Designing Zones for Cancer Surveillance Reporting

Dave Stinchcomb, Zaria Tatalovich, Matt Airola, Mandi Yu, Li Zhu, Denise Lewis, Scarlett Gomez, Salma Shariff-Marco, Lauren Maniscalco, Yong Yi, Rocky Feuer

NAACCR Annual Conference, June 13, 2019

Preface – notes about this project

NCI is working on the development of a set of cancer reporting zones across the US that are more suitable for cancer data reporting than counties. In each respective state, the zones will be custom crafted to represent areas that:

1) are meaningful to stakeholders in terms of cancer reporting and cancer interventions;

2) comprise adjacent census tracts and smaller counties (or portions of counties) that sum to population sizes that are sufficiently large to support stable rates;

3) collectively cover the entire population of the state;

3) are homogeneous with respect to important socio-demographic characteristics and are compact in size;

4) have large enough case counts for data reporting, without compromising confidentiality; and

5) result in a relatively small proportion of areas with suppressed values, although for rarer cancer sites suppression will be inevitable especially when producing rates stratified by sex and/or race.

Research data released with these zones should be easy to access, with no special data use provisions.

The pilot study using cancer data from several cancer registries has been completed and the resulting zones satisfied the predefined criteria. These zones subdivide large population urban counties and are collections of smaller counties (or portions of counties) and have a minimum population size of 50,000. Our goal is to expand these “cancer-centric” zones to other registries and work with our cancer surveillance partners to release cancer statistics, and other socio-demographic factors relevant to understanding the cancer burden and identifying areas in need of interventions.

Westat

Improving Lives Through Research®

Designing Zones for Cancer Surveillance Reporting

Dave Stinchcomb[1], Zaria Tatalovich[2], Matt Airola[1], Mandi Yu[2], Li Zhu[2], Denise Lewis[2], Scarlett Gomez[3], Salma Shariff-Marco[3], Lauren Maniscalco[4], Yong Yi[4], Rocky Feuer[2]

1. Westat
2. National Cancer Institute
3. Greater Bay Area Cancer Registry
4. Louisiana Tumor Registry

June 13, 2019

Acknowledgements

- Louisiana
  - Xiao-Cheng Wu
  - Lauren Maniscalco
  - Yong Yi
  - Tina Lefante
  - Mei-chin Hsieh
  - Qingzhao Yu
- California
  - Scarlett Gomez
  - Salma Shariff-Marco
  - Debby Oh
  - Jenn Jain
- NCI
  - Zaria Tatalovich
  - Rocky Feuer
  - Mandi Yu
  - Denise Lewis
  - Li Zhu
- Westat
  - Matt Airola
  - Chichi Orji

Agenda

Background

Goals and objectives

Initial activities

- Tool evaluation
- Initial zone construction tests
- Picking a target population size

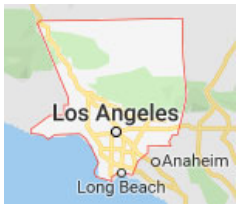
California and Louisiana testing

- The differencing problem and a 2-step process
- Recent results

Background / motivation

- In the U.S. most geospatial cancer reporting is based on counties
  - Large differences in population from hundreds to millions
  - Larger counties often have very heterogeneous populations
  - Data for smaller counties often suppressed due to small numbers
- U.S. census tracts, the next smaller full coverage census area, are too small
  - Target population of about 4,000
  - Too few people to support stable cancer rates
- Some states have developed their own sub-county areas
  - Vary in size and purpose

Loving County, TX
Pop: 134



Los Angeles County, CA
Pop: over 10 million

Goals

- Explore the feasibility of developing a set of cancer reporting zones to:
  - Provide greater spatial resolution for large counties
  - Reduce suppression of data for small counties
  - Provide more meaningful data for communities and stakeholders
- Work through details for California and Louisiana with registry representatives
- Develop a general process that could be applied to all U.S. states (and perhaps Canadian provinces and beyond)

Specific objectives

- Zones should be collections of neighboring census tracts
- Zones should have a similar number of people (with a minimum)
- Zones should be relatively *compact*
  - The distance from the center to any boundary does not vary significantly

- Zones should have a homogeneous population

Agenda

Background

Goals and objectives

Initial activities

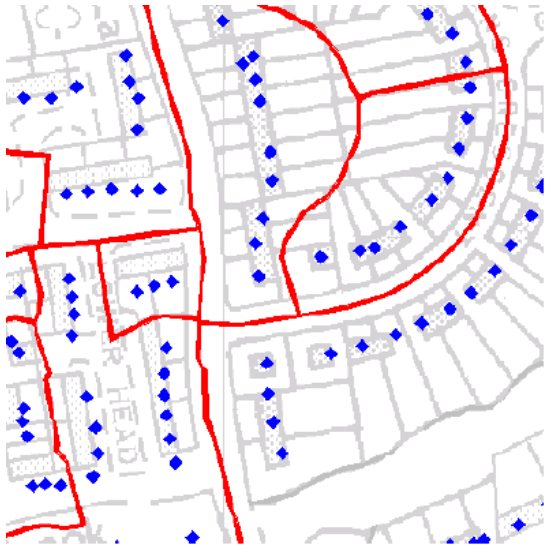
- Tool evaluation
- Initial zone construction tests
- Picking a target population size

California and Louisiana testing

- The differencing problem and a 2-step process
- Recent results

Existing zone design tools

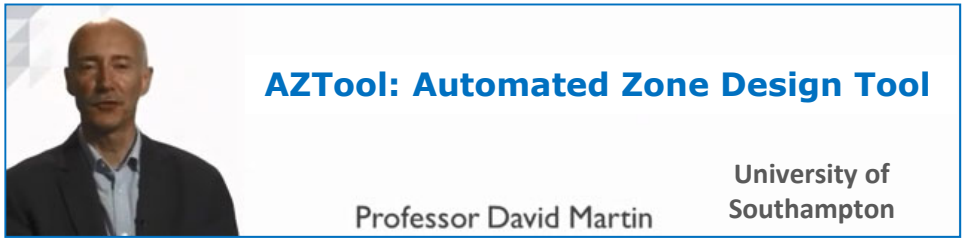
- Goals: combine spatially contiguous areas to achieve an objective function
  - Minimum / maximum population threshold
  - Homogeneity
  - Compactness
- Typical uses:
  - Statistical disclosure control
  - Survey sampling
  - Voting and electoral districts



