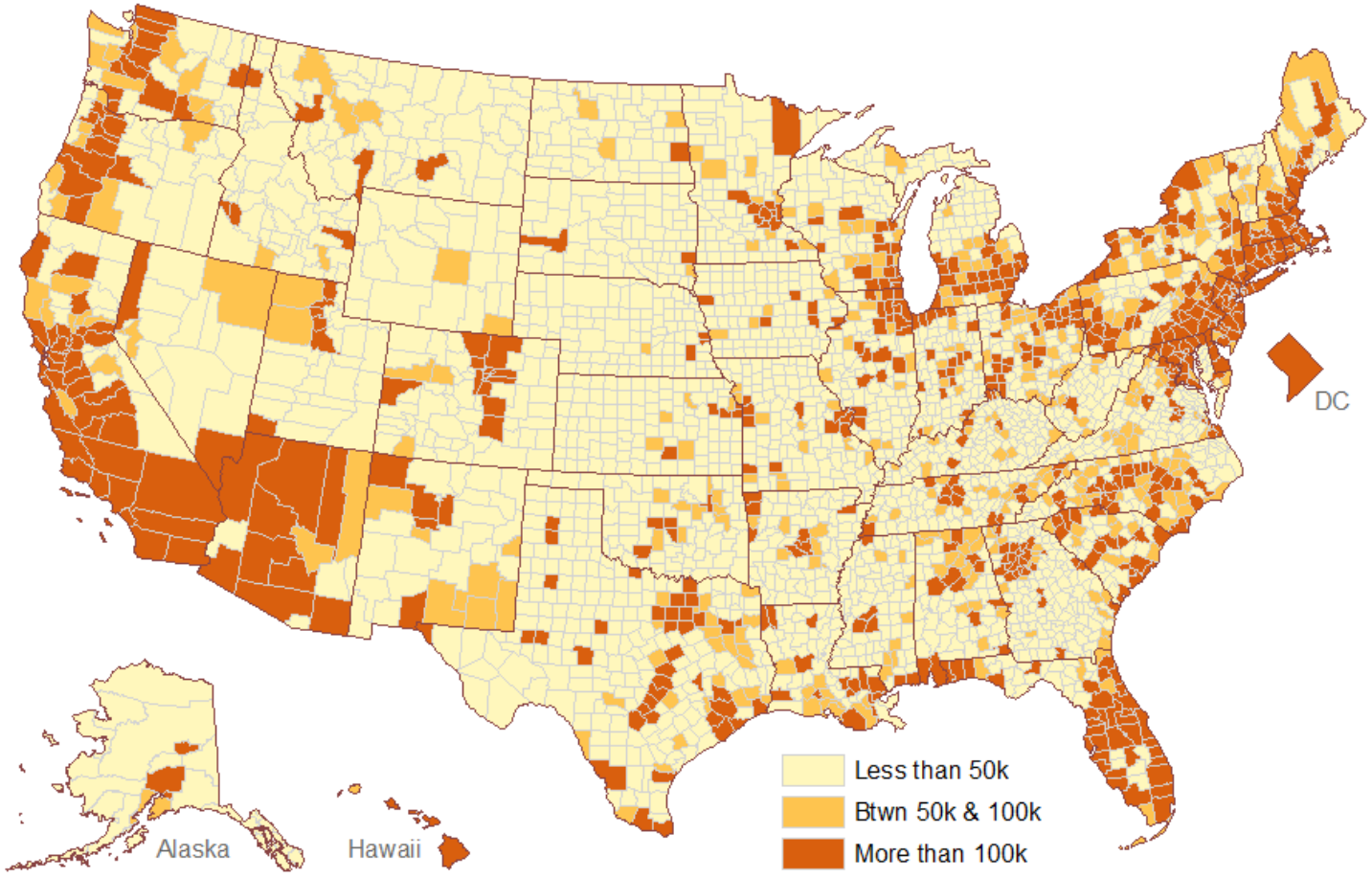
- The 2-step process results in zones with larger populations:



