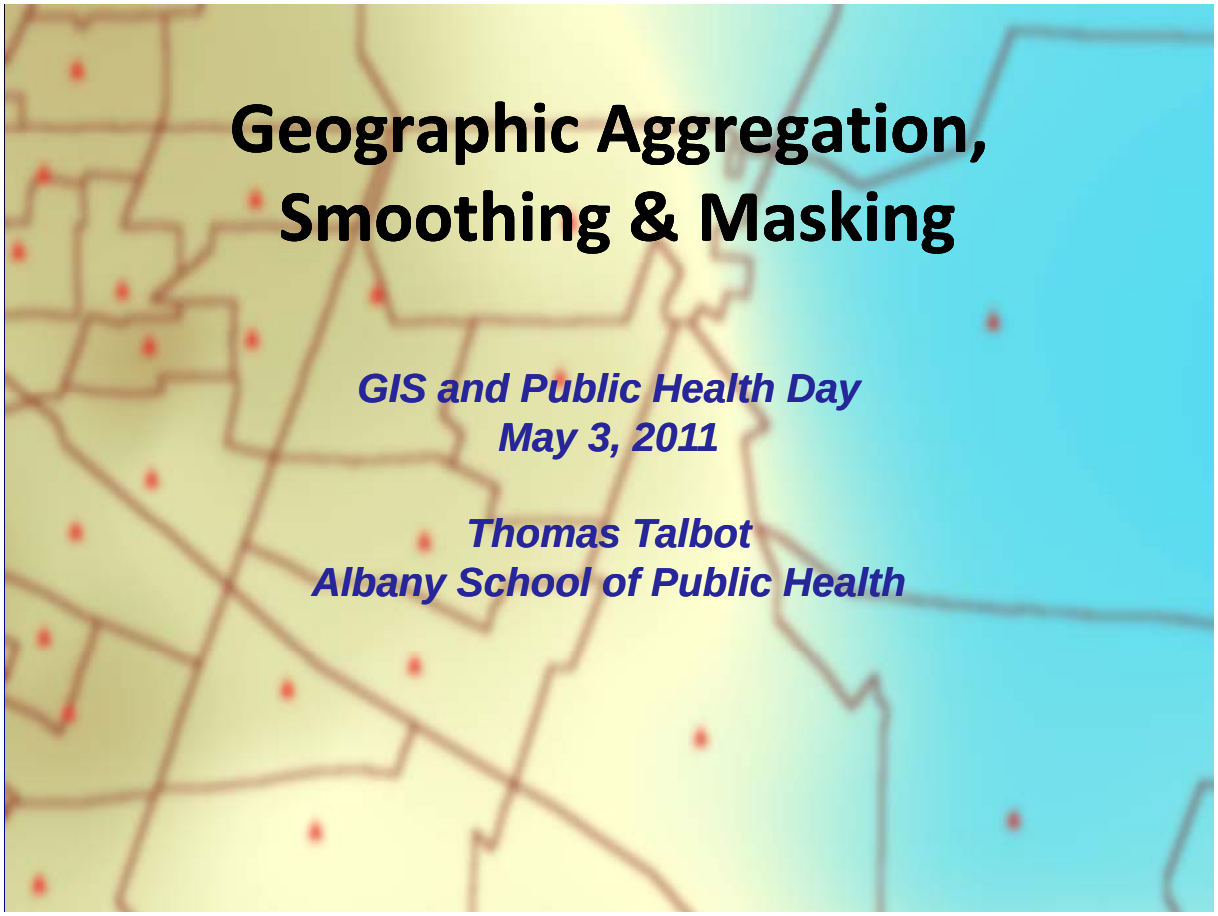


Geographic Aggregation, Smoothing & Masking

GIS and Public Health Day
May 3, 2011

Thomas Talbot
Albany School of Public Health

Community Health Mapping

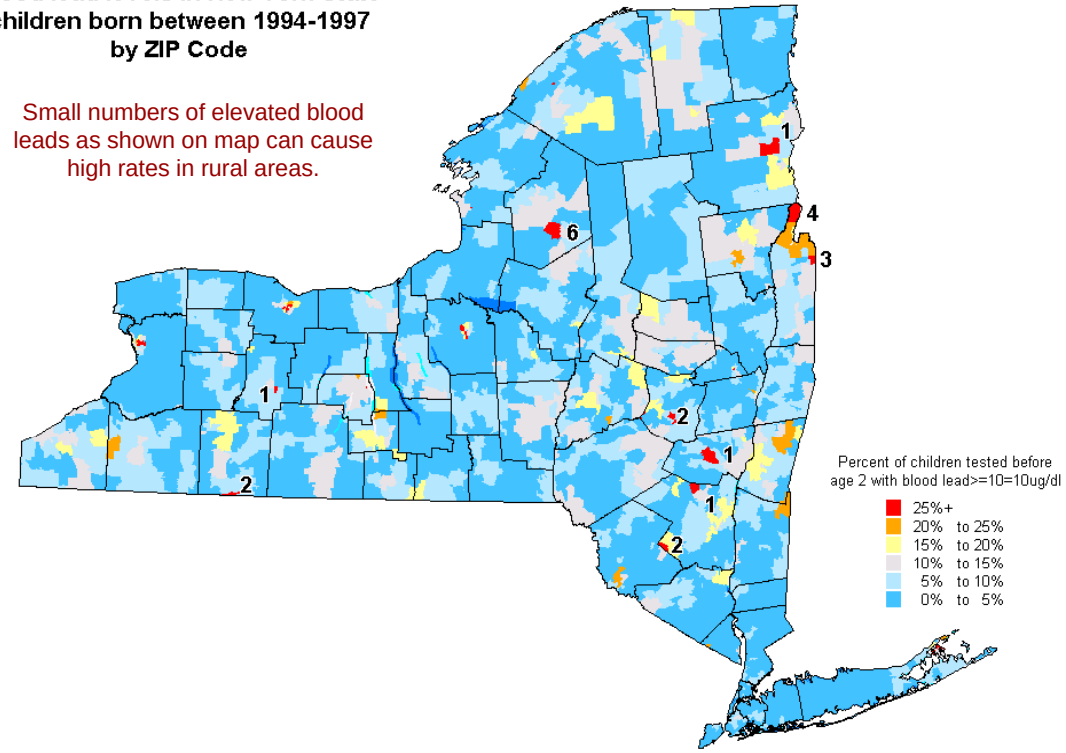
- Increasing demand to produce local community health maps.
- Risk of disclosure of confidential information when showing small area data.
- Rates of disease can be unreliable due to small numbers.

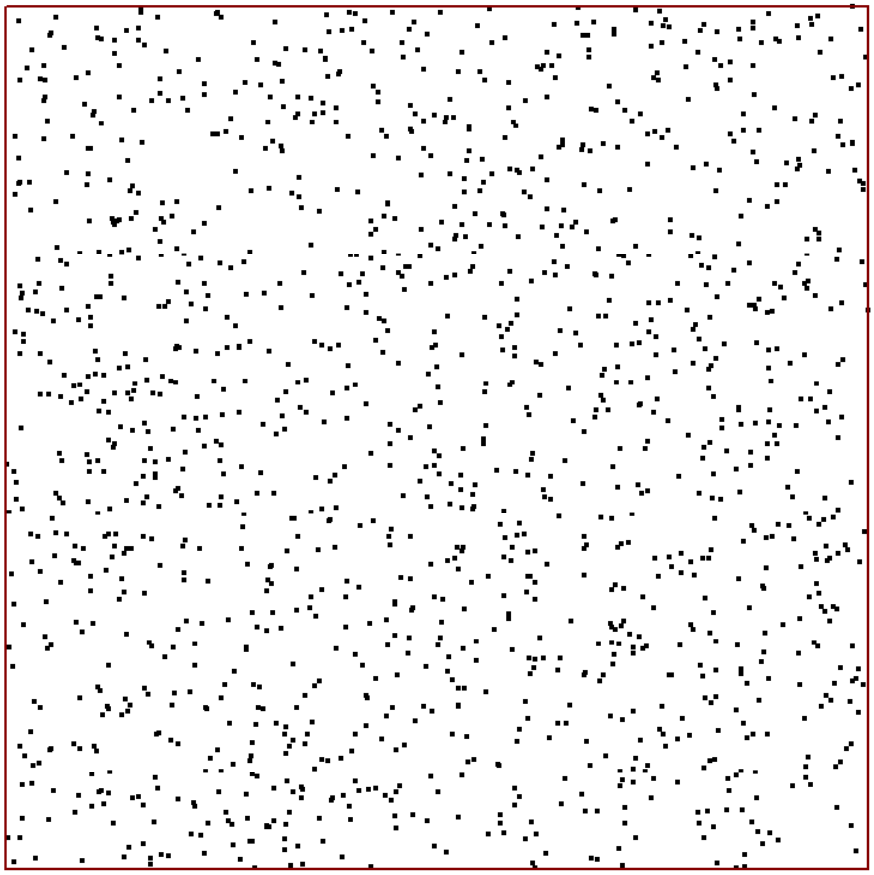
Disclosure of confidential information

Census Blocks

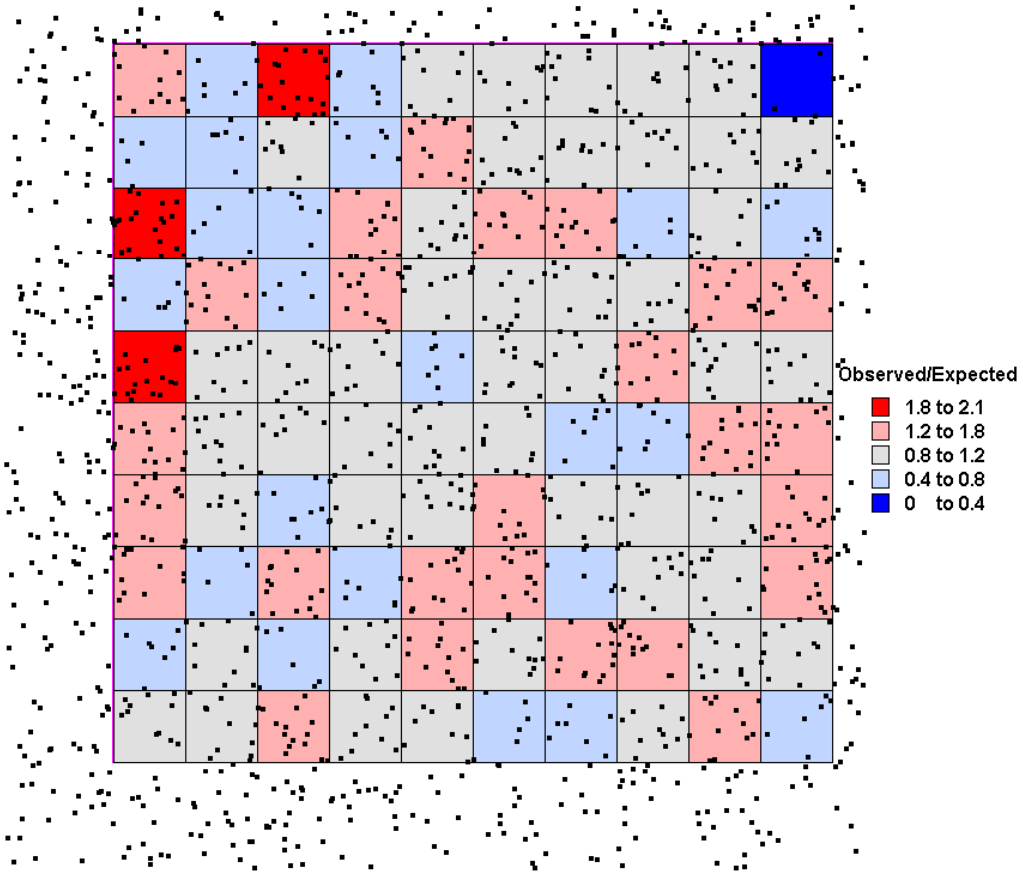
Geographic distribution of elevated blood lead levels in New York State children born between 1994-1997 by ZIP Code

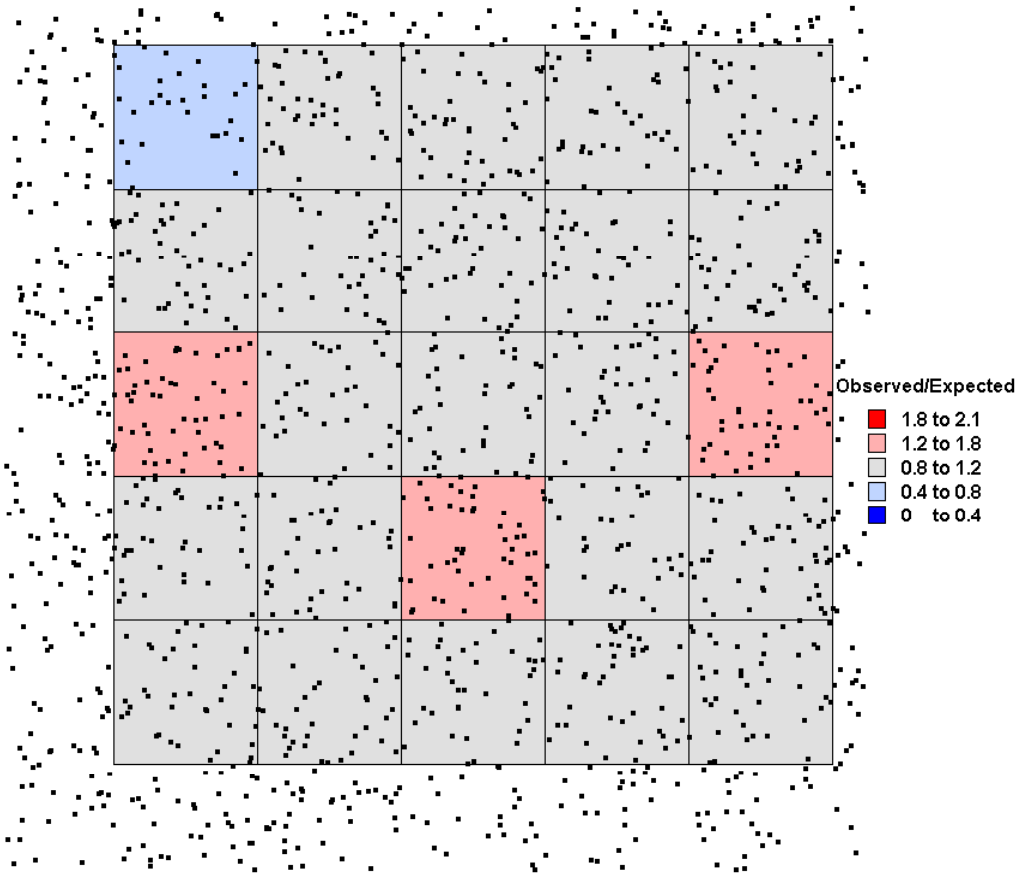
Small numbers of elevated blood leads as shown on map can cause high rates in rural areas.