- Other names: spatial aggregation, regionalization, spatial clustering

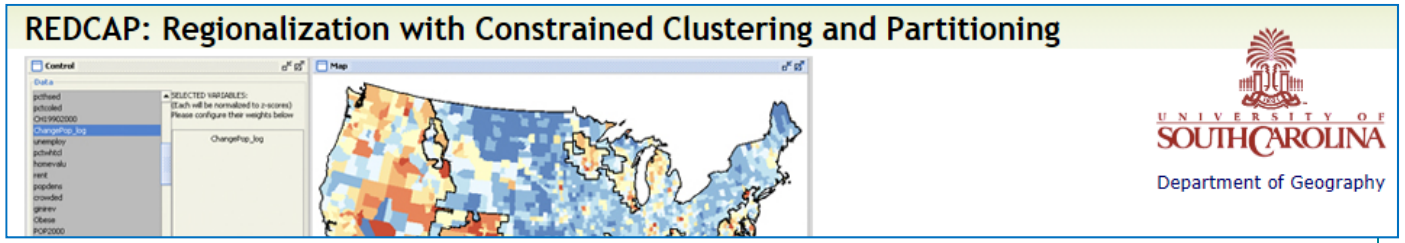
Evaluated three zone design tools

- AZTool

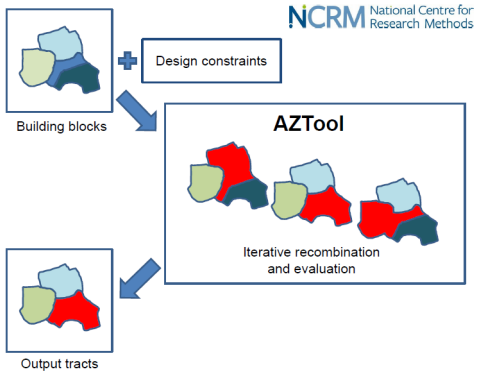


- GAT

- 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

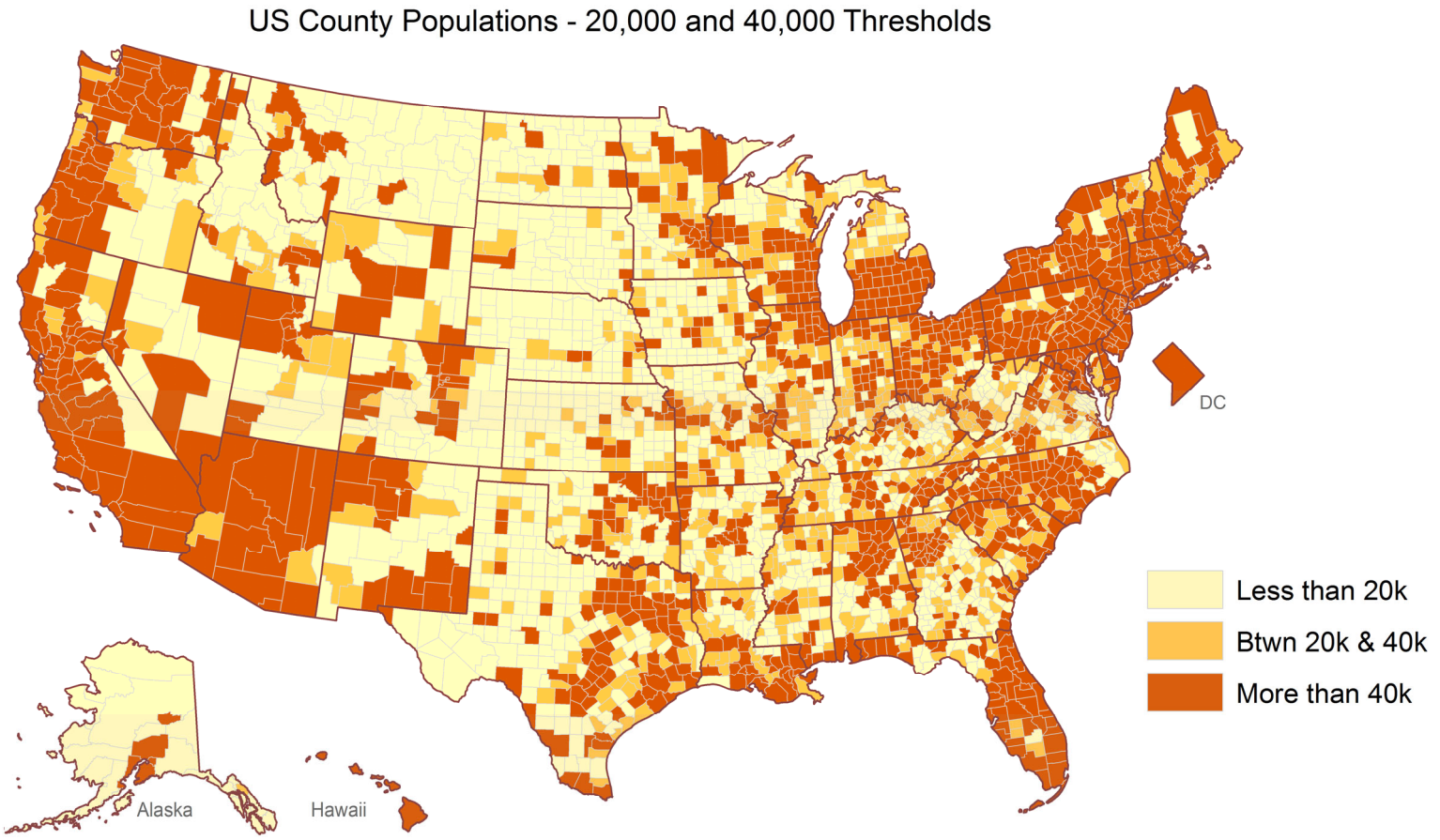
Tool comparison summary

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