- An advantage of the larger populations is less suppression

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

County population sizes – 50,000 target



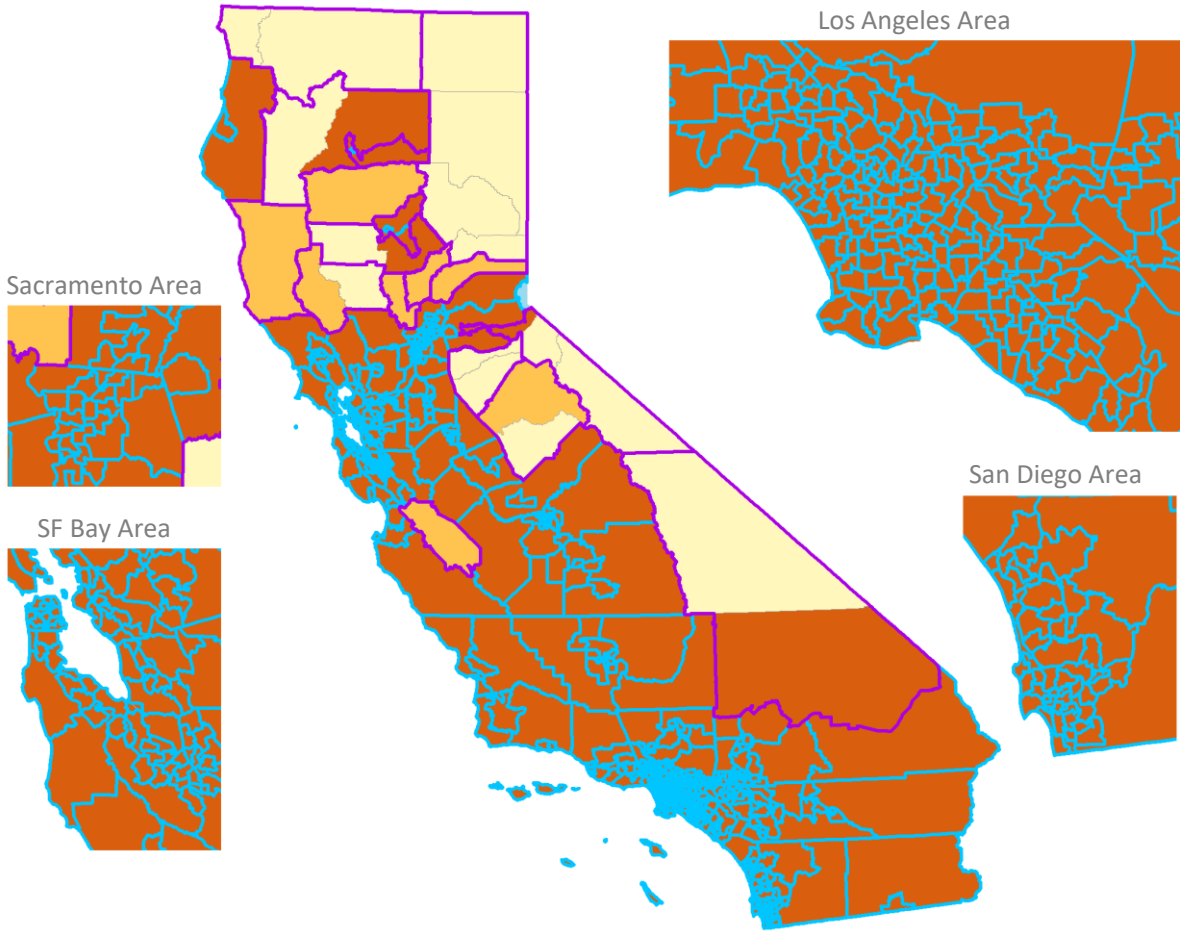
Results – 2-step zones in California