Random Distribution of points



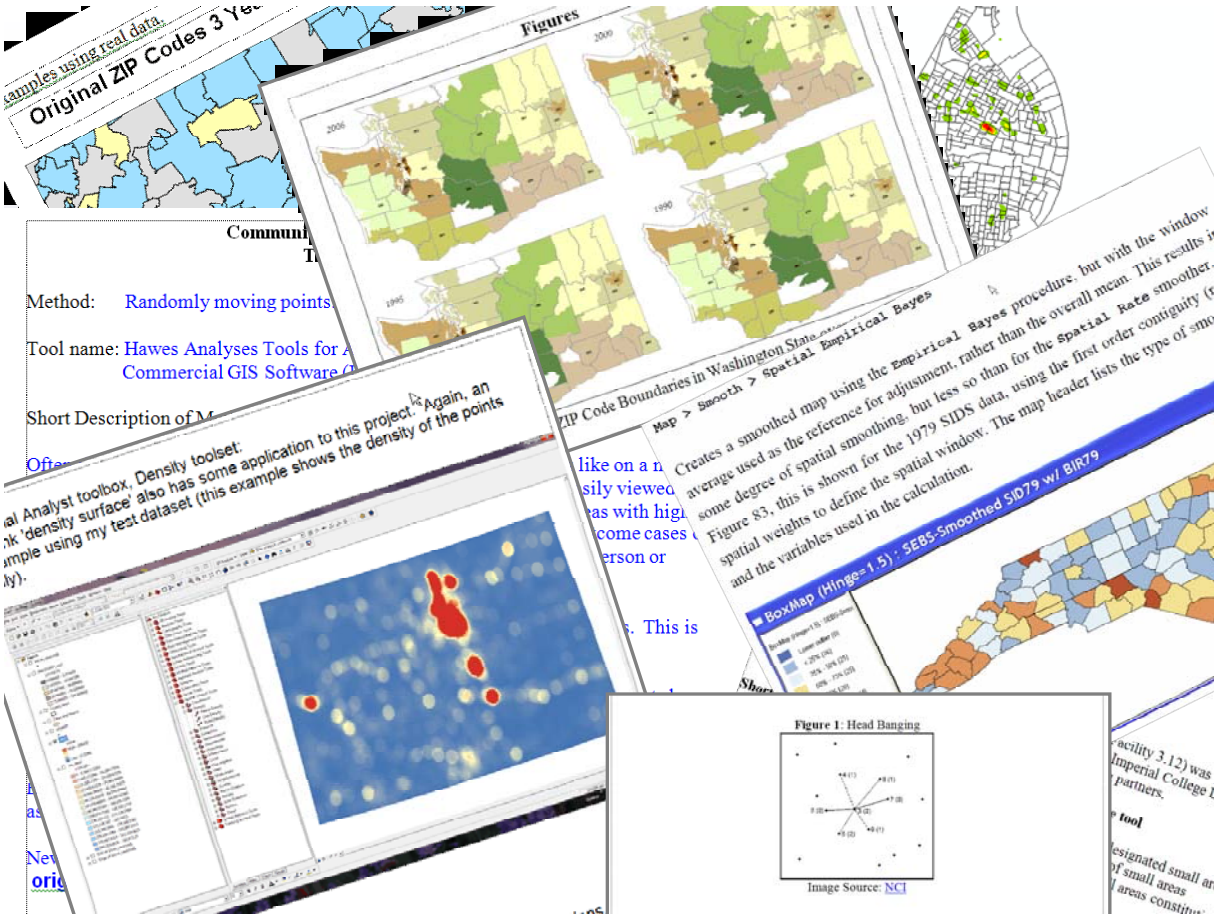


EPHT
Community Mapping Team Goals

- Identify and recommend methods and tools to maintain confidentiality and reduce unstable rates due to small numbers when mapping health data at the community level.
- Help build capacity in the state and national programs in the use of these tools.

Tasks

- Develop criteria for selecting tools & methods.
- Pilot test tools.
- Help organize webinars & training.



Criteria for Evaluating Tools

- Cost & Licencing
- Platform
- Software Requirements
- Ease of Use
- Open Source
- Training Support
- Confidentiality Issues
- Performance Issues

Working Definitions

•Spatial Aggregation: Merge spatial units to yield stable rate estimates and protect confidentiality

•Spatial Smoothing: Borrow data from neighboring areas to yield stable rate estimates and protect confidentiality

•Masking: Obscure specific data elements by replacing sensitive data with realistic but not real data.

Aggregating health data to existing geographic units

- The most commonly used method for producing maps of health outcomes.
  - Example EPHT County Maps
- Cases and population are assigned to a geographic area.
- Thematic maps of disease rates can then be produced using a number of software tools.
  - SAS, ArcGIS, MapInfo

Examples

- State
- County
- Town
- Census Tract
- ZIP Code
- Health Service Areas
- School Districts
- EMS Regions
- Health Department Districts
- Metropolitan Areas (MSA)

Problems using Pre-existing Regions

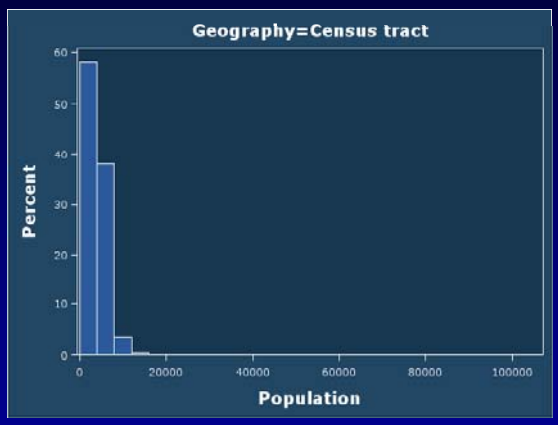
Unequal populations

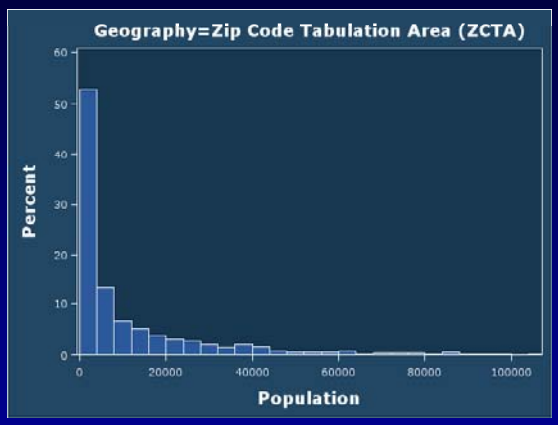
Populations are too large. Difficult to see variations in rates between local communities.

or

Populations are too small so data is suppressed or rates are unstable due to chance.

Population sizes vary.



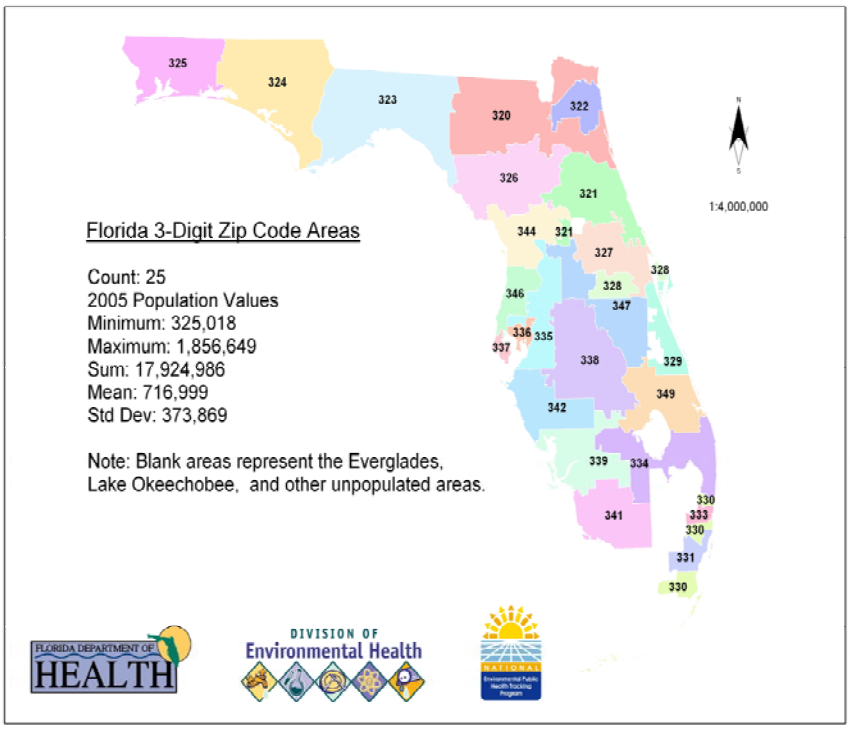


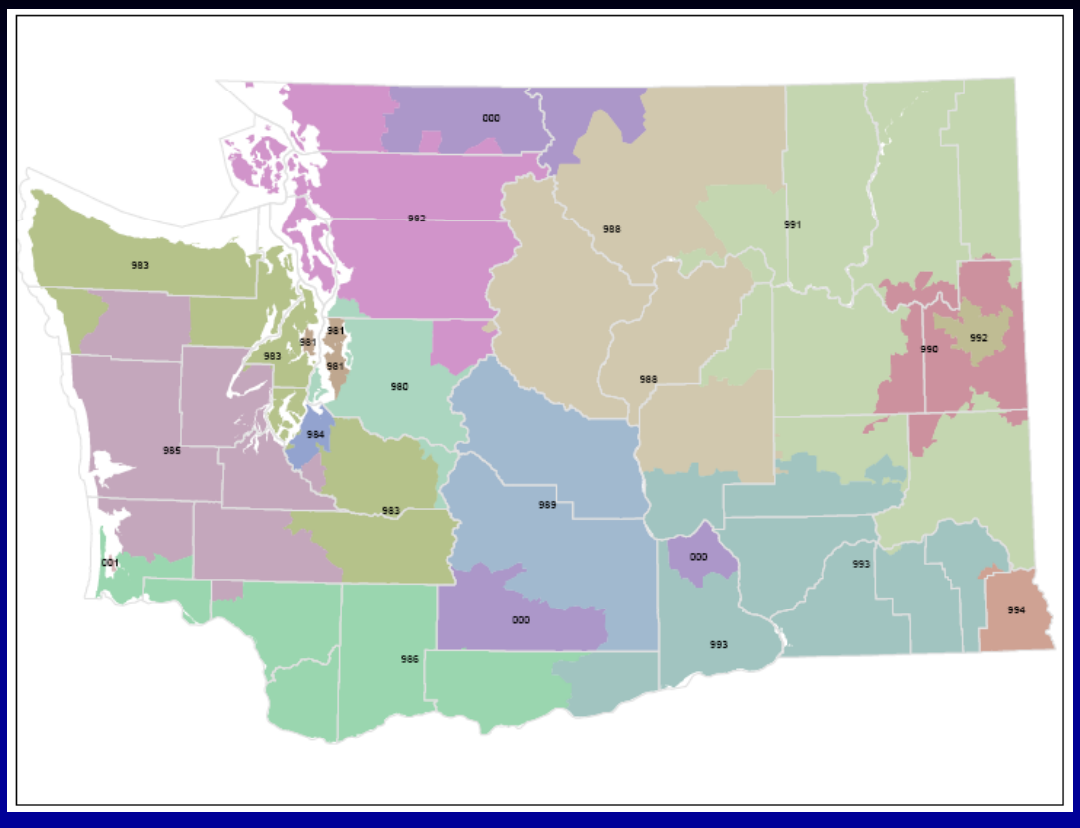
Number of Tracts	4,907
Population Range	1 - 24,523
Population Median	3,624