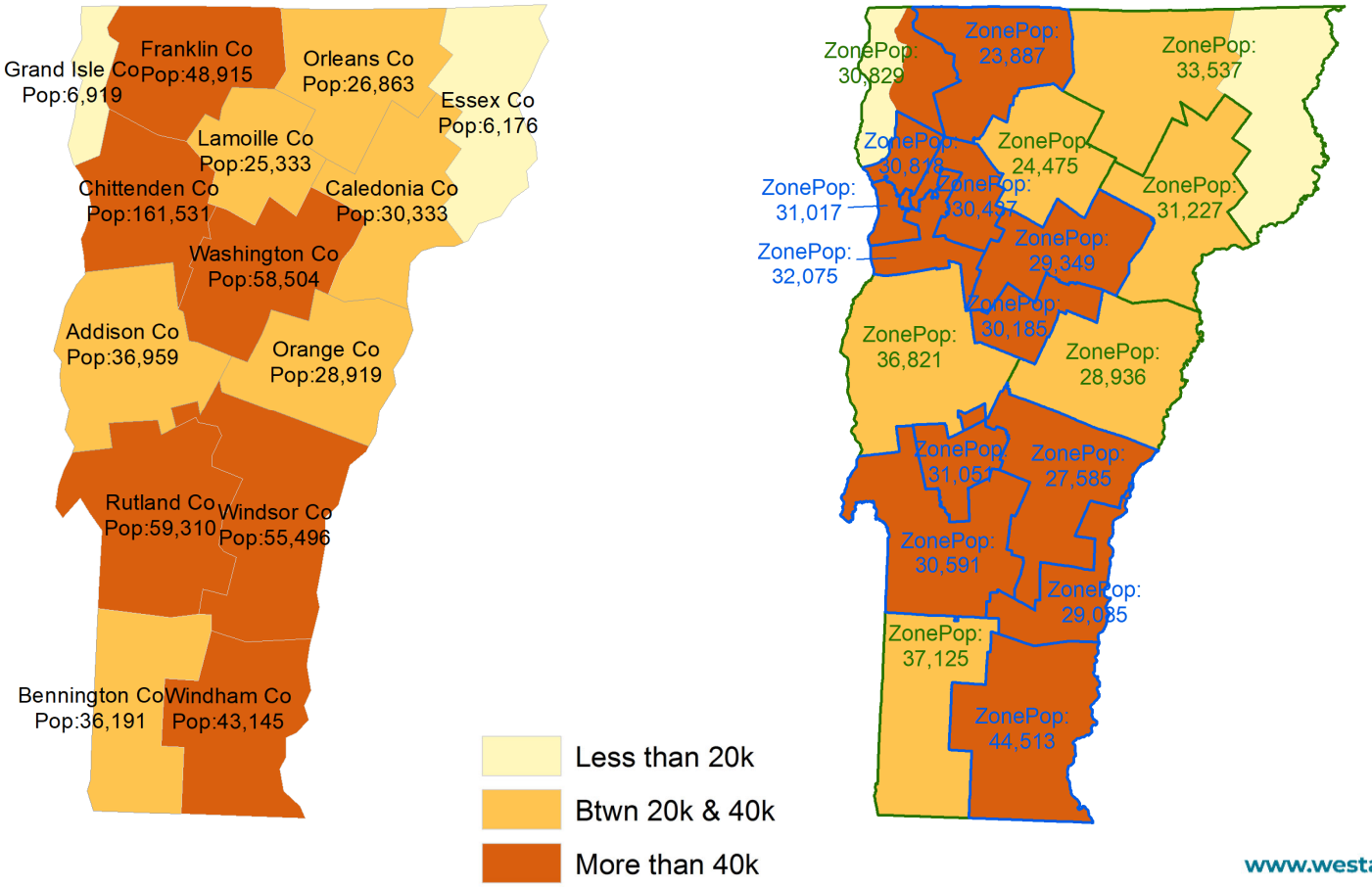
Our Choice

- 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

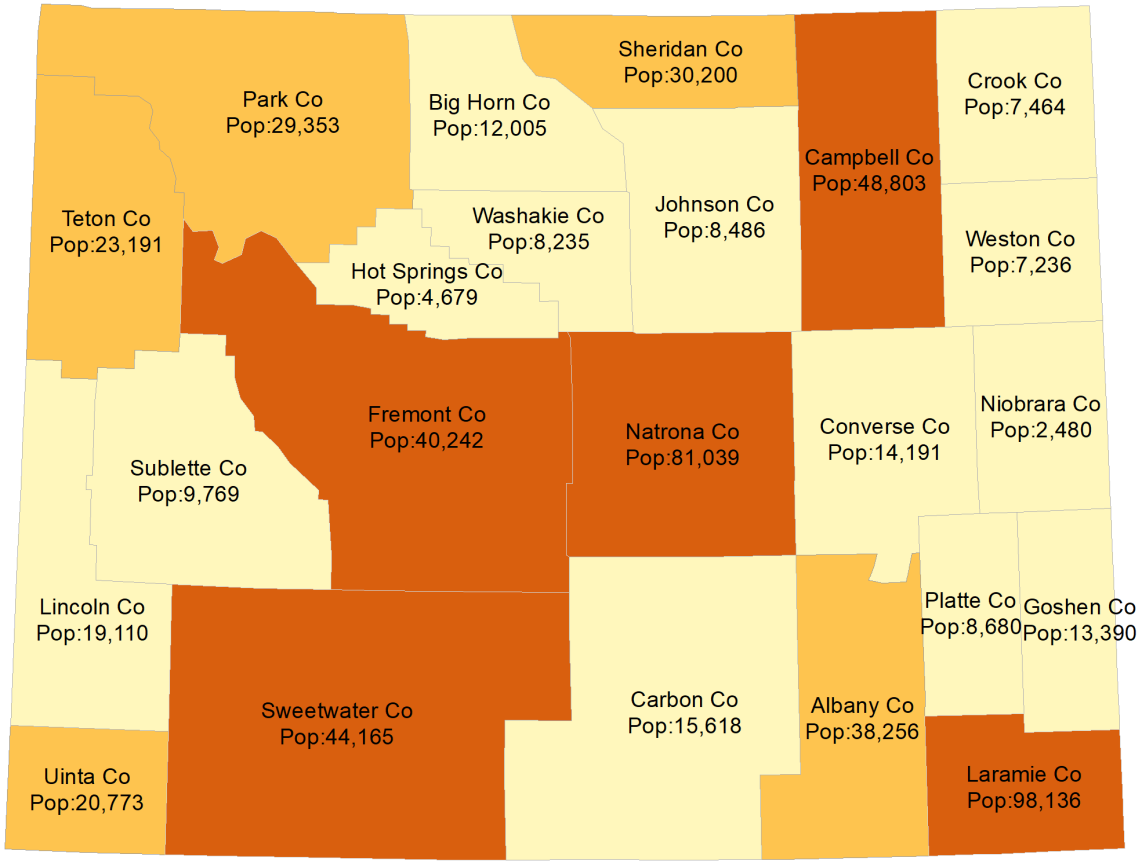
U.S. county population categories



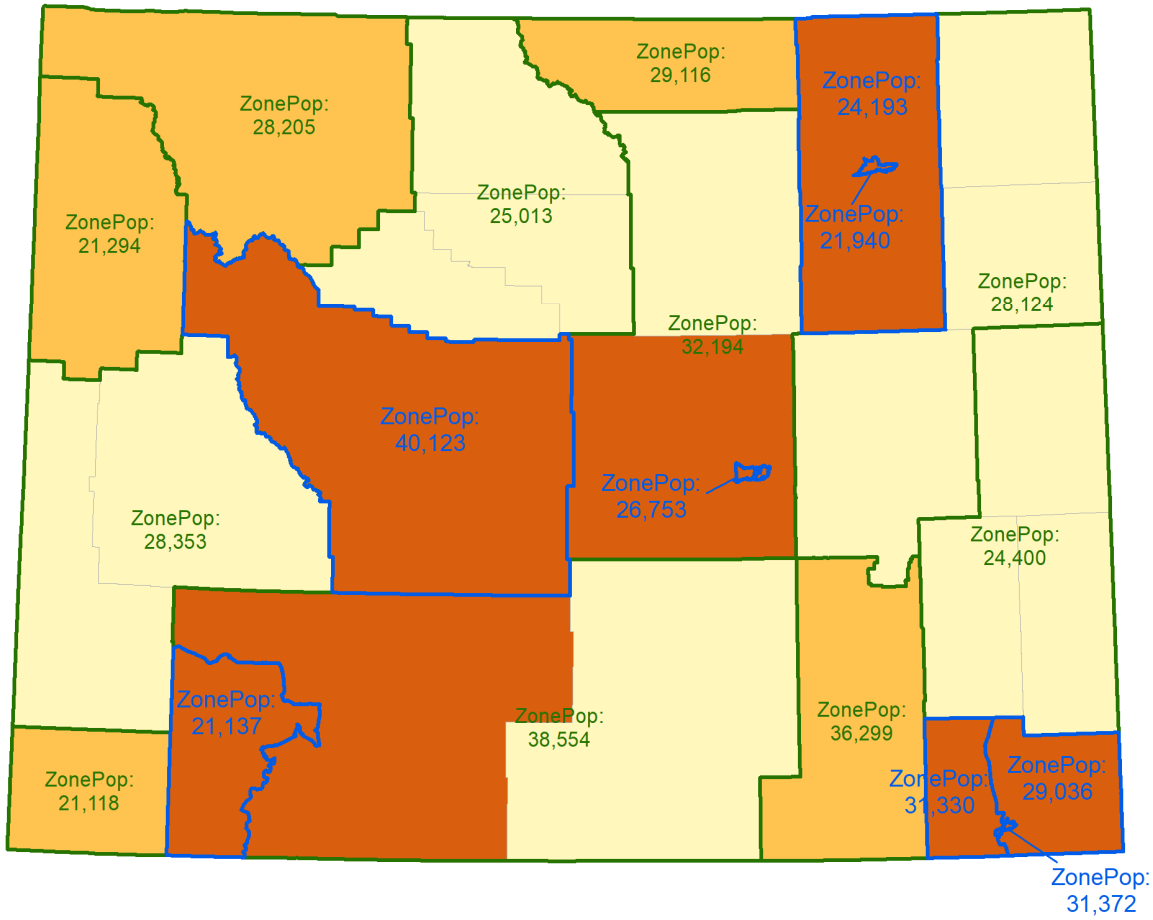
Initial zone construction tests - Vermont



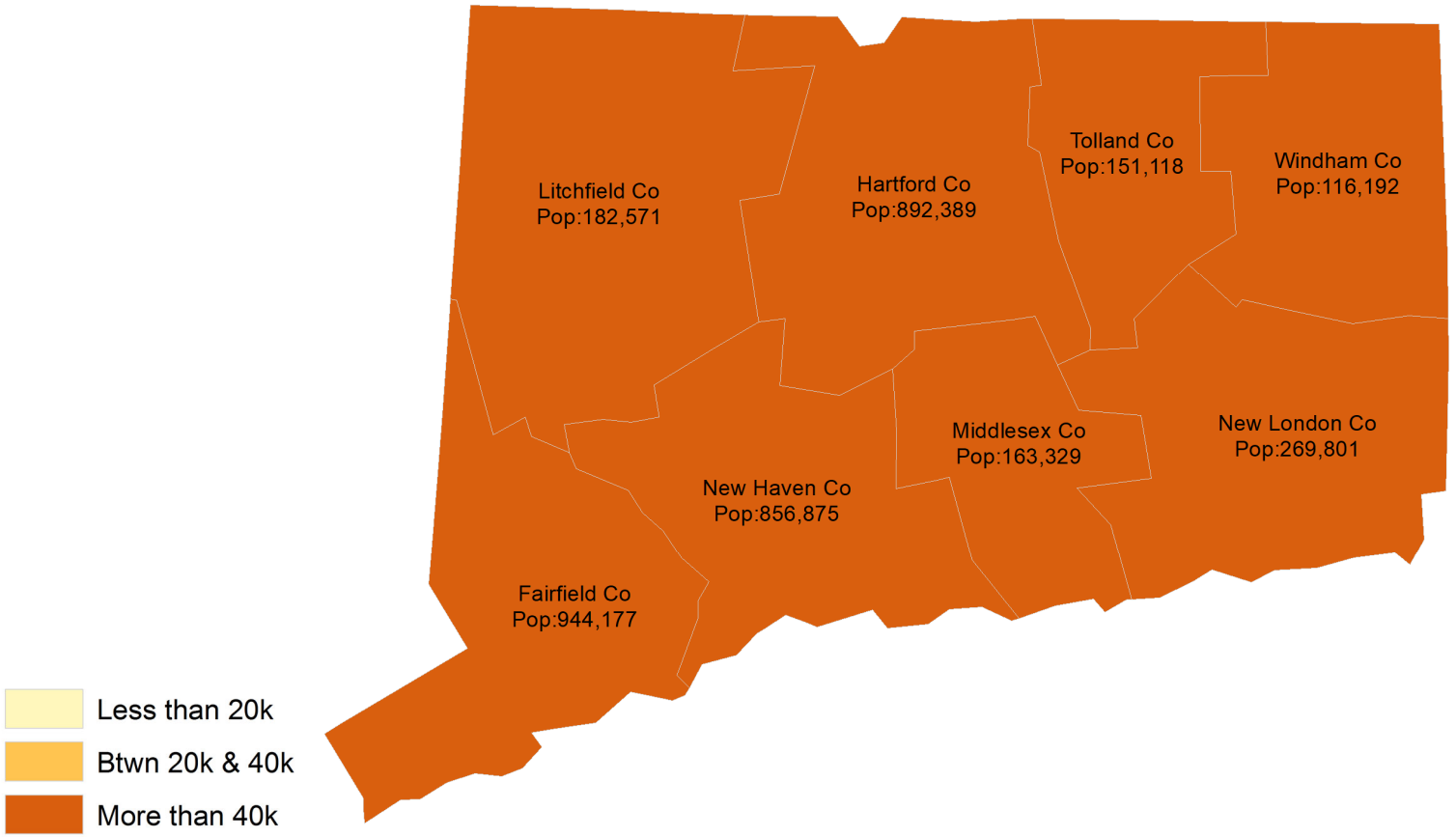
Initial zone construction tests - Wyoming