Step A: split up large counties

Step A Zone

Step B: combine with small and medium counties

Step B Zone



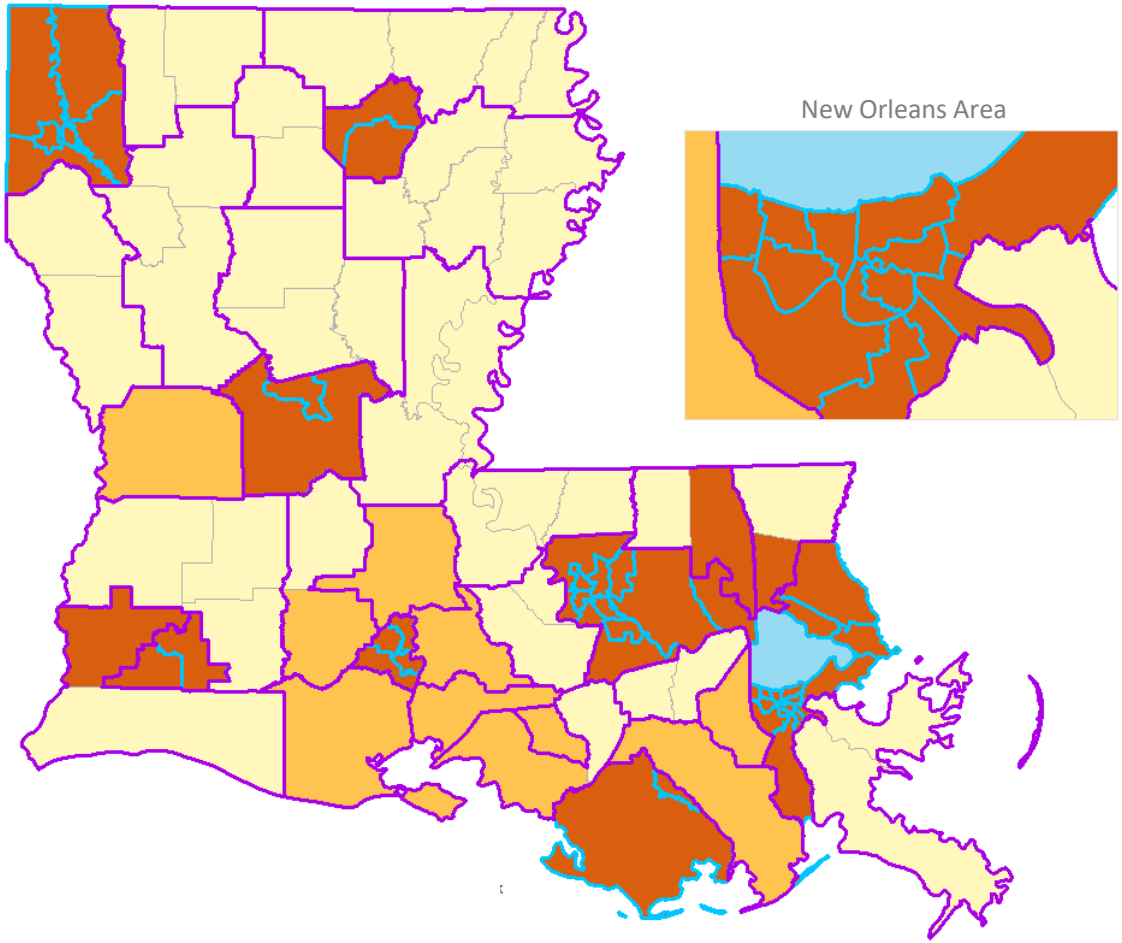
Results – 2-step zones in Louisiana

Step A: split up large counties

Step A Zone

Step B: combine with small and medium counties