Number of ZIP Codes	1,606
Population Range	1 - 100,995
Population Median	3,936

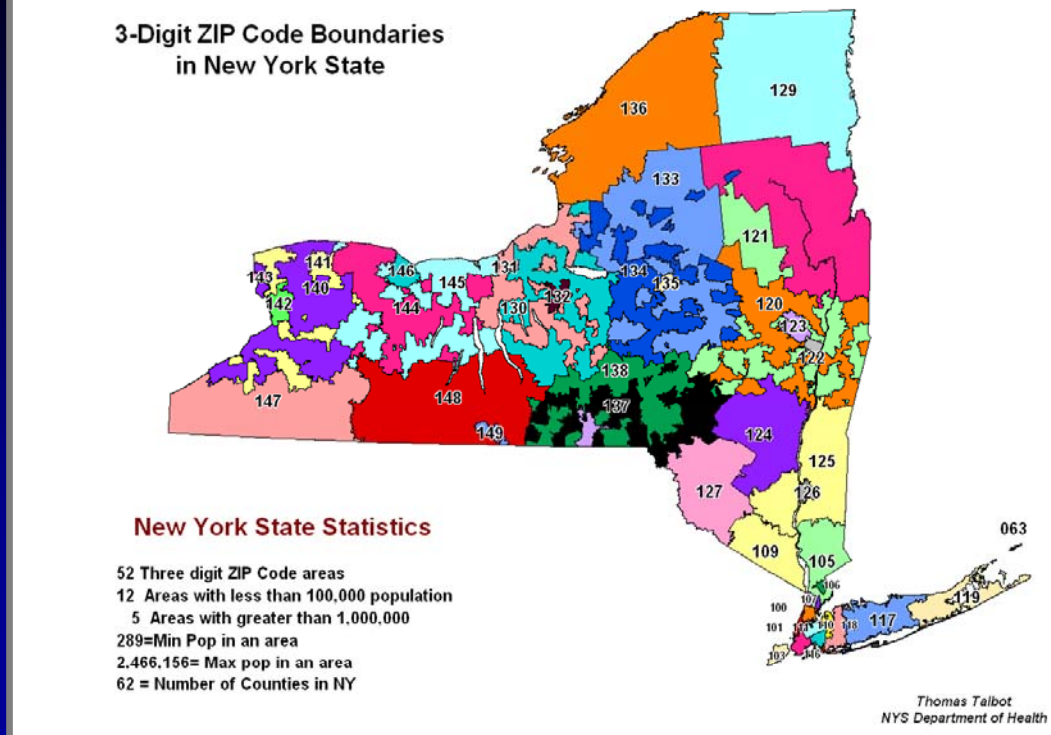
2000 US Census, New York State

Health Departments Looked at 3-Digit ZIP Code Merging





Washington State EPHT



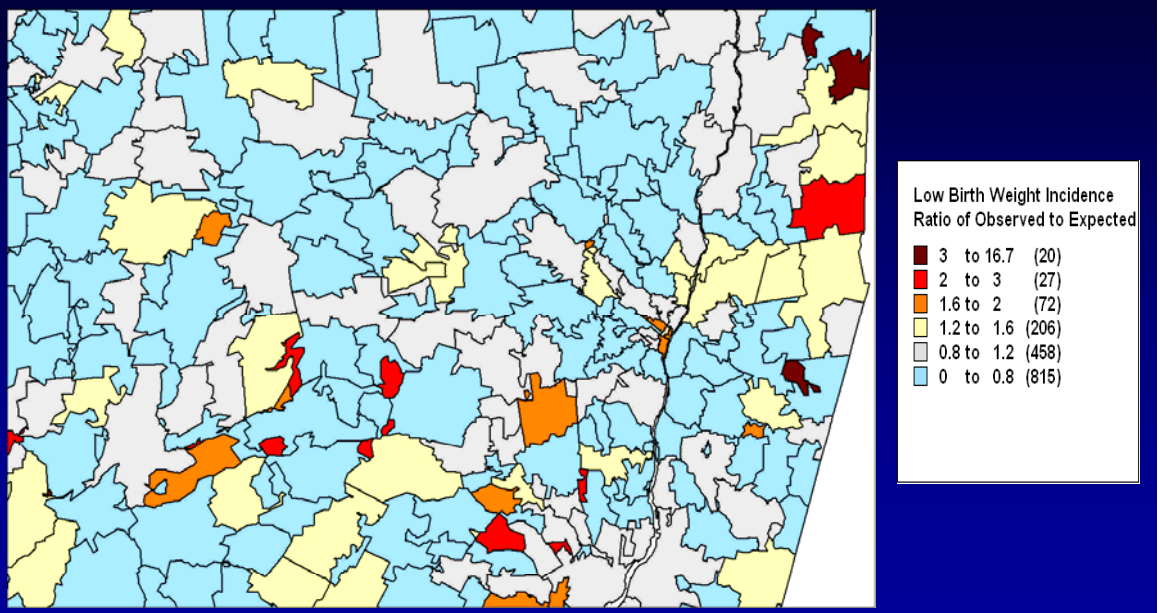
Need for an Aggregation Tool

- Merge small areas with neighboring areas to provide more stable rates of disease and/or protect confidentiality.
  - Aggregation can be done manually.
  - Existing automated tools were difficult to use or did not fulfill requirements.

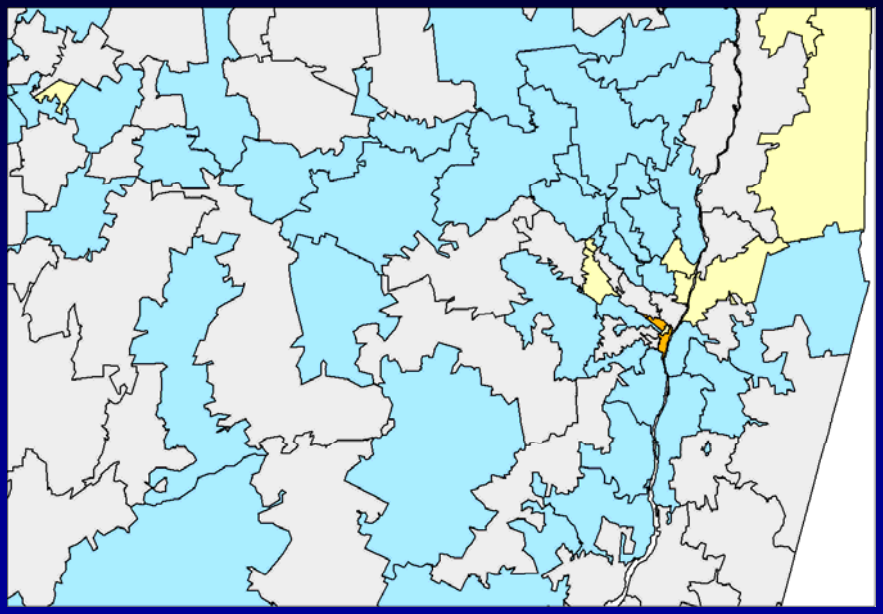
NYSDOH Tool Requirements

- Aggregate small areas into larger ones.
- User decides how much aggregation is needed.
  - Based on cases and/or underlying population
  - Example 250 births and at least 3 low birth weight births
- Works with various levels of geography.
  - Census blocks, tracts, towns, ZIP codes etc.
  - Can nest one level of geography in another
    - Example: Census tracts are aggregated. Aggregated areas do not cross county boundaries
- Uses open source free software (R).
- Outputs results for use in mapping programs.

Original ZIP Codes
3 Years Low Birth Weight Incidence Ratios



Aggregated to 250 Births per ZIP Code Group



Aggregation Tool

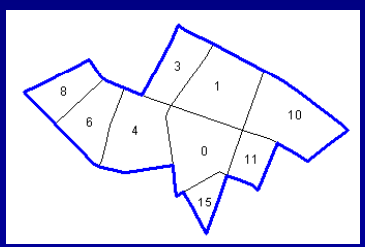
Original Block Data †

Block	Cases
122300/2004	10
122300/2005	11
014500/3005	3
014500/3007	4
014500/3008	0
014500/3009	1
014500/3010	15
103202/2001	8
103202/2002	6

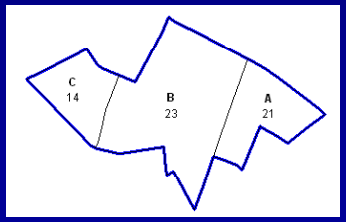
SAS or R Tool

Regions

Block	Cases	Region
122300/2004	10	A
122300/2005	11	A
014500/3005	3	B
014500/3007	4	B
014500/3008	0	B
014500/3009	1	B
014500/3010	15	B
103202/2001	8	C
103202/2002	6	C