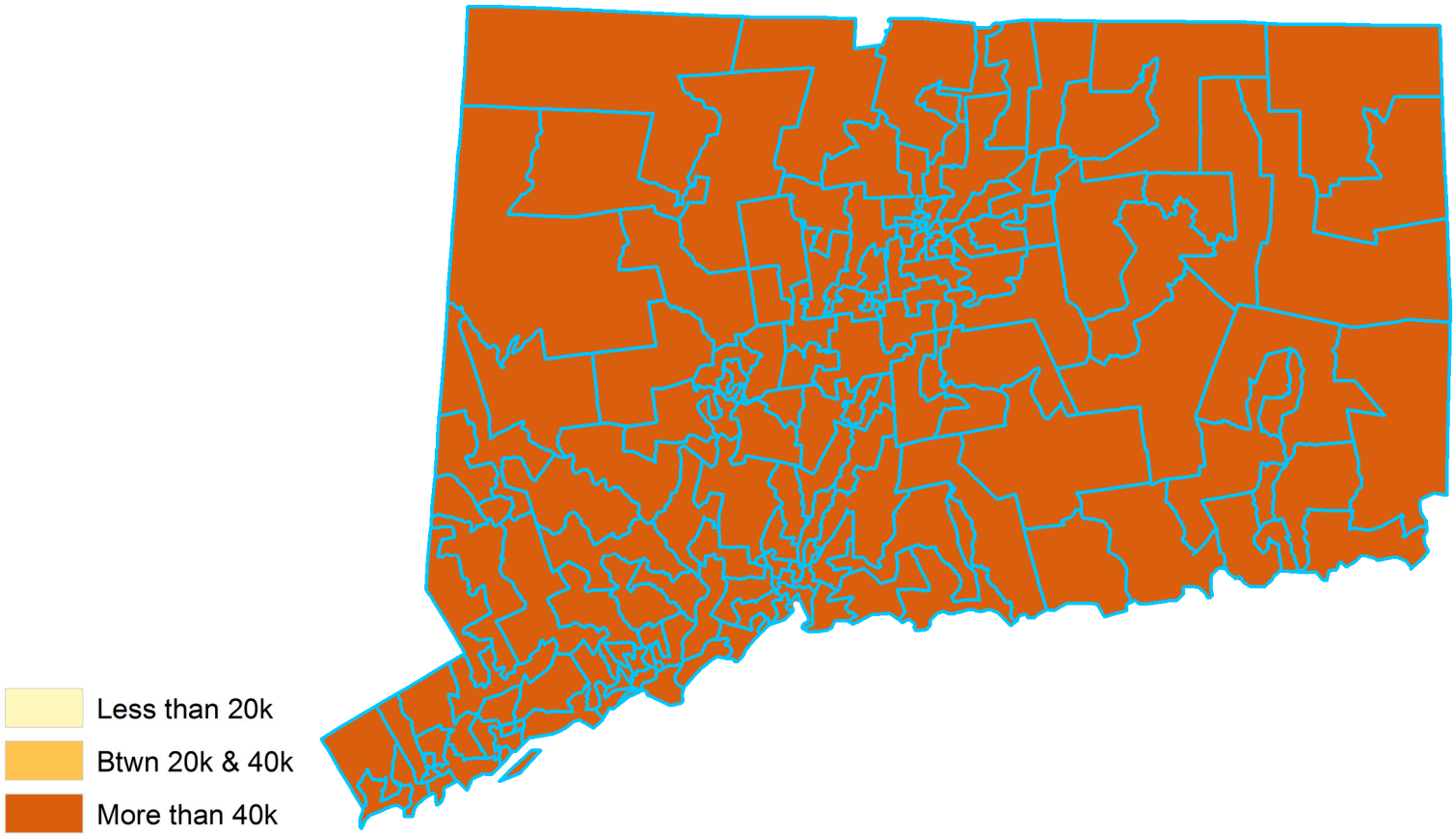
Initial zone construction tests - Wyoming



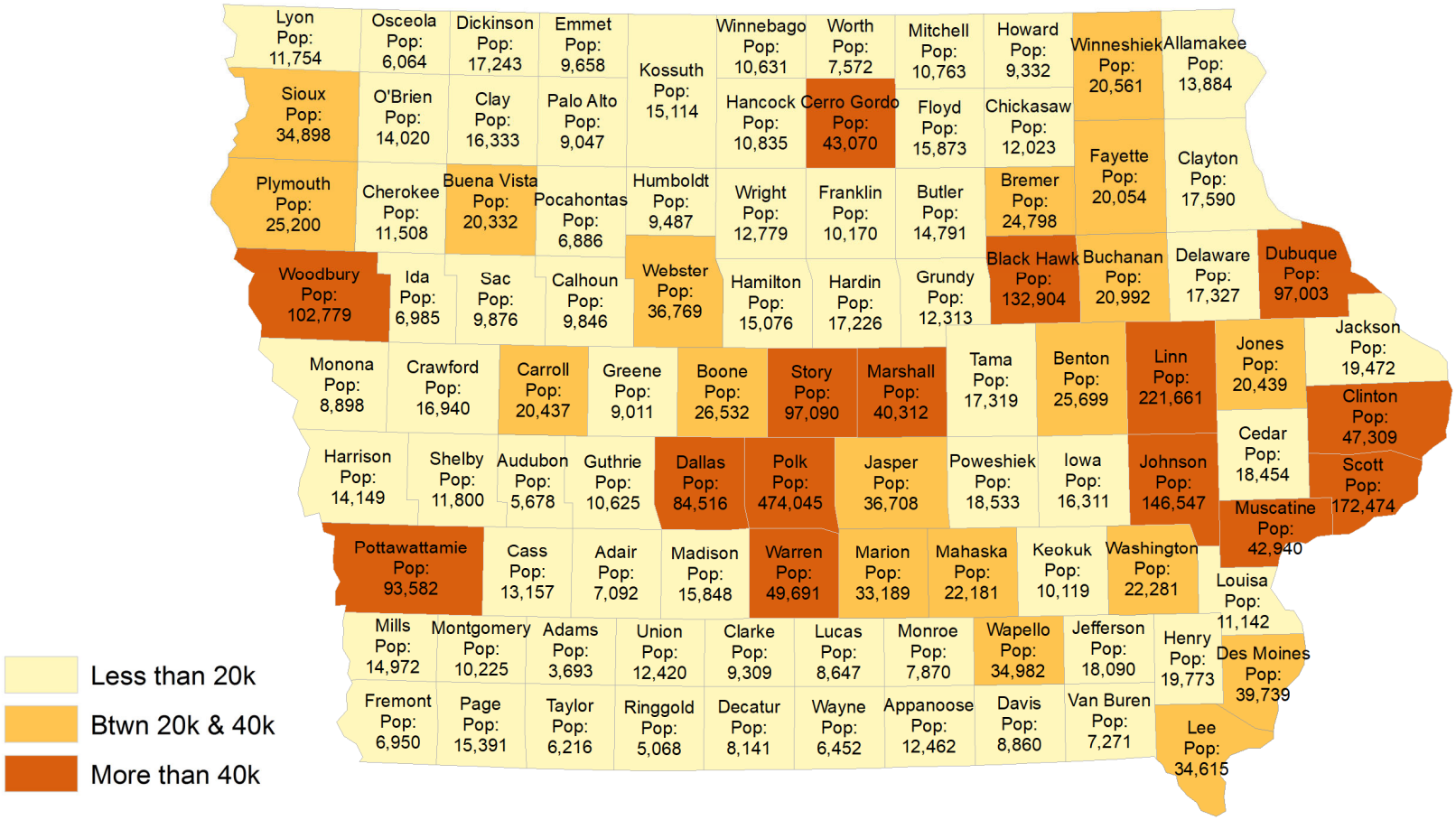
Initial zone construction tests - Connecticut