Step B Zone



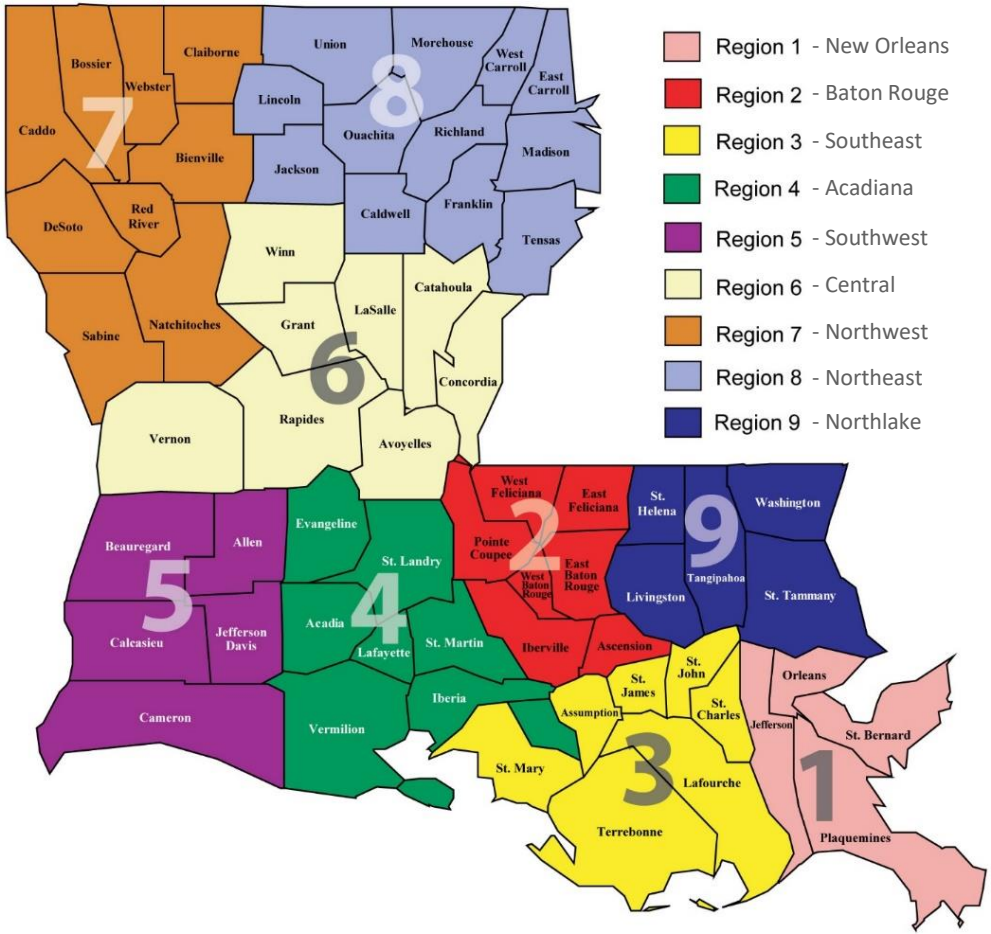
County populations

Less than 50k

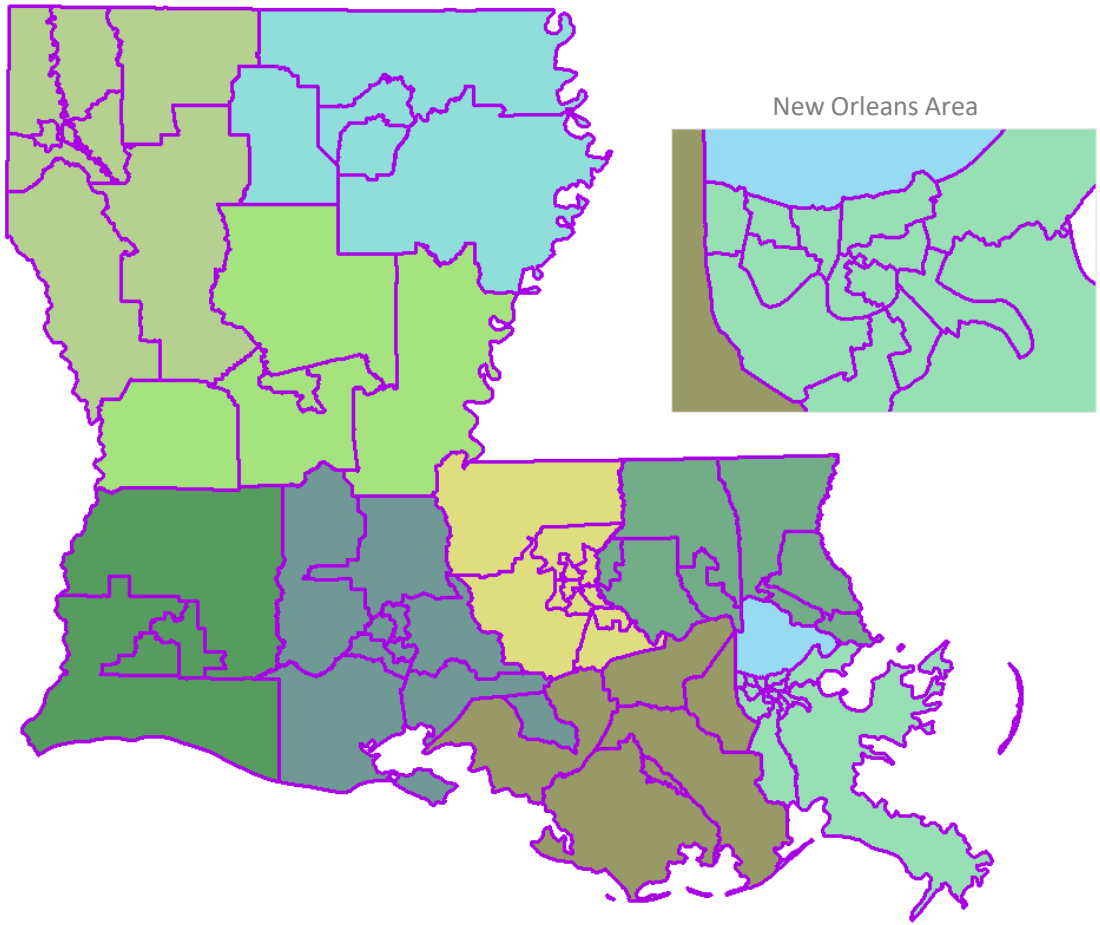
Btwn 50k & 100k

More than 100k

Louisiana Health Regions

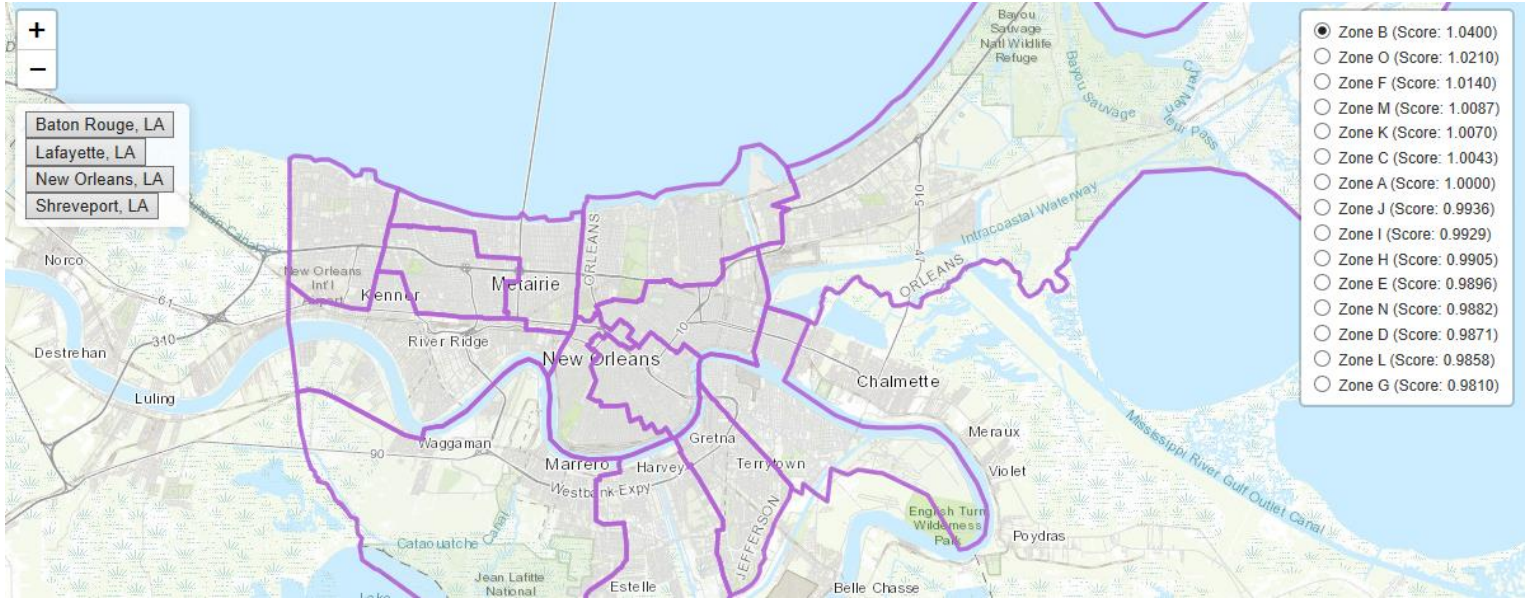


Louisiana zones respect Health Region boundaries



Alternative Zone Solutions

- Different runs of AZTool can produce slightly different zone design solutions
- States can select preferred alternative
  - Interactive maps to facilitate review