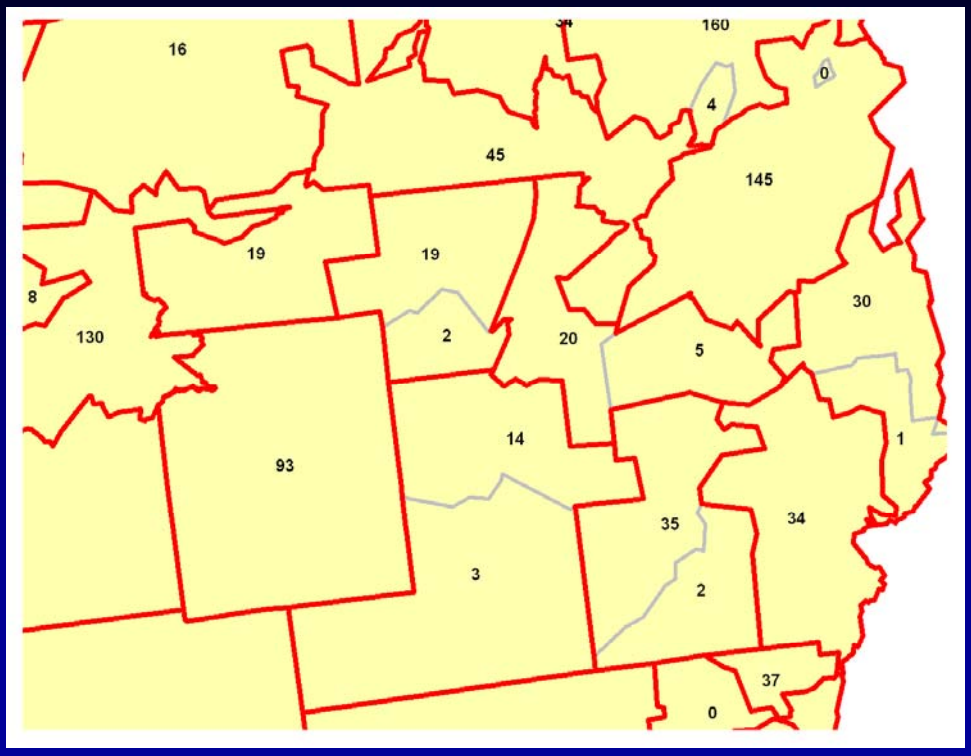


Cases	Region
21	A
23	B
14	C

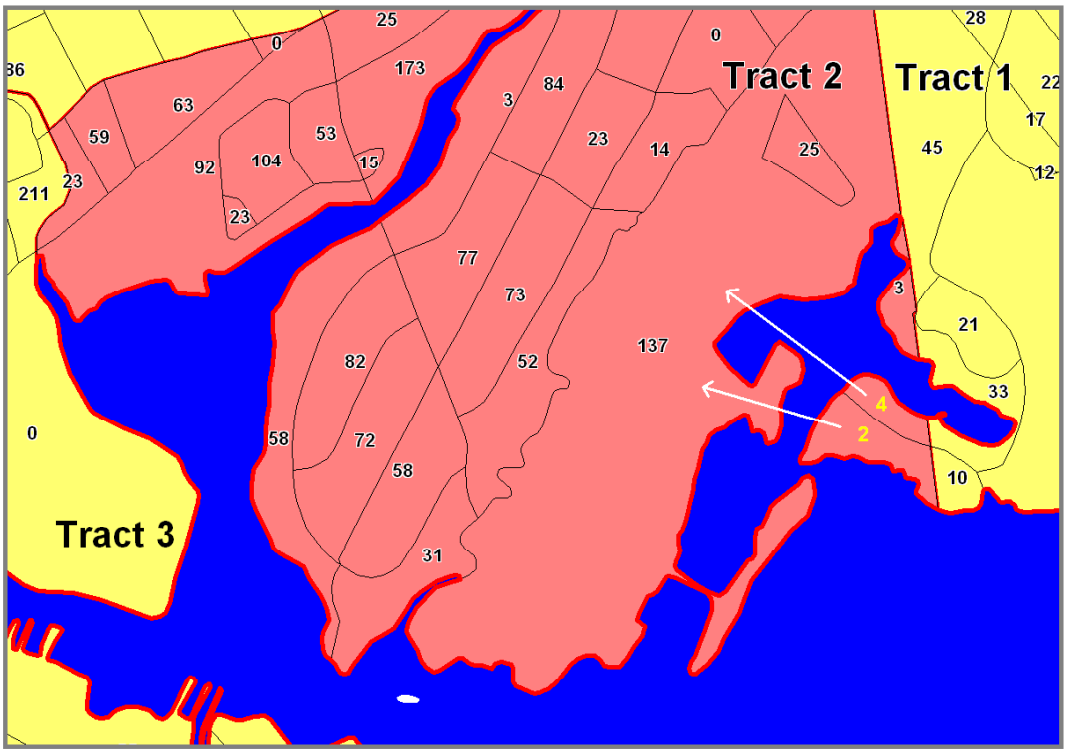


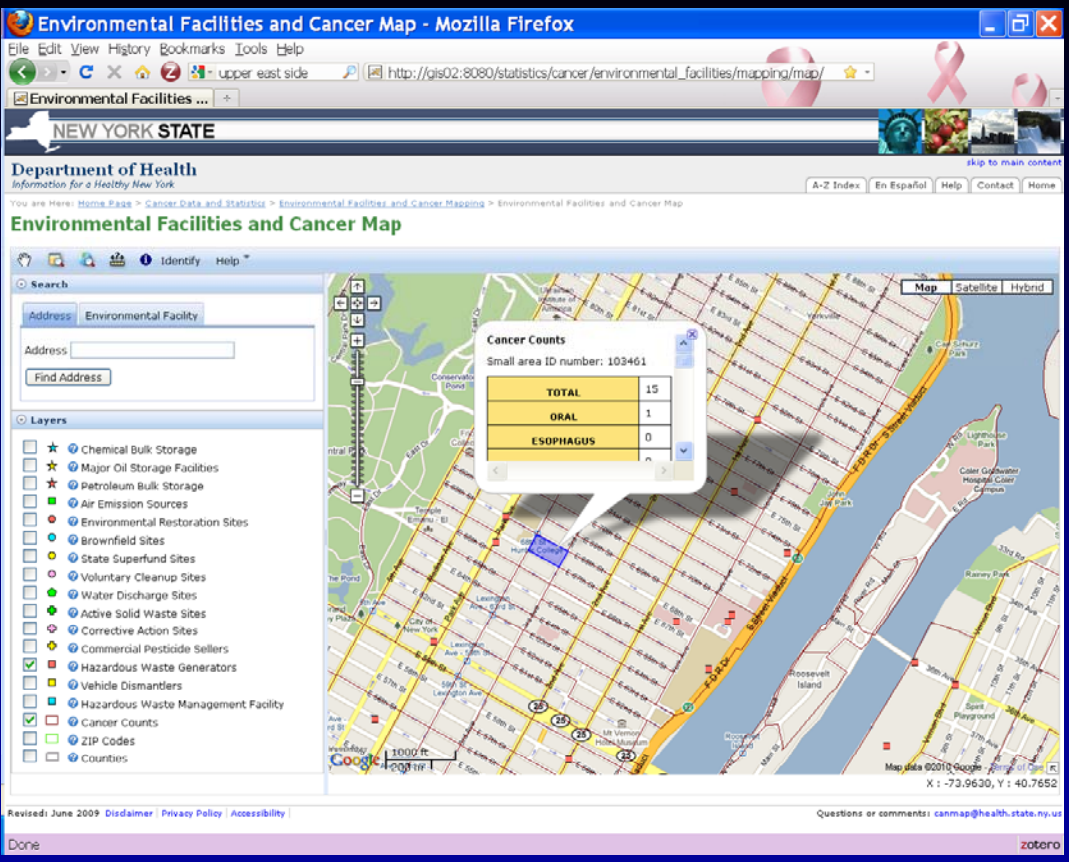
† Simulated data

Merging ZIP Codes to avoid Data Suppression & Provide Stable Rates



Merging blocks with noncontigous blocks in same tract.





Performance Measures

- Compactness
- Similar population sizes.
- Number of aggregated areas.
- Aggregated zones are contained within larger areas.
- Tool can handle large numbers of polygons
- Speed

New York State Descriptive Statistics

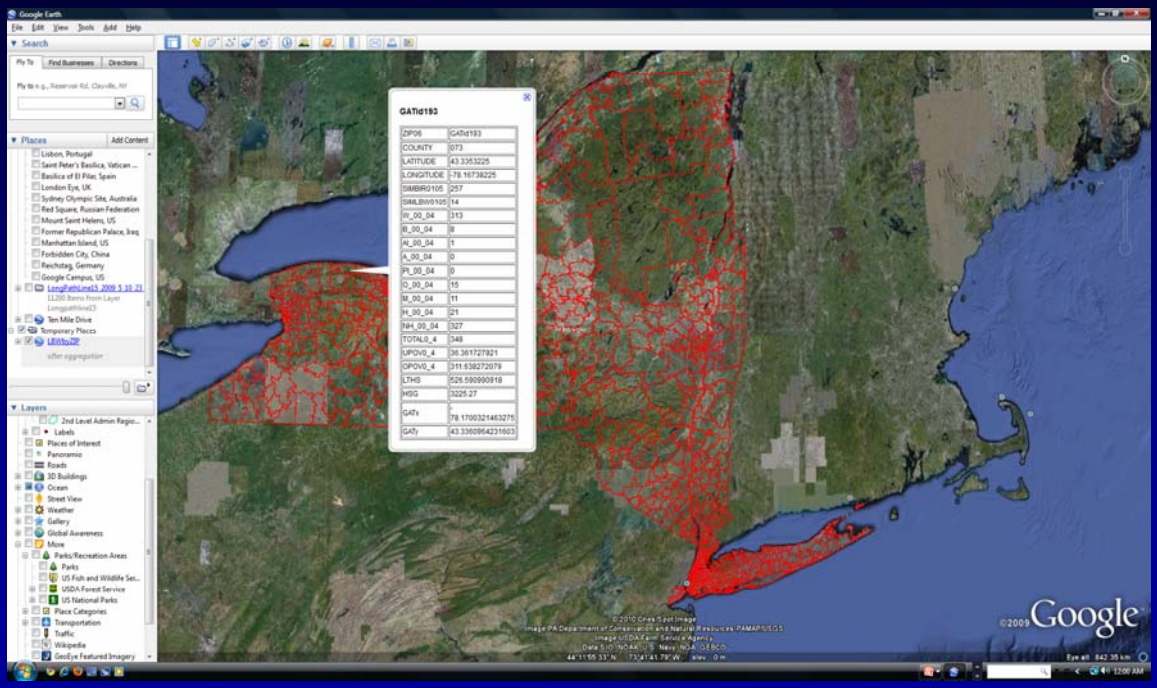
Year 2000 populated census blocks

Statistic (calculated using populated regions only)	Original Census Blocks	New Regions: Level of Aggregation		
		6 cases	12 cases	24 cases
Number of regions	225,167	39,748	21,525	11,381
Median Population	39	385	770	1,467
Median number of cases	1	10	20	38
Median number of blocks	1	4	7	14

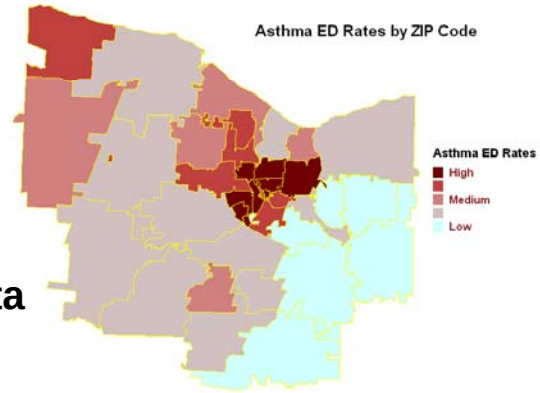
NYS number of cases (5 yrs) 470,000
NYS population 2000 18,976,457

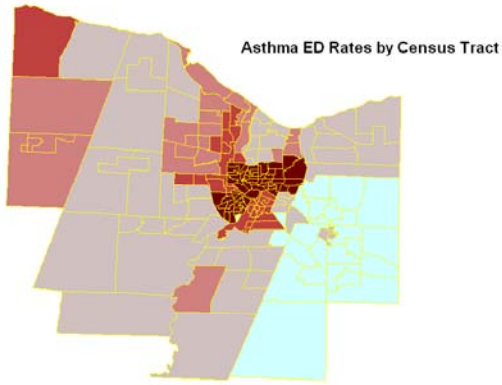
Note: The range in the census block populations is 0 - 23,373 Persons