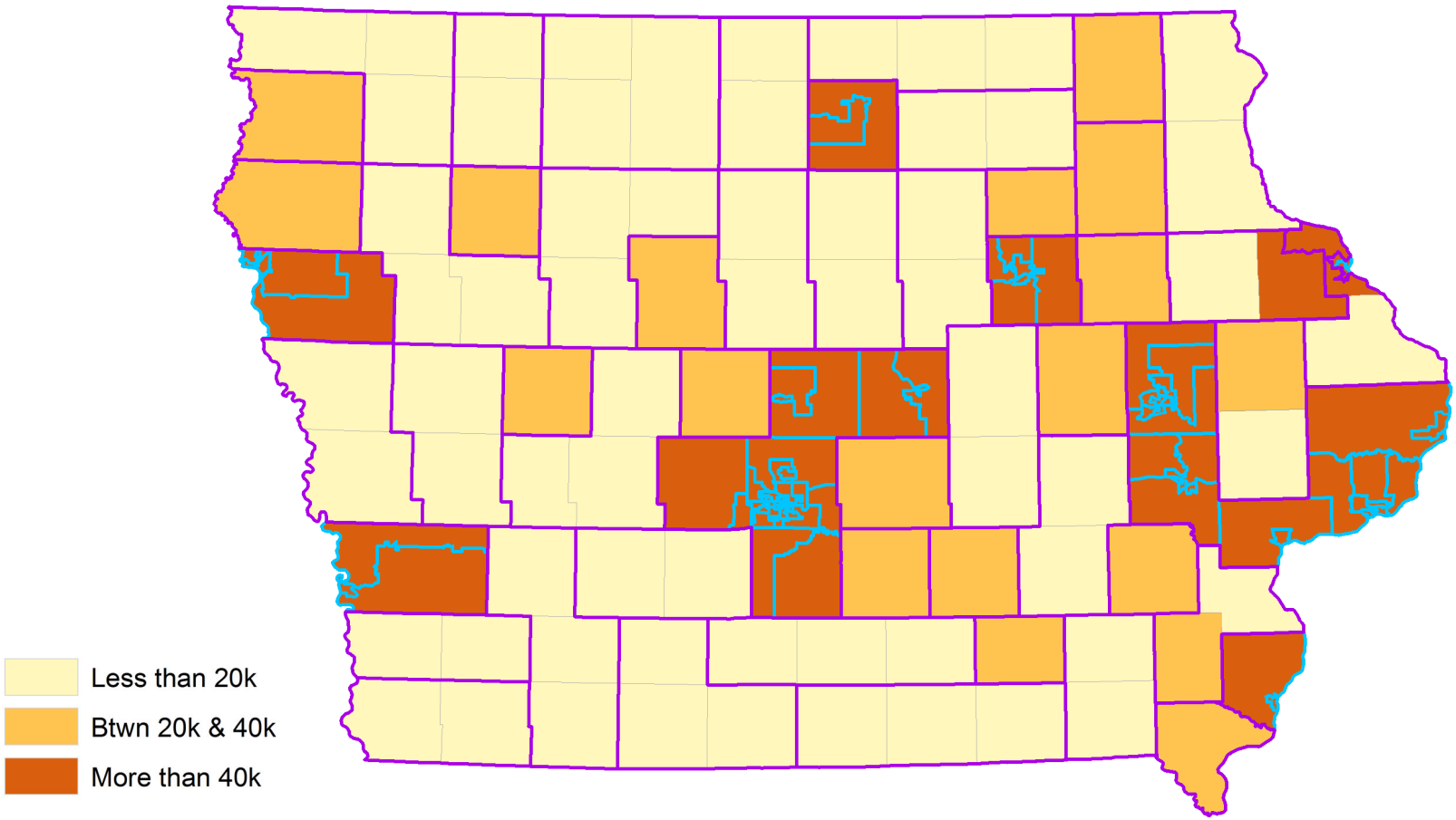
Initial zone construction tests - Connecticut



Initial zone construction tests - Iowa



Initial zone construction tests - Iowa



Target population size

- What should the target population be for our zones?
  - 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
- We can reduce suppression by aggregating more years of data
  - Case count estimates 1-year, 5-years, 10-years

Estimate population needed to have 16 cases based on 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

Minimum population of 20,000

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

Minimum population of 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

Agenda

Background

Goals and objectives

Initial activities

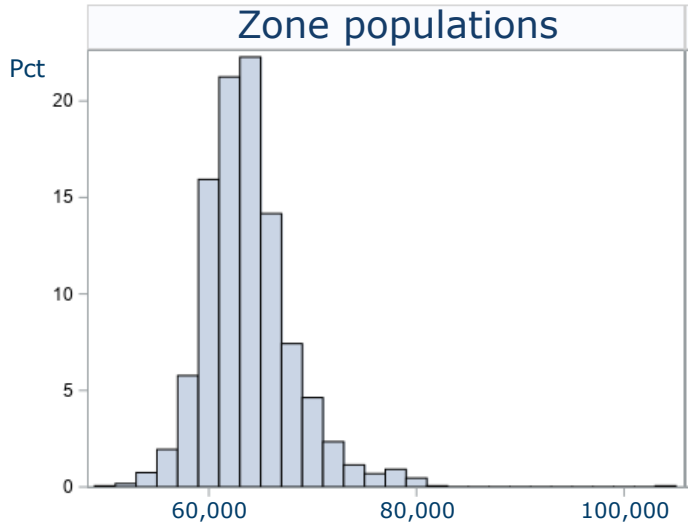
- Tool evaluation
- Initial zone construction tests
- Picking a target population size

California and Louisiana testing

- The differencing problem and a 2-step process
- Recent results

Simple approach – a single step

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



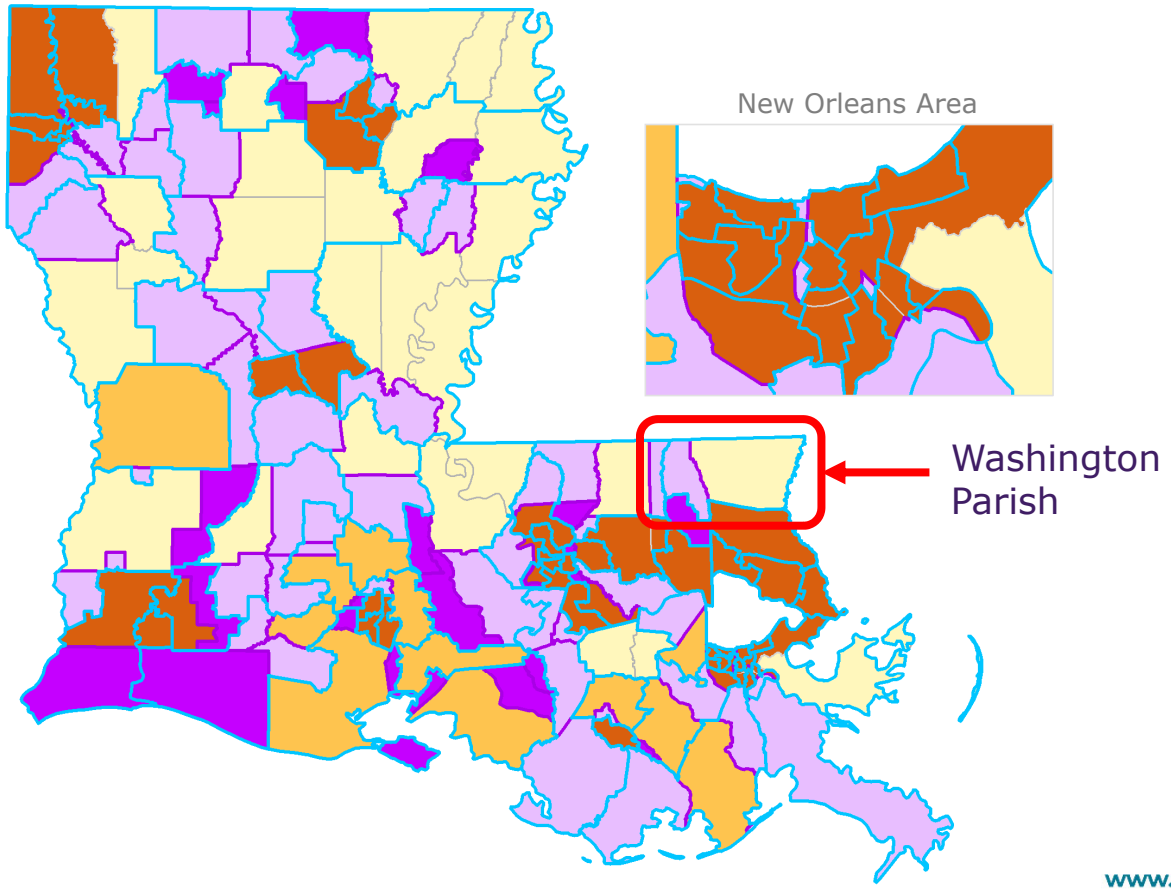
The differencing problem

- 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?