Current status

- So far, we've agreed that everyone will:
  - Use the 2-step process
  - Set the minimum population to 50,000
  - Seek homogeneous zones based on
    - Urbanicity
    - % below poverty
    - % minority
  - Include a compactness objective
- With flexibility for state-specific options:
  - Configure zones within existing health regions
  - Select preferred solution among alternatives

Plans for zone-level reporting

- Websites with cancer rates by zone
  - California example: CaliforniaHealthMaps.org
- SEER*Stat database support
- Report data by:
  - Site
  - Site and gender
  - Site, gender, and race/ethnicity
- Range of reporting years can vary to meet suppression requirements
  - 1 year for common cancers
  - 5-10 years for less common cancers or more detailed breakdowns

1. *Background and goals*
2. *Tool evaluation*
3. *Target population size*
4. *The differencing issue*
5. *California and Louisiana results*
6. *Next steps – invitation to participate*

Invitation to participate

- Partnering with NAACCR, we are inviting all U.S. registries to participate
  - Interested registries fill out a questionnaire
  - NCI/NAACCR selects 5-6 registries to start with
- We expect the process to take about 3 months
  - Rolling set of registries – when one finishes, a new one would be added
- Registries will need to:
  - Provide a point of contact, attend kickoff meeting
  - Provide info about relevant existing geographic areas
  - Review alternative zone solutions
  - Assist with zone naming
  - Help work through options for zone deployment

Outline of the process

- Phase 1 (about 4 weeks)
  - Kick-off meeting
  - Input/discussion about existing geographic areas and special considerations
  - Development of zone alternatives (Westat)
- Phase 2 (about 8-12 weeks)
  - Review alternative zone solutions, identify preferred
  - Assist with zone naming
  - Deploy an interactive tool for use with your registry
    - Based on a template provided by NCI/SEER
- Phase 3
  - Help us work through common methods for zone reporting across registries

Questions / discussion

www.cancer.gov **www.cancer.gov/espanol**

References

- Cockings, S., Harfoot, A., Martin, D., & Hornby, D. Maintaining existing zoning systems using automated zone design techniques: Methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A*, 2011 43, 2399–2418.
- Duke-Williams O, Rees P. Can census offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure, *International Journal of Geographical Information Science*. 1998 12:6, 579-605
- Flowerdew R, Manley DJ, Sabel CE. Neighbourhood effects on health: does it matter where you draw the boundaries? Soc Sci Med. 2008 Mar;66(6):1241-55.
- Garfinkel SL, Abowd JM, Powazek S. Issues encountered deploying differential privacy." In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 2018, WPES'18, 133-137
- Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP), *International Journal of Geographical Information Science*, 2008 22:7,801-823.
- Martin, D. Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 2003, 17:2, 181-196.
- Rossen, LM, Khan, D. Mapping Suicide Death Rates: Geographic Aggregation Tools and Spatial Smoothing with Hierarchical Bayesian Models. Presented at the FCSM Geospatial Interest Group Workshop, November 18, 2016.
- Sabel CE, Kihal W, Bard D, Weber C. Creation of synthetic homogeneous neighbourhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France. *Soc Sci Med.* 2013 Aug;91:110-21.
- Talbot TO, Done DH, Babcock GD. Calculating census tract-based life expectancy in New York state: a generalizable approach. *Popul Health Metr.* 2018 Jan 26;16(1):1.
- Tatalovich Z, Wilson JP, Milam JE, Jerrett ML, McConnell R. Competing definitions of contextual environments. *Int J Health Geogr.* 2006 Dec 7;5:55.
- Wang F, Guo D, McLafferty S. Constructing Geographic Areas for Cancer Data Analysis: A Case Study on Late-stage Breast Cancer Risk in Illinois. *Appl Geogr*. 2012 Nov;35(1-2):1-11