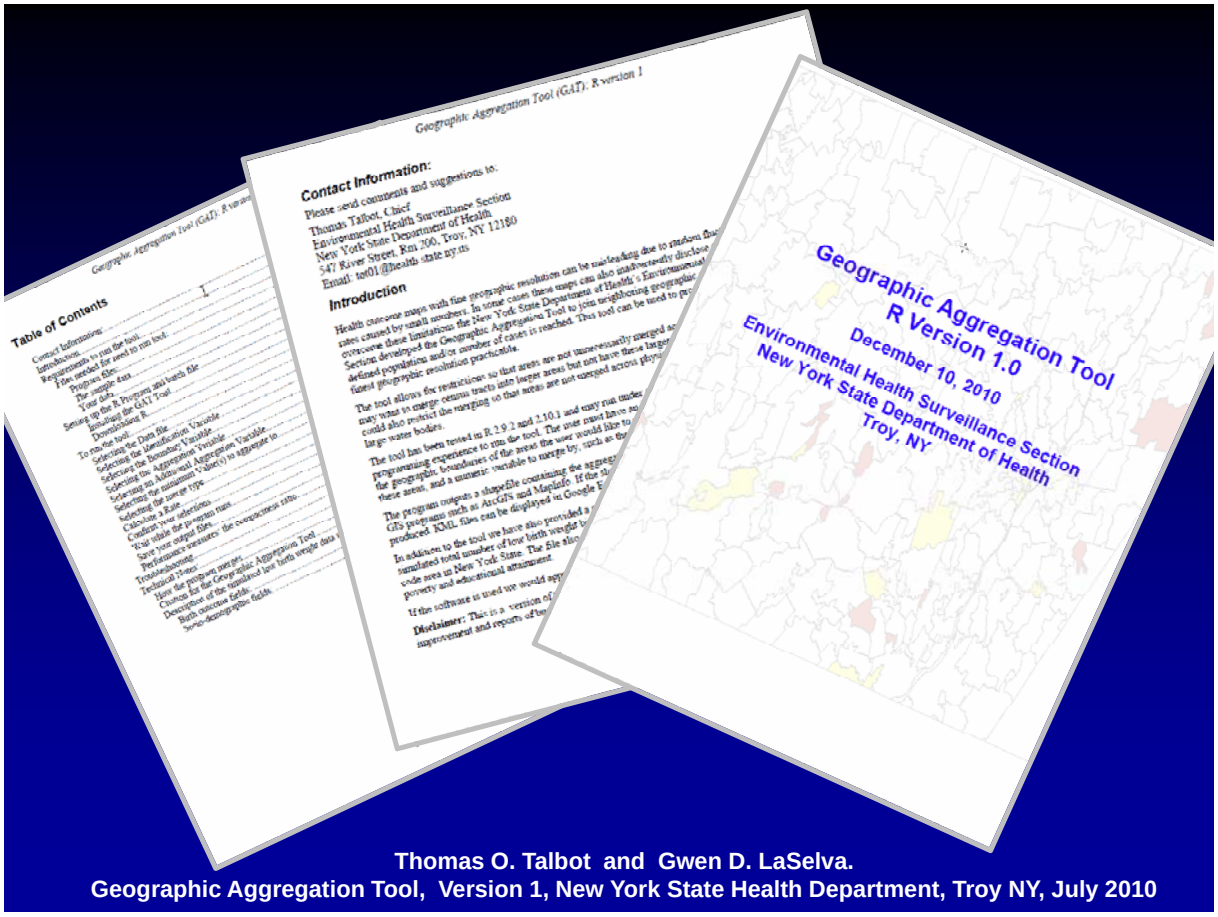
GAT Outputs SHP & KML Files



Converting ZIP Code Level Data to Census Tracts



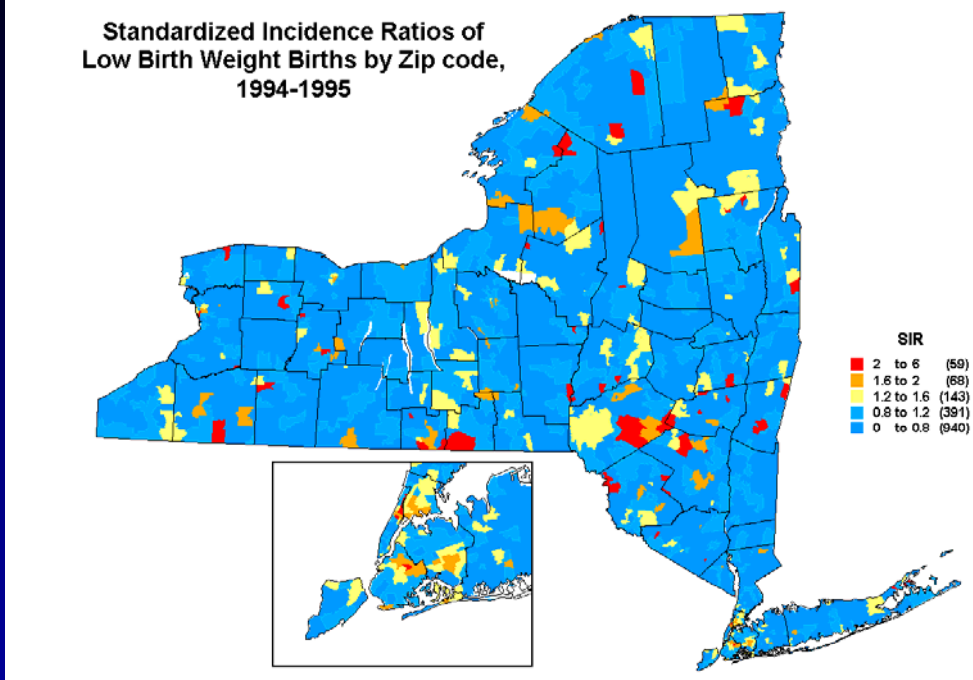




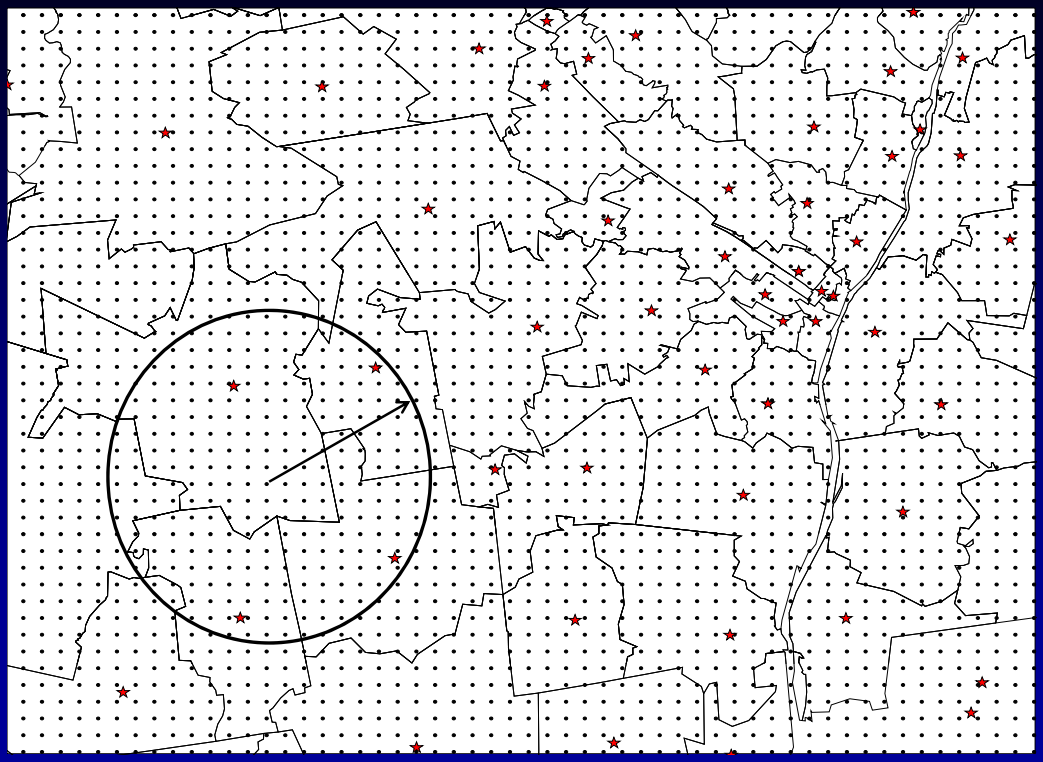
Thomas O. Talbot and Gwen D. LaSelva.
Geographic Aggregation Tool, Version 1, New York State Health Department, Troy NY, July 2010

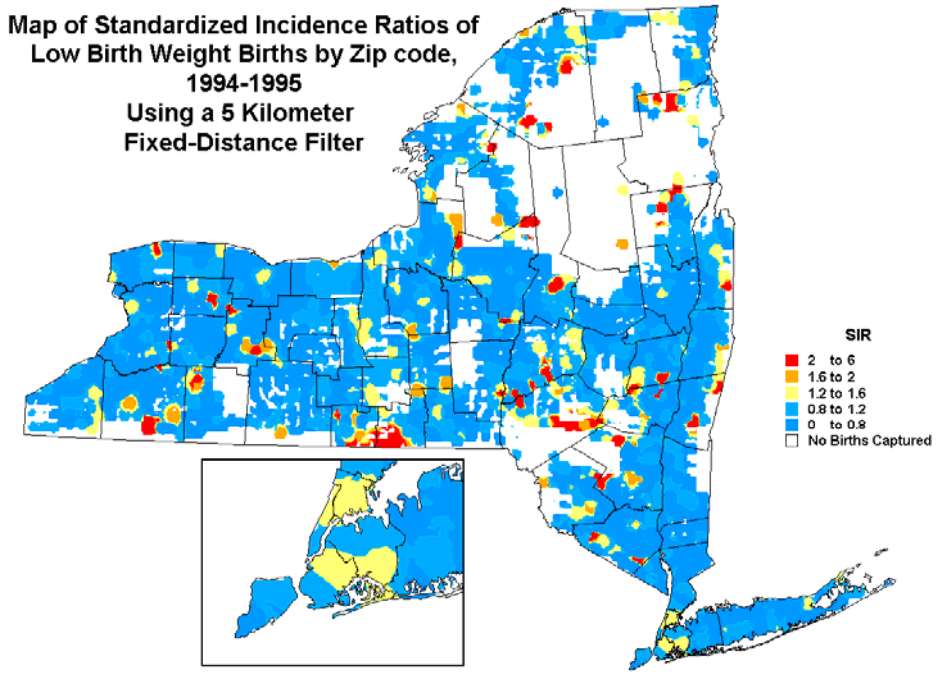
Smoothed Rate Maps

- Borrow data from neighboring areas to provide more stable rates of disease.
  - Spatial Filters
    - Fixed filter size
    - Adaptive spatial filter
  - Bayes
    - Empirical
    - Hierarchal

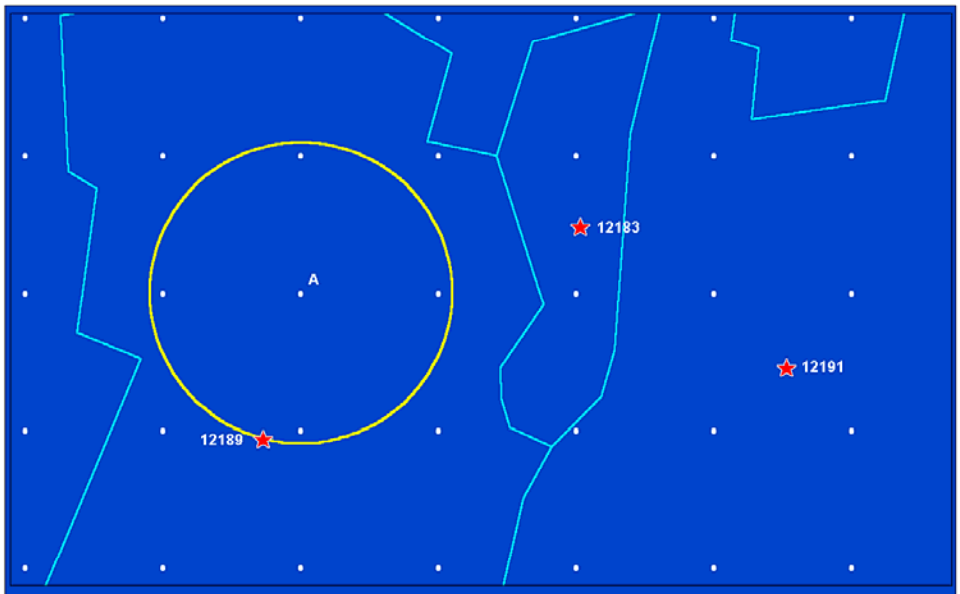


Spatial Filtering 1: fixed distance



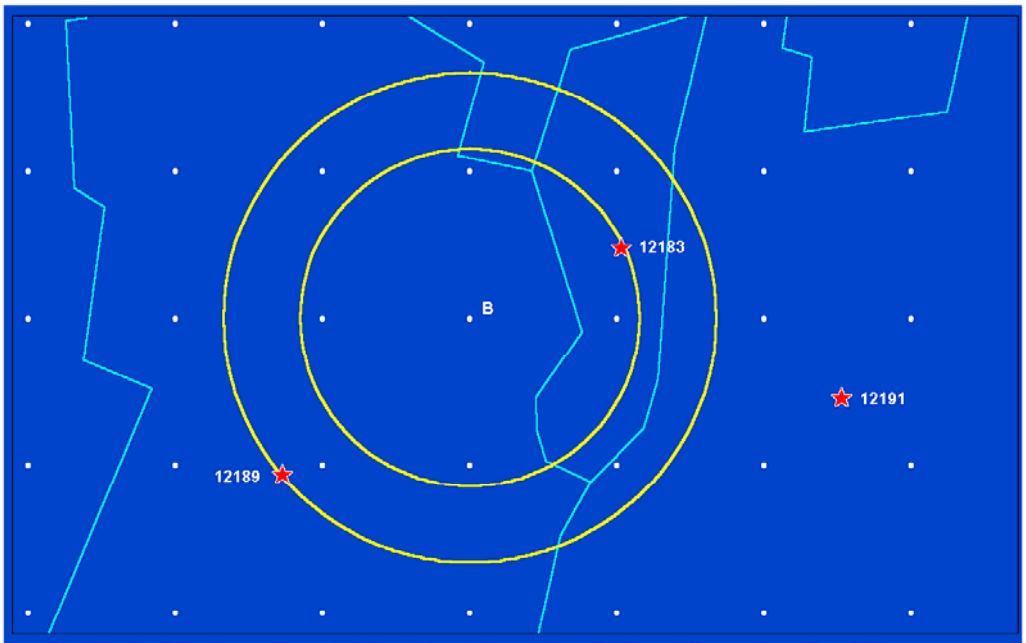


Spatial Filtering 2: minimum population



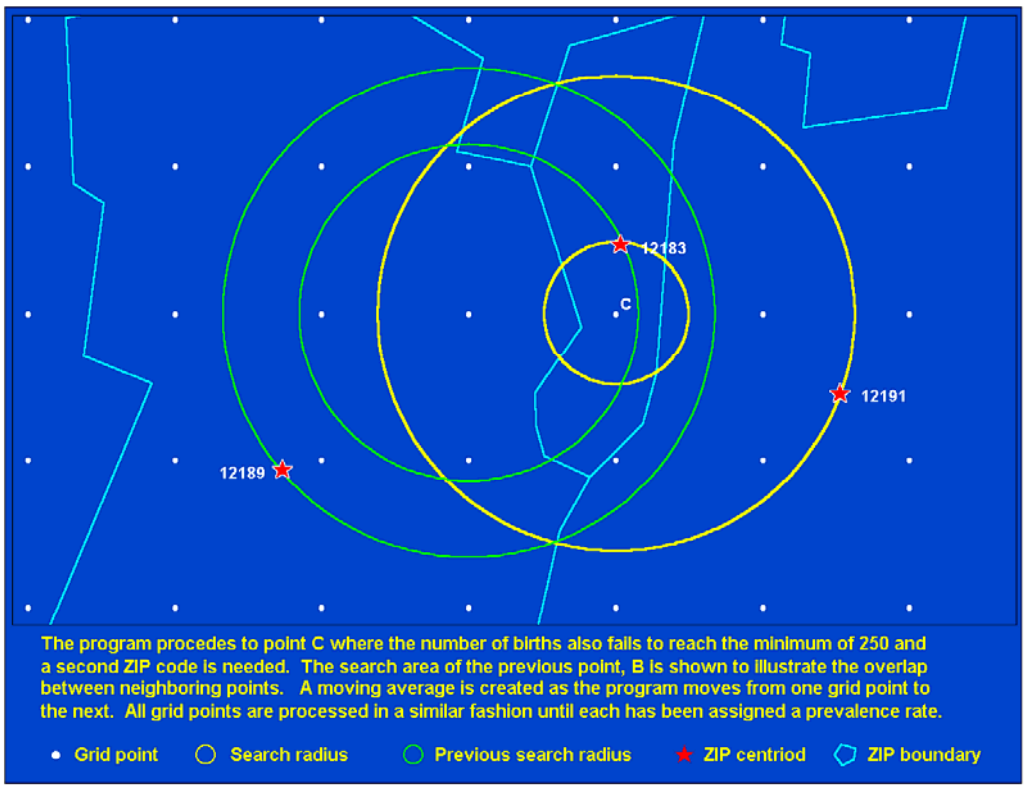
First a grid point (A) is selected and the nearest ZIP code centroid is identified. The number of births number of low birth weight births are tabulated. If the number of births exceeds 250 then the proportion of lbw births is calculated for the grid point and the program moves on to the next grid point (B).