Potential differencing issues – Louisiana

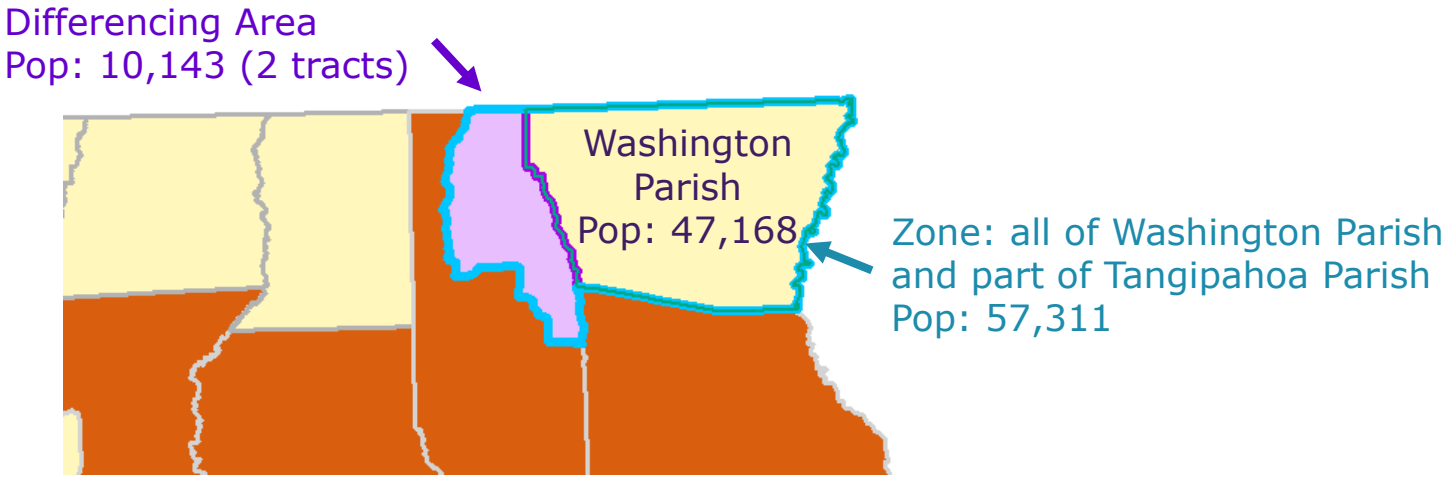


Areas with pop < 20,000

Single tract

2 or more tracts

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

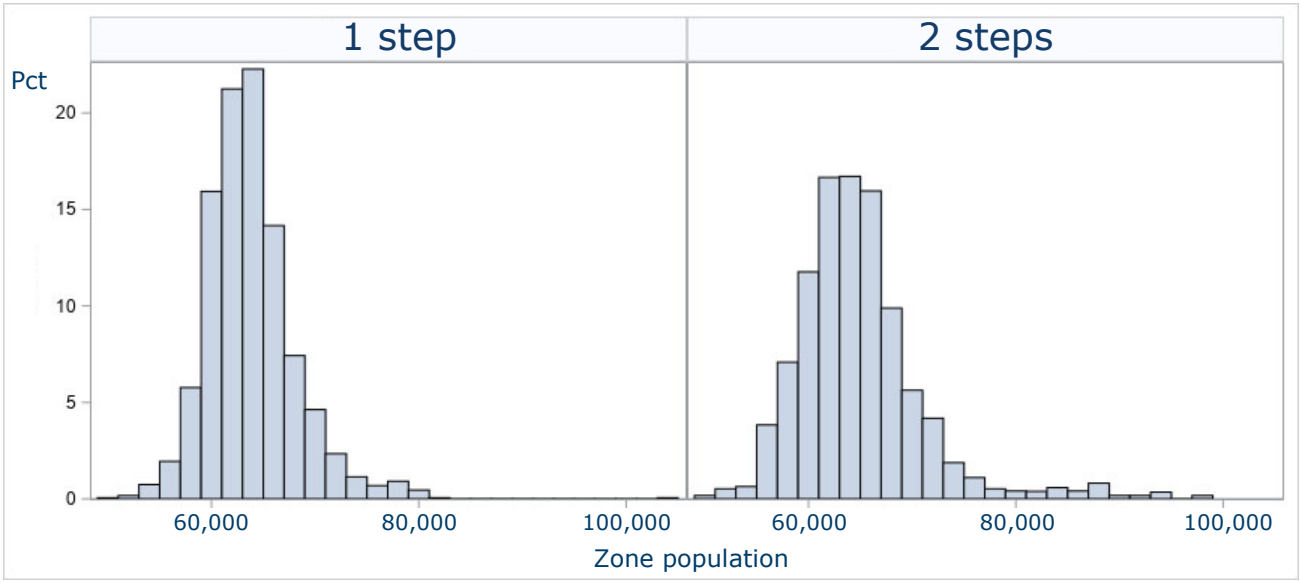
* Populations are real but incidence rates and case counts are made up

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

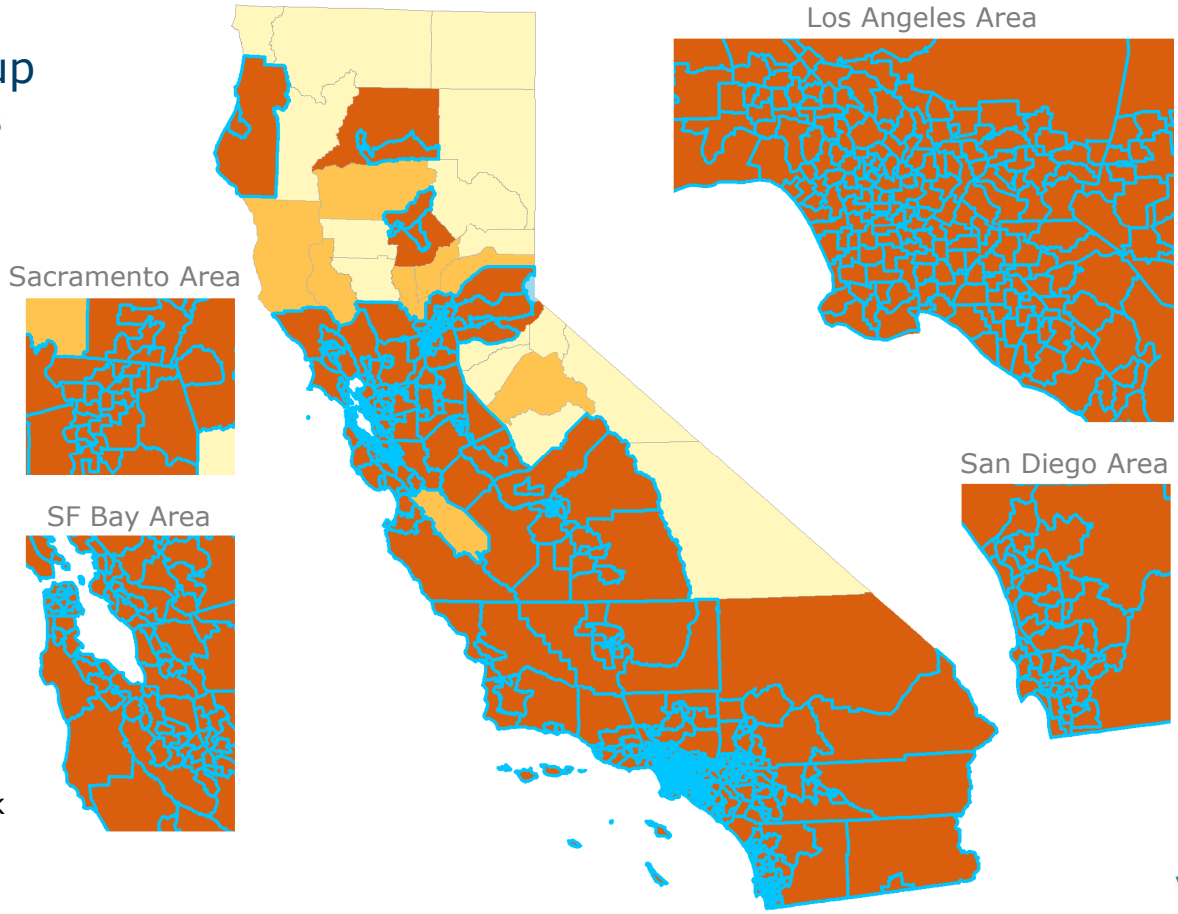
› The 2-step process results in zones with larger populations:



› An advantage of the larger populations is less suppression

Recent results – 2-step zones in California

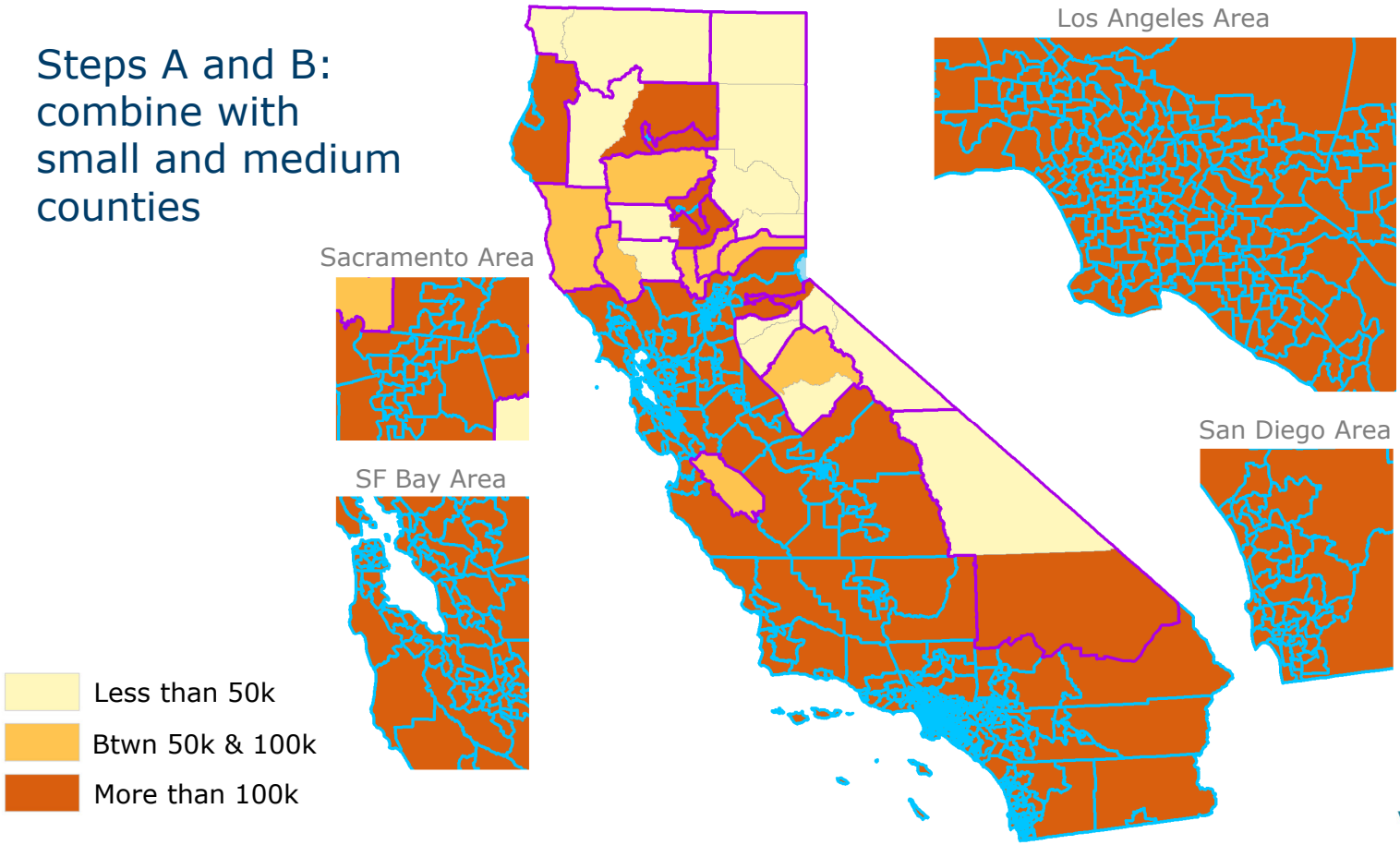
Step A: split up large counties



Less than 50k
Btwn 50k & 100k
More than 100k

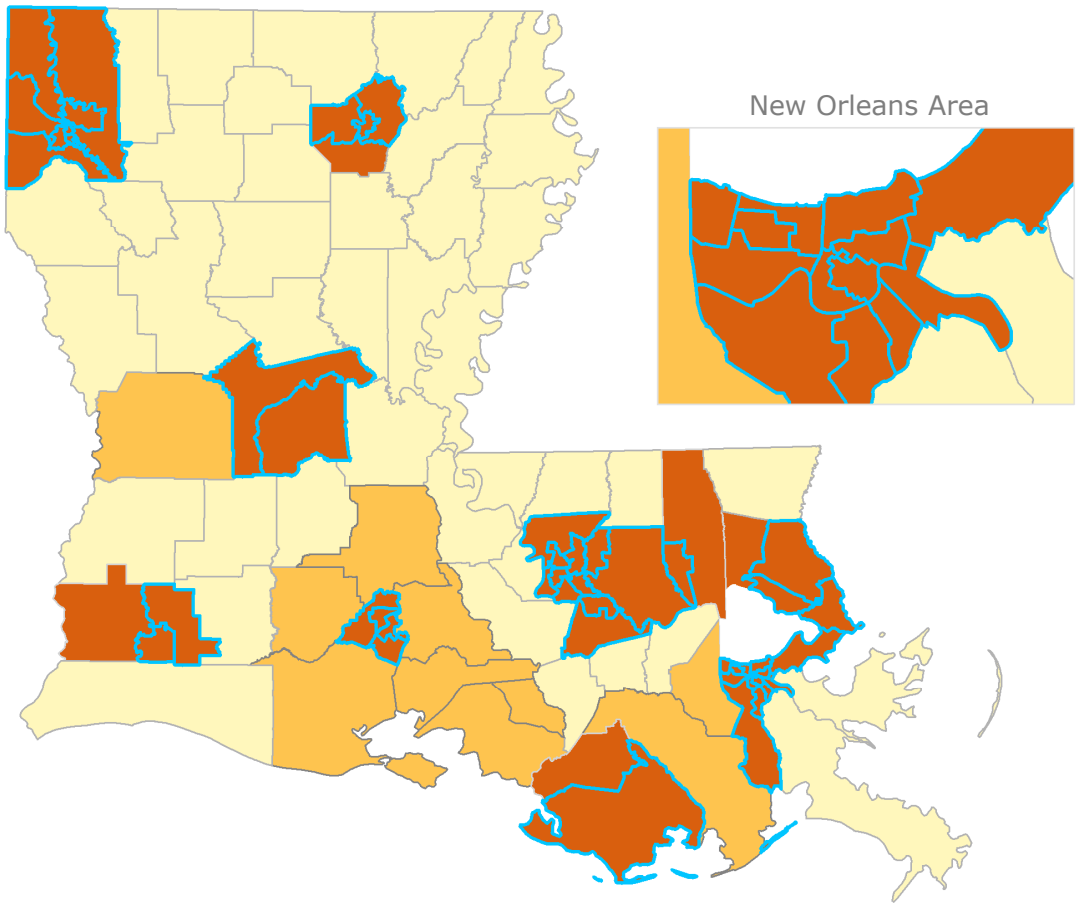
Recent results – 2-step zones in California