Grid point Search radius ZIP code centriod ZIP code boundary



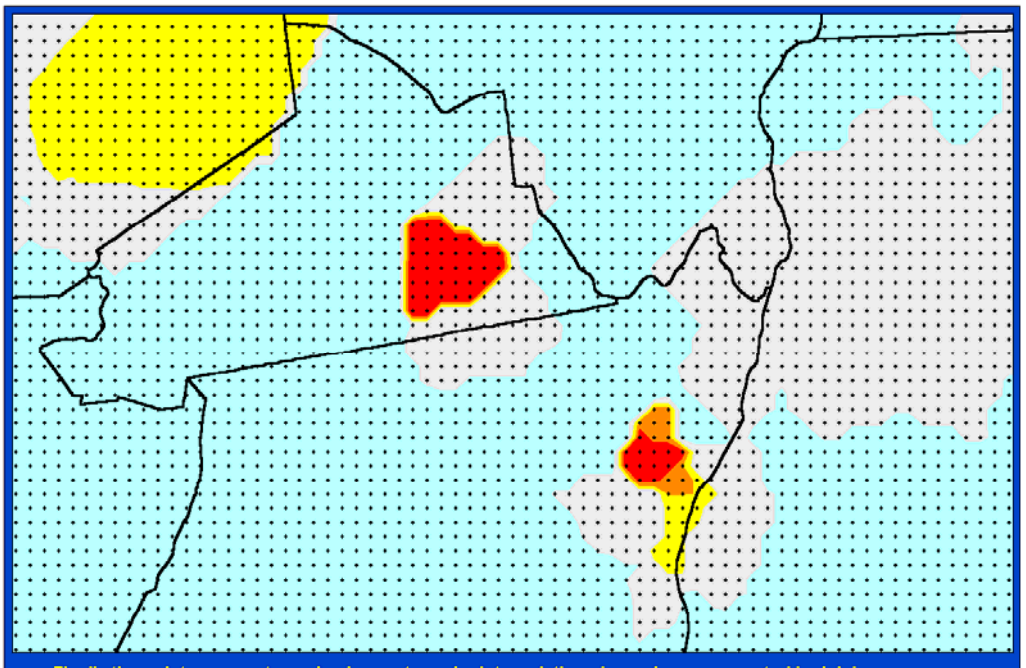
At grid point B the closest Zip code is located and the number of births and the number of low birth weight births are tabulated as in the previous step. In this case however, the number of births is less than 250 so the next nearest Zip code centroid is located and the numbers are summed. This continues until the number of births reaches 200 at which time the rate of lbw is calculated.

Grid point Search radius ZIP code centriod ZIP code boundary



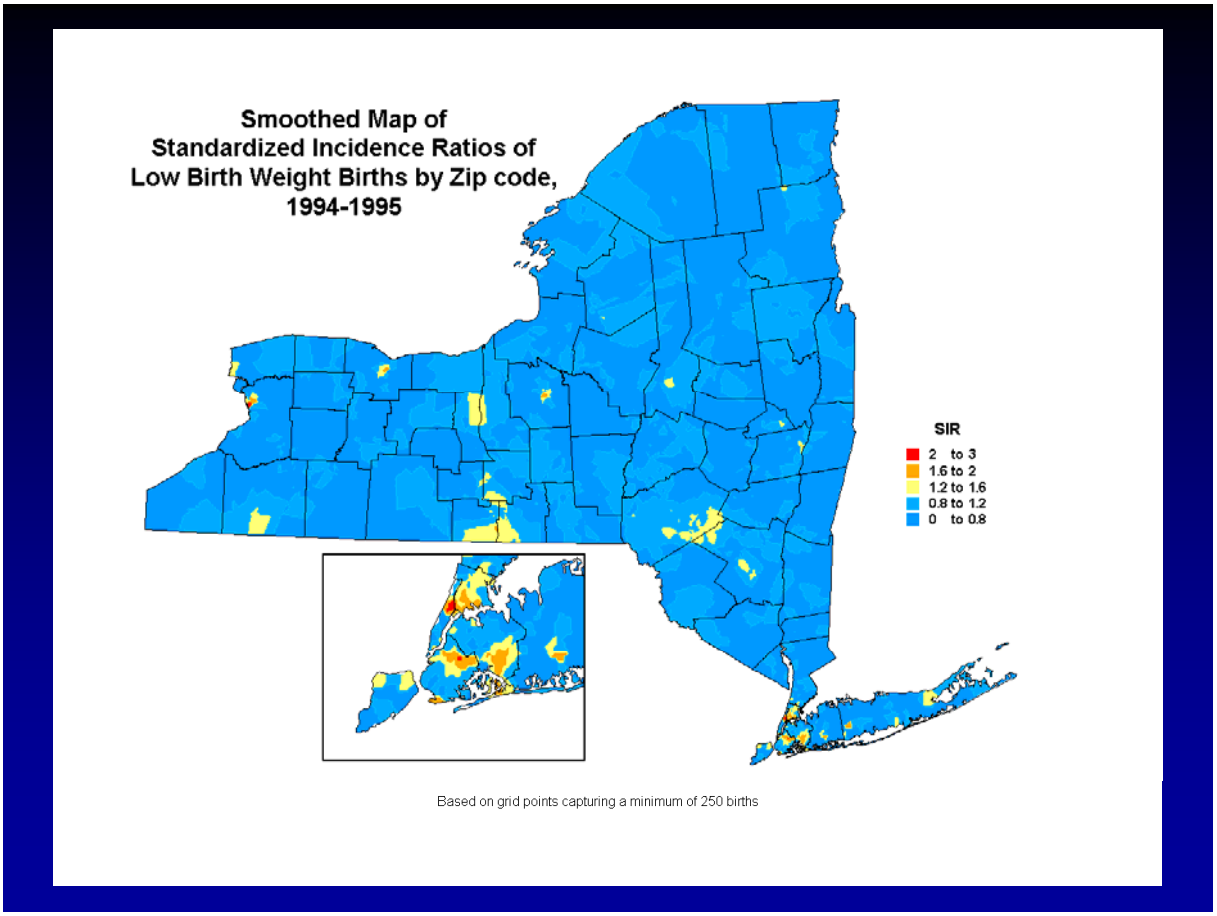
The program procedes to point C where the number of births also fails to reach the minimum of 250 and a second ZIP code is needed. The search area of the previous point, B is shown to illustrate the overlap between neighboring points. A moving average is created as the program moves from one grid point to the next. All grid points are processed in a similar fashion until each has been assigned a prevalence rate.

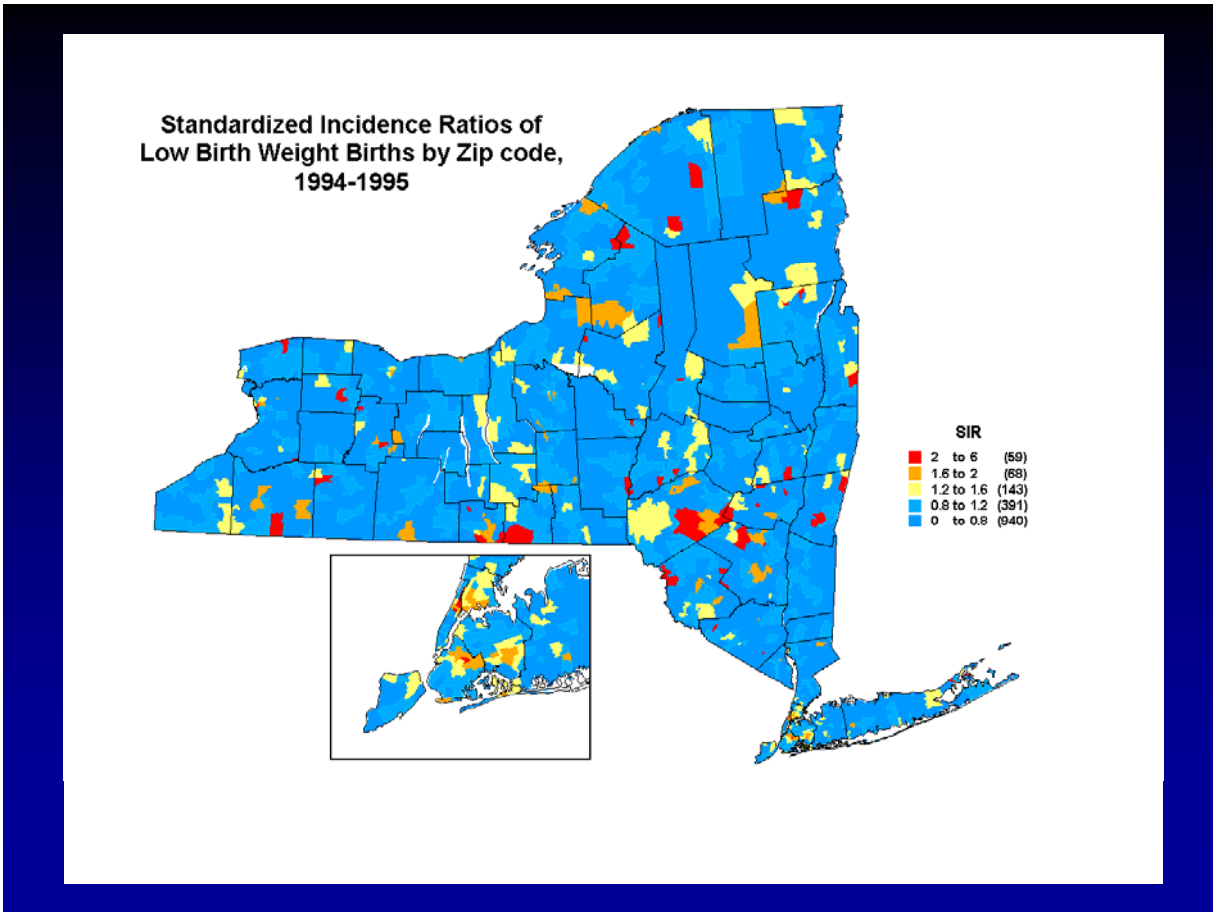
Grid point | Search radius | Previous search radius | ZIP centriod | ZIP boundary



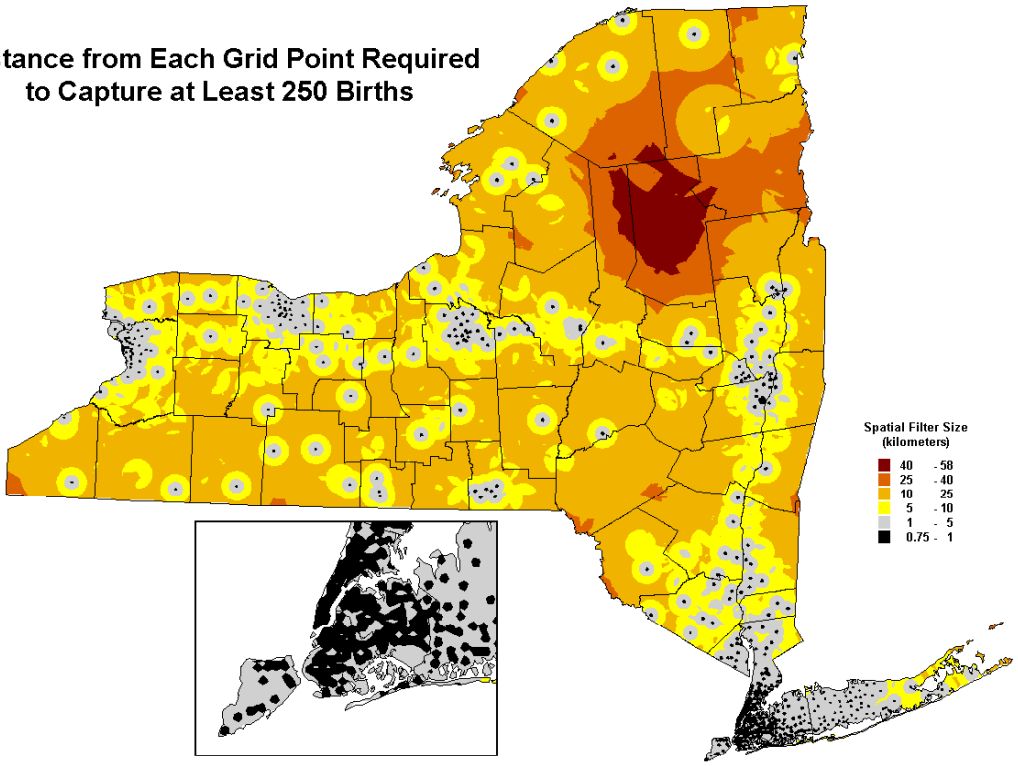
Finally the points are contoured using rectangular interpolation. Isoregions are created by joining areas with similar disease rates.

Grid points | County boundary | Isoregions

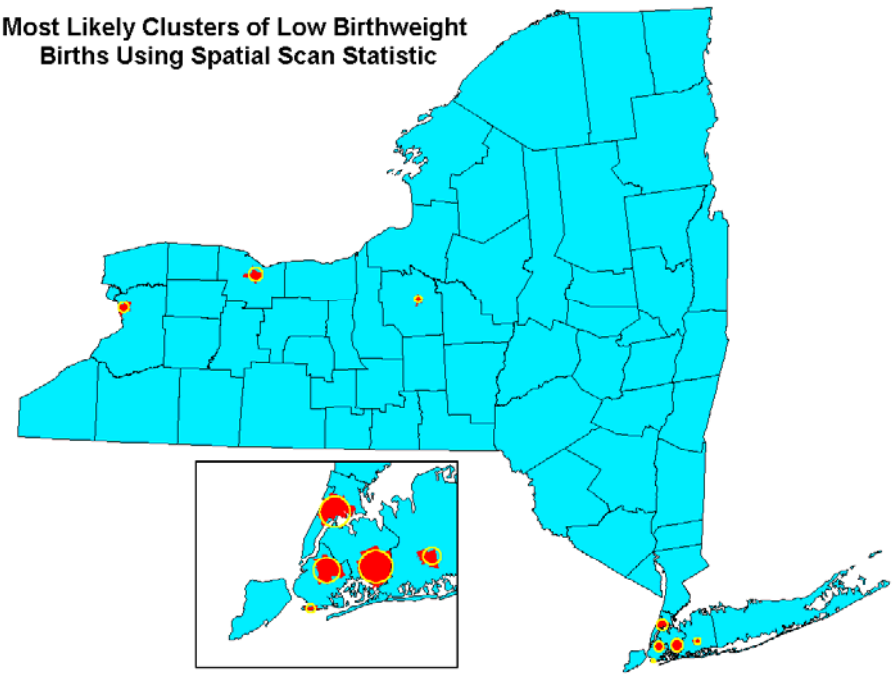

Smoothed Map of
Standardized Incidence Ratios of
Low Birth Weight Births by Zip code,
1994-1995
SIR
2 to 3
1.6 to 2
1.2 to 1.6
0.8 to 1.2
0 to 0.8
Based on grid points capturing a minimum of 250 births


Standardized Incidence Ratios of
Low Birth Weight Births by Zip code,
1994-1995
SIR
2 to 6 (59)
1.6 to 2 (68)
1.2 to 1.6 (143)
0.8 to 1.2 (391)
0 to 0.8 (940)

Distance from Each Grid Point Required to Capture at Least 250 Births



Most Likely Clusters of Low Birthweight Births Using Spatial Scan Statistic



p<0.05 Restictions; no cluster can contain more than 10% of births.

Empirical Bayesian (EB)

- Similar to spatial filters, but process of "pooling" of information more formal
- Difference: The degree to which area rate is modified by depends on how much information is available for the area.
  - Areas with large populations or many events will not be altered as much
  - Areas with small populations or few events will be smoothed towards the mean of the entire area.

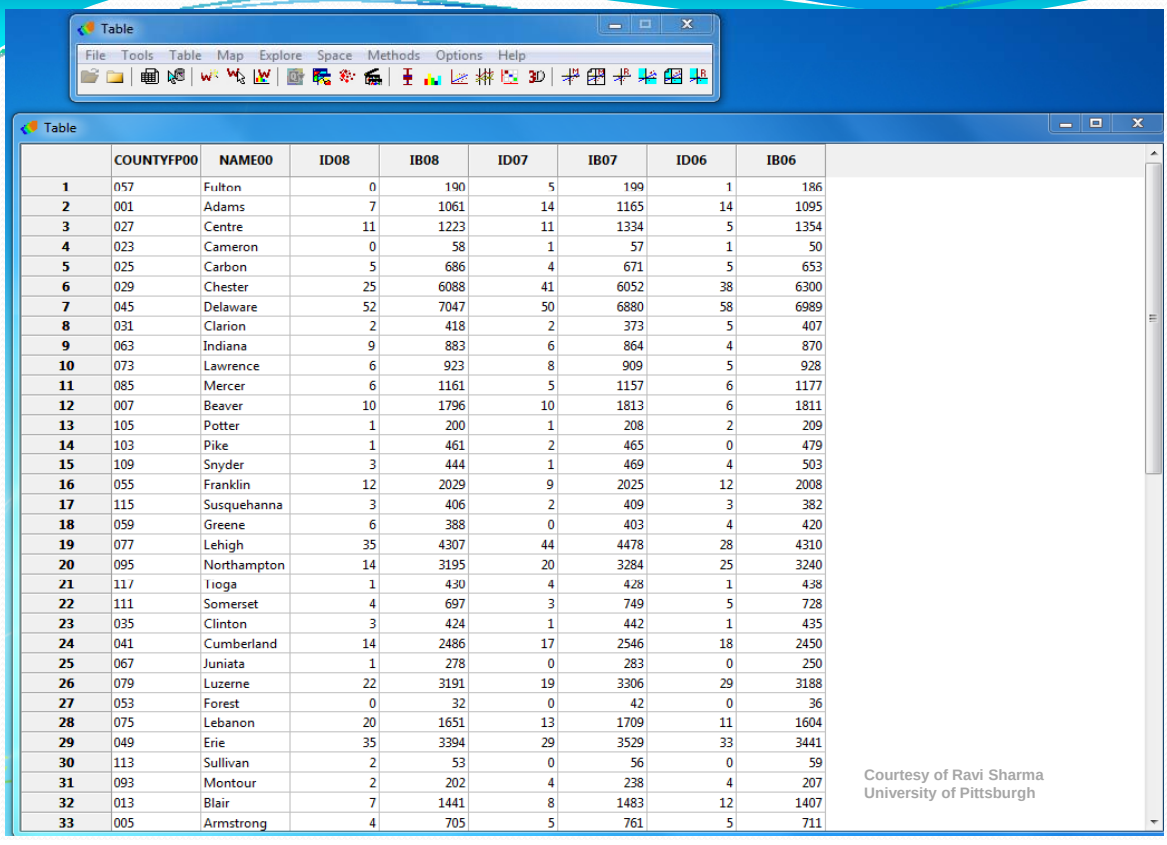
A localized version smoothes unstable rates towards a local neighborhood mean only, therefore preserving more of the spatial pattern.
(Marshall, 1991, Applied Statistics, 40(2):283-294)

Both of these empirical Bayes smoothers can be easily applied in GeoDa software.

What is GeoDa?

- GeoDa is a free software program and OpenGeoDa is the cross-platform, open source version of Legacy GeoDa. OpenGeoDa runs on different versions of Windows (including XP, Vista and 7), Mac OS, and Linux.
- GeoDa can be downloaded from: http://geodacenter.asu.edu/

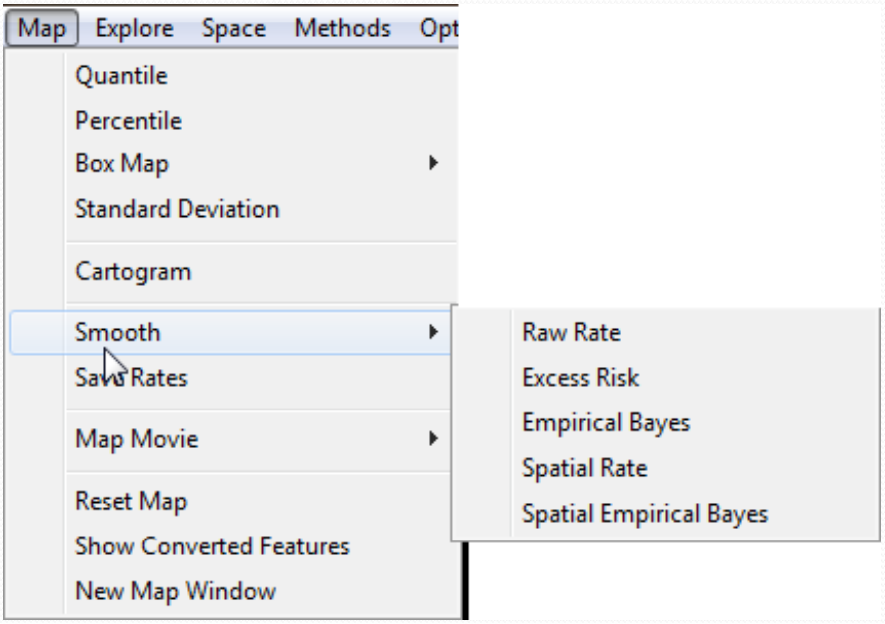
Infant deaths & Births by Pennsylvania counties, 2006-2008



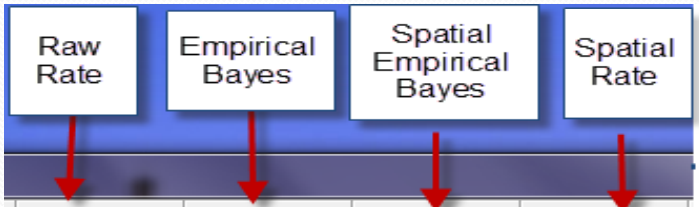
	COUNTYFP00	NAME00	ID08	IB08	ID07	IB07	ID06	IB06
1	057	Fulton	0	190	5	199	1	186
2	001	Adams	7	1061	14	1165	14	1095
3	027	Centre	11	1223	11	1334	5	1354
4	023	Cameron	0	58	1	57	1	50
5	025	Carbon	5	686	4	671	5	653
6	029	Chester	25	6088	41	6052	38	6300
7	045	Delaware	52	7047	50	6880	58	6989
8	031	Clarion	2	418	2	373	5	407
9	063	Indiana	9	883	6	864	4	870
10	073	Lawrence	6	923	8	909	5	928
11	085	Mercer	6	1161	5	1157	6	1177
12	007	Beaver	10	1796	10	1813	6	1811
13	105	Potter	1	200	1	208	2	209
14	103	Pike	1	461	2	465	0	479
15	109	Snyder	3	444	1	469	4	503
16	055	Franklin	12	2029	9	2025	12	2008
17	115	Susquehanna	3	406	2	409	3	382
18	059	Greene	6	388	0	403	4	420
19	077	Lehigh	35	4307	44	4478	28	4310
20	095	Northampton	14	3195	20	3284	25	3240
21	117	Tioga	1	430	4	428	1	438
22	111	Somerset	4	697	3	749	5	728
23	035	Clinton	3	424	1	442	1	435
24	041	Cumberland	14	2486	17	2546	18	2450
25	067	Juniata	1	278	0	283	0	250
26	079	Luzerne	22	3191	19	3306	29	3188
27	053	Forest	0	32	0	42	0	36
28	075	Lebanon	20	1651	13	1709	11	1604
29	049	Erie	35	3394	29	3529	33	3441
30	113	Sullivan	2	53	0	56	0	59
31	093	Montour	2	202	4	238	4	207
32	013	Blair	7	1441	8	1483	12	1407
33	005	Armstrong	4	705	5	761	5	711

Courtesy of Ravi Sharma
University of Pittsburgh

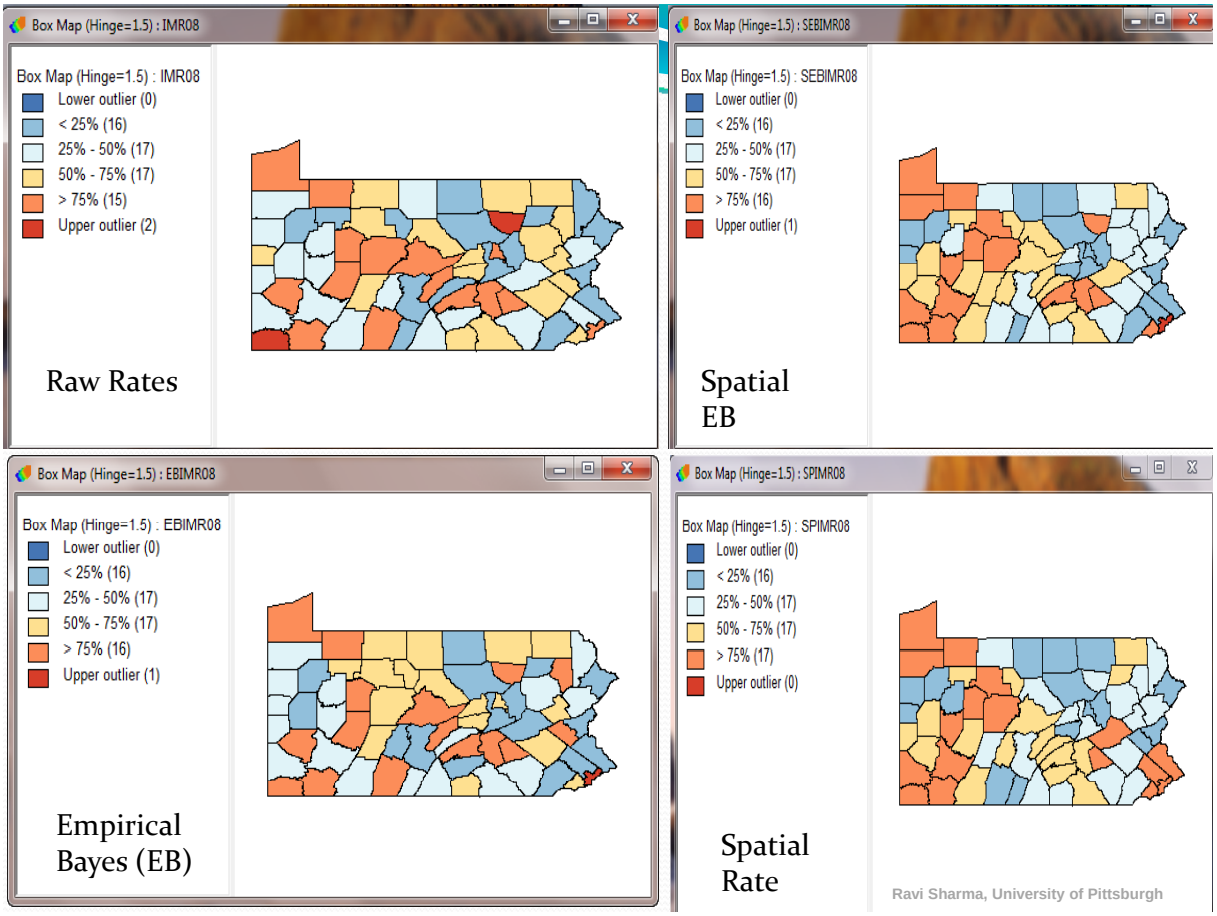
GeoDa Rate Smoothing Options



Raw & Smoothed Rates



	IMR08	EBIMR08	SEBIMR08	SPIMR08
)	0.0000000	6.7880700	5.5658600	5.5658600
L	6.5975500	7.0995500	6.2684400	6.2684400
3	8.9942800	7.8795600	7.0789900	7.0789900
3	0.0000000	7.1481100	7.0827800	7.0827800
5	7.2886300	7.3120600	6.3895600	6.3895600
3	4.1064400	5.0230000	5.2913300	5.9387800
7	7.3790300	7.3635500	7.5009100	8.2195300
3	4.7846900	6.9468500	5.6166600	5.6166600
3	10.1925000	8.0844800	7.3872500	7.3872500
3	6.5005400	7.0935100	5.4431000	5.4431000
L	5.1679600	6.6234600	5.3172600	5.3172600
5	5.5679300	6.5747500	7.4366100	7.4366100
)	5.0000000	7.1424000	4.5470400	4.5470400
L	2.1692000	6.4977300	4.7713700	4.7713700
4	6.7567600	7.2318900	5.4957700	5.4190800
9	5.9142400	6.6797100	5.7934200	5.7934200
5	7.3891600	7.3287700	6.9394600	6.9394600
3	15.4639000	8.4399000	9.6219100	8.9947100
7	8.1263100	7.8349500	6.3785700	5.6776000