Steps A and B: combine with small and medium counties



Recent results – 2-step zones in Louisiana

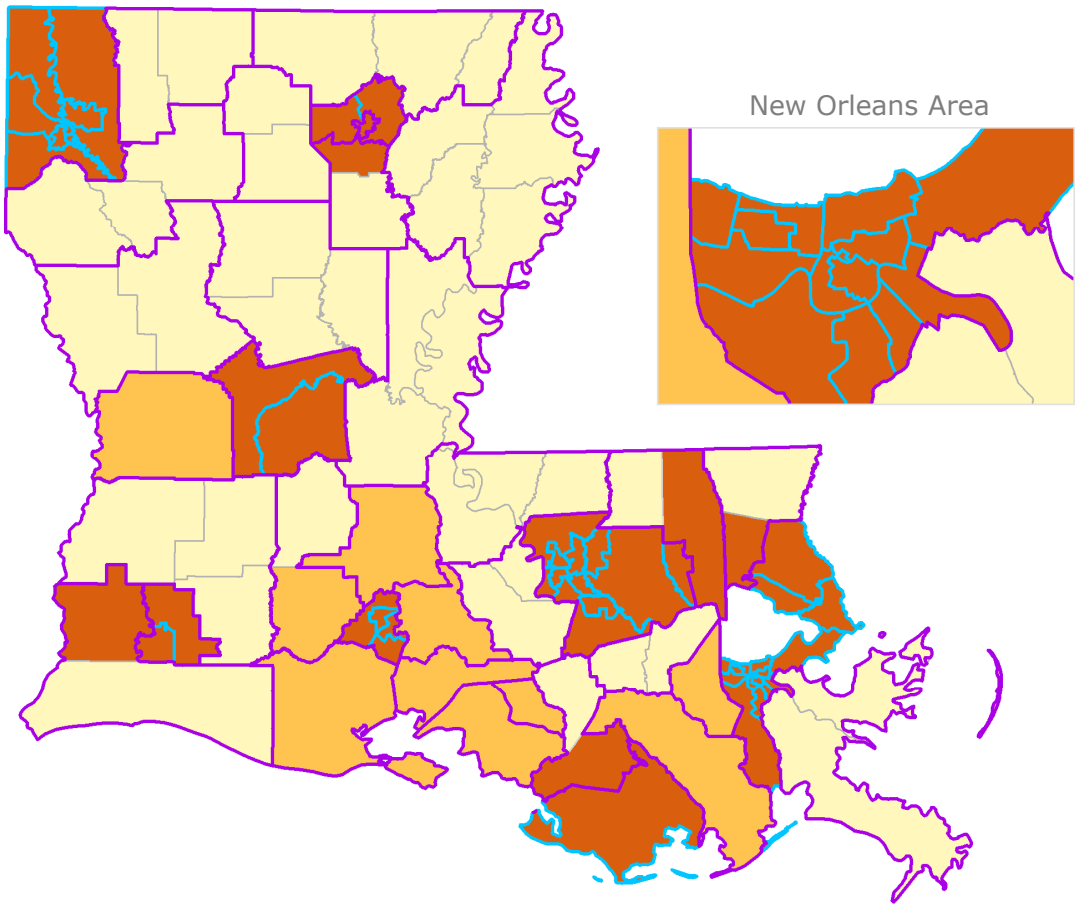
Step A: split up large parishes



Less than 50k
Btwn 50k & 100k
More than 100k

Recent results – 2-step zones in Louisiana

Steps A and B: combine with small and medium parishes

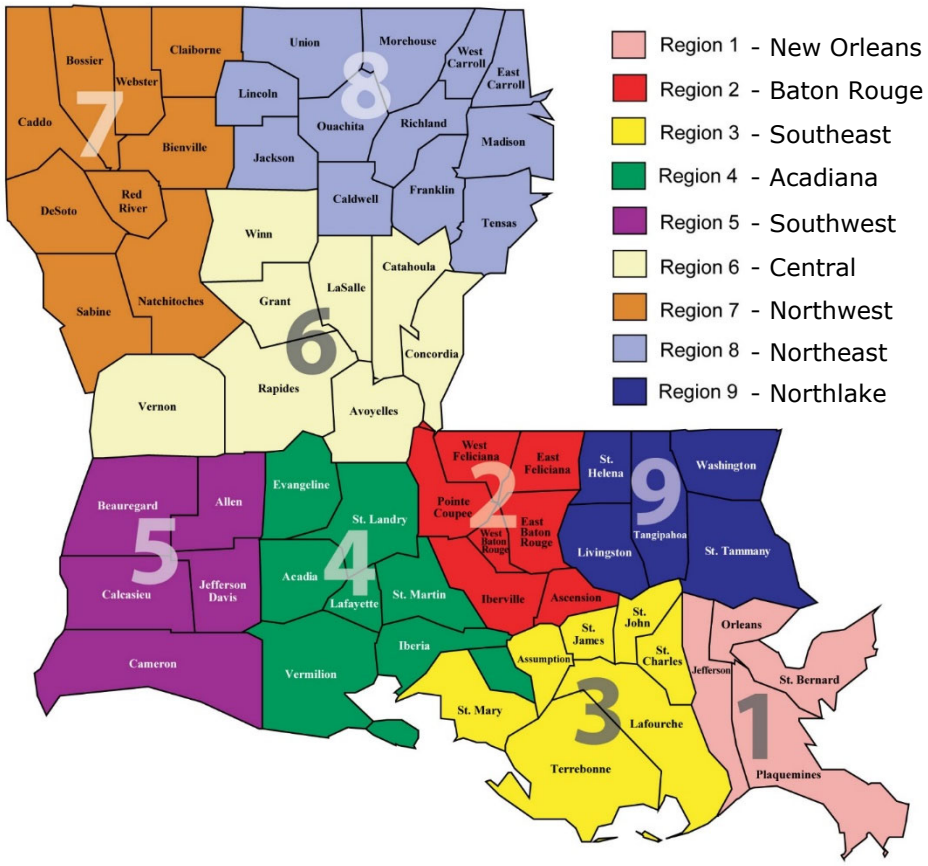


Less than 50k

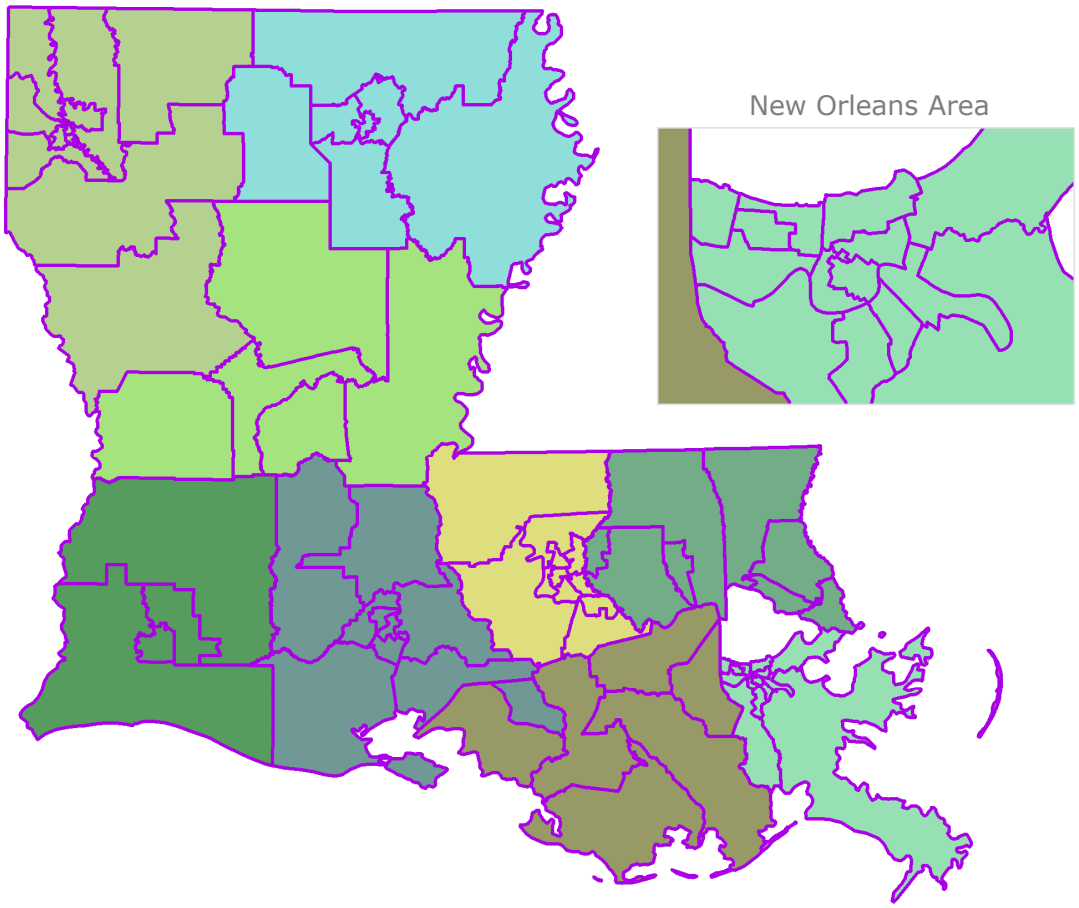
Btwn 50k & 100k

More than 100k

Louisiana Health Regions



Louisiana zones respect health region boundaries



Conclusions

- So far, we've agreed to:
  - Use a 2-step process
  - Set the minimum population to 50,000
  - Seek homogeneous zones based on
    - Urbanicity
    - % below poverty
    - % minority
  - Include a compactness objective
- State-specific options: any existing health regions to consider?

Next steps

- Still working with the California and Louisiana registries
  - Are these zones appropriate?
  - Are these zones useful for cancer reporting?
- Options for zone-level reporting
  - Website with rates by zone (tables and maps)
  - SEER*Stat database
  - Site, site x gender, site x gender x 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

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.
- 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.

Thank You

DavidStinchcomb@Westat.com