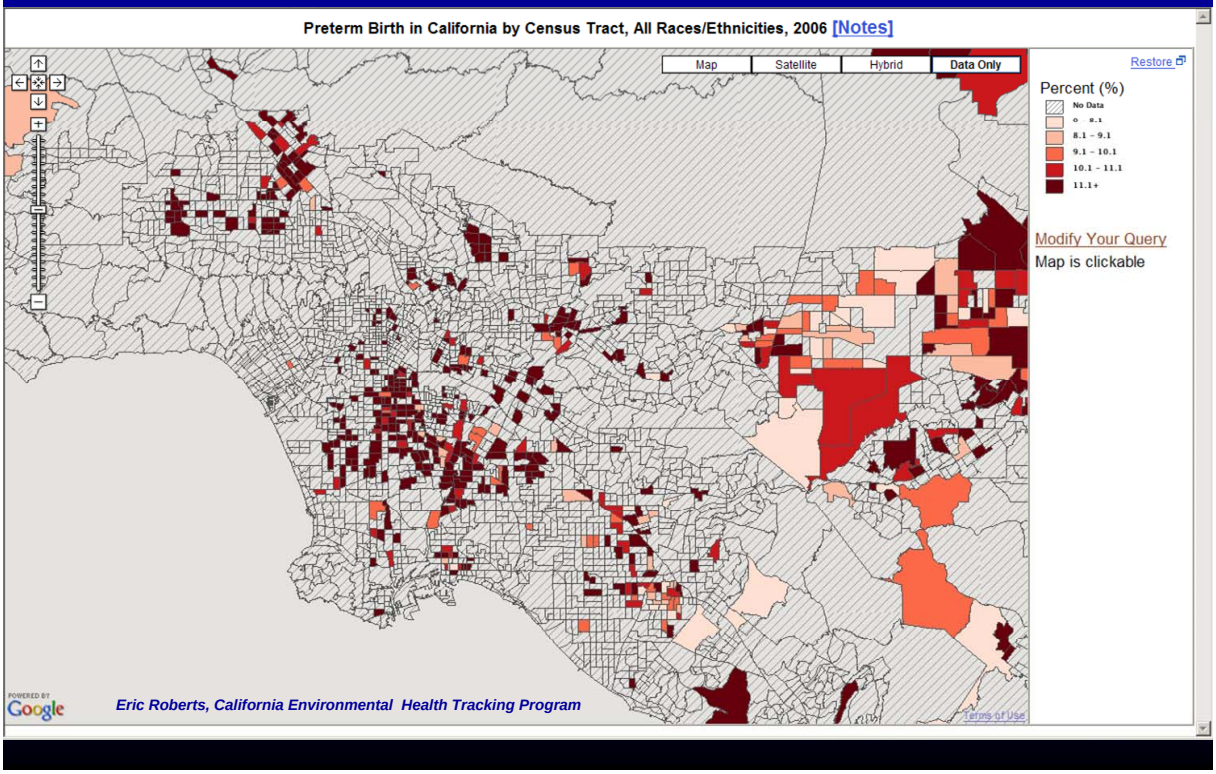
Hierarchical Bayesian (HB)

- Reasoning is very similar to EB
- The difference:
  - Bandwidth is considered an **unknown variable** that is calculated from the data themselves
  - Data with little spatial structure will be "flattened out" more
  - Data with more spatial structure will be allowed to show local variability

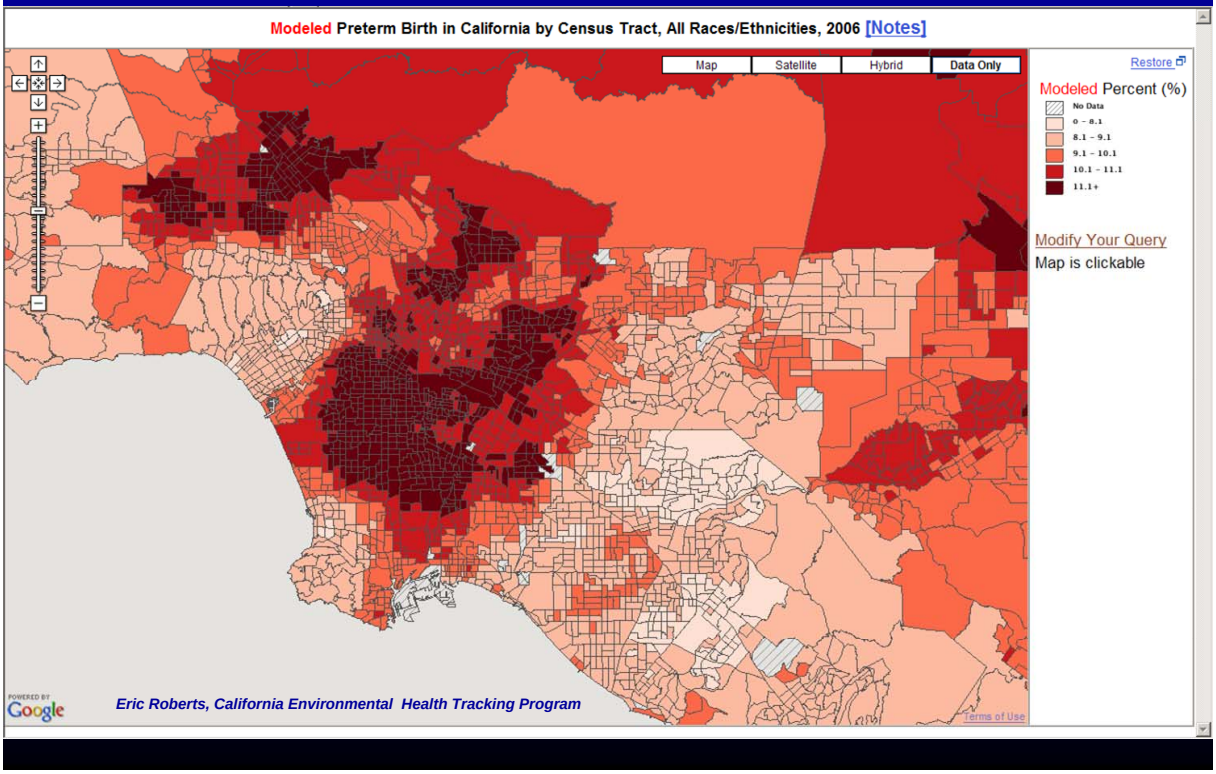
Hierarchical Bayesian

- Note: Functionally, a single bandwidth is still calculated for the entire map
- Generally speaking, HB results in bandwidths that are more conservative (flattening) than people expect

Data suppression example



Rate stabilized version (HB)



Bayesian Modeling References

- Waller, L.A. and Gotway, C.A. 2004. Applied Spatial Statistics for Public Health Data. Wiley. 494 pp.
- Johnson, G.D. 2004. Smoothing Small Area Maps of Prostate Cancer Incidence in New York State (USA) using Fully Bayesian Hierarchical Modelling. Int. J. Health Geographics 2004, 3:29 (http://www.ij-healthgeographics.com/content/3/1/29)
- Elliot, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. 2000. Spatial Epidemiology: Methods and Applications. Oxford. 475 pp.
- Statistics in Medicine. 2000. Vol. 19 (special issue on disease mapping)
- Lawson, A. et al. 1999. Disease Mapping and Risk Assessment for Public Health. Wiley. 482 pp.

Smoothing Software Sources

Spatial Filters / Kernel DensitySmoothing

DMap 4 **http://www.uiowa.edu/~gishlth/DMAP4/**

ArcGis www.esri.com

CrimeStat III http://www.icpsr.umich.edu/CrimeStat

Empirical Bayes

GeoDa http://geodacenter.asu.edu/

Hierarchal Bayes

WINBUGS http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

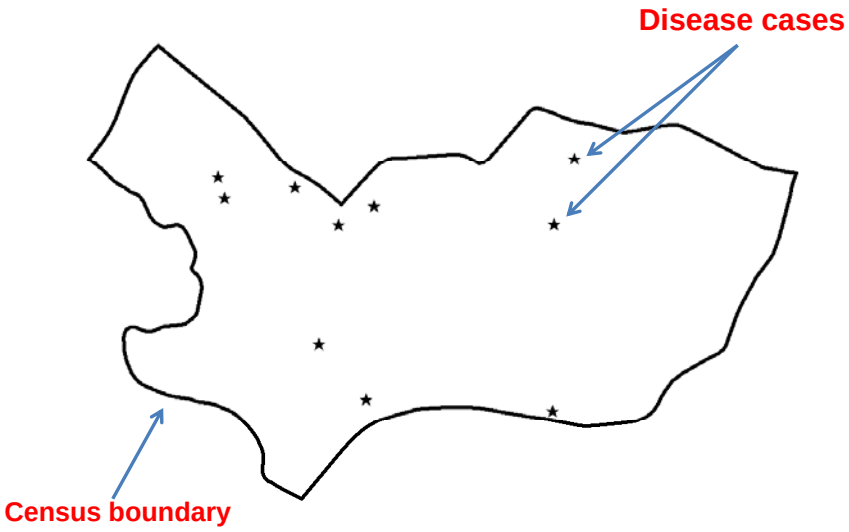
Head-Banging

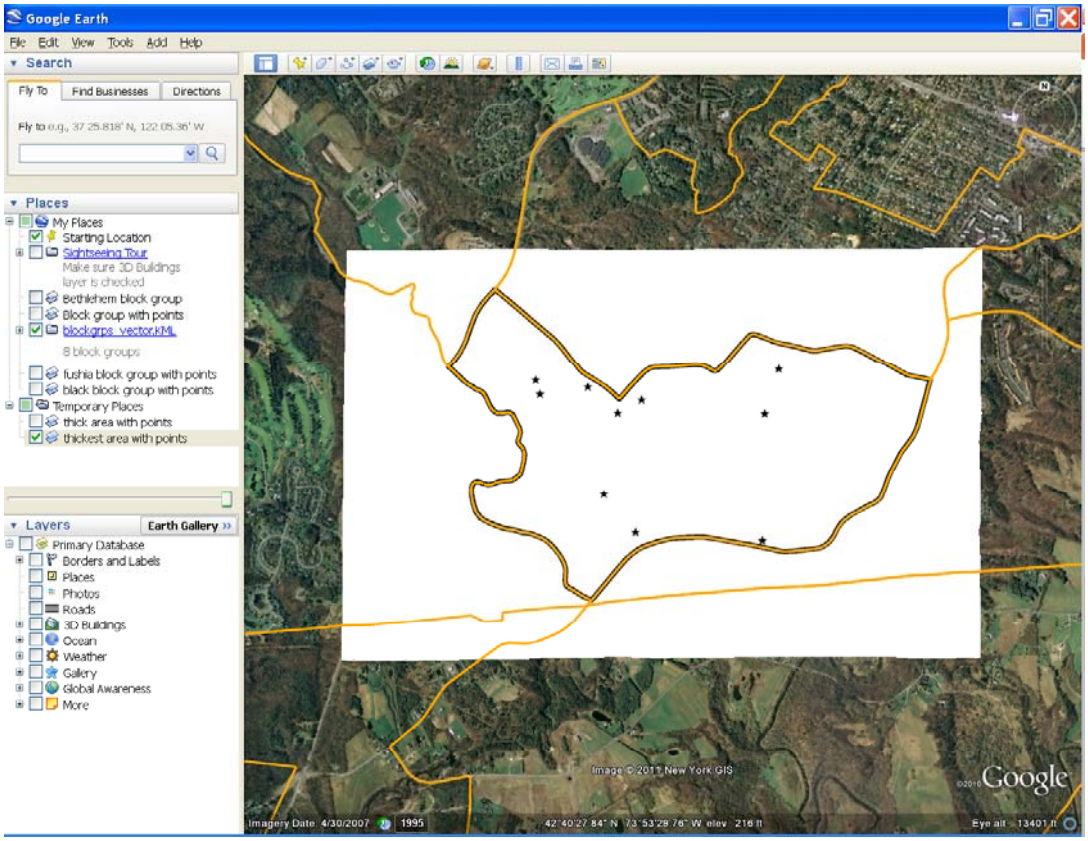
Head Bang http://surveillance.cancer.gov/headbang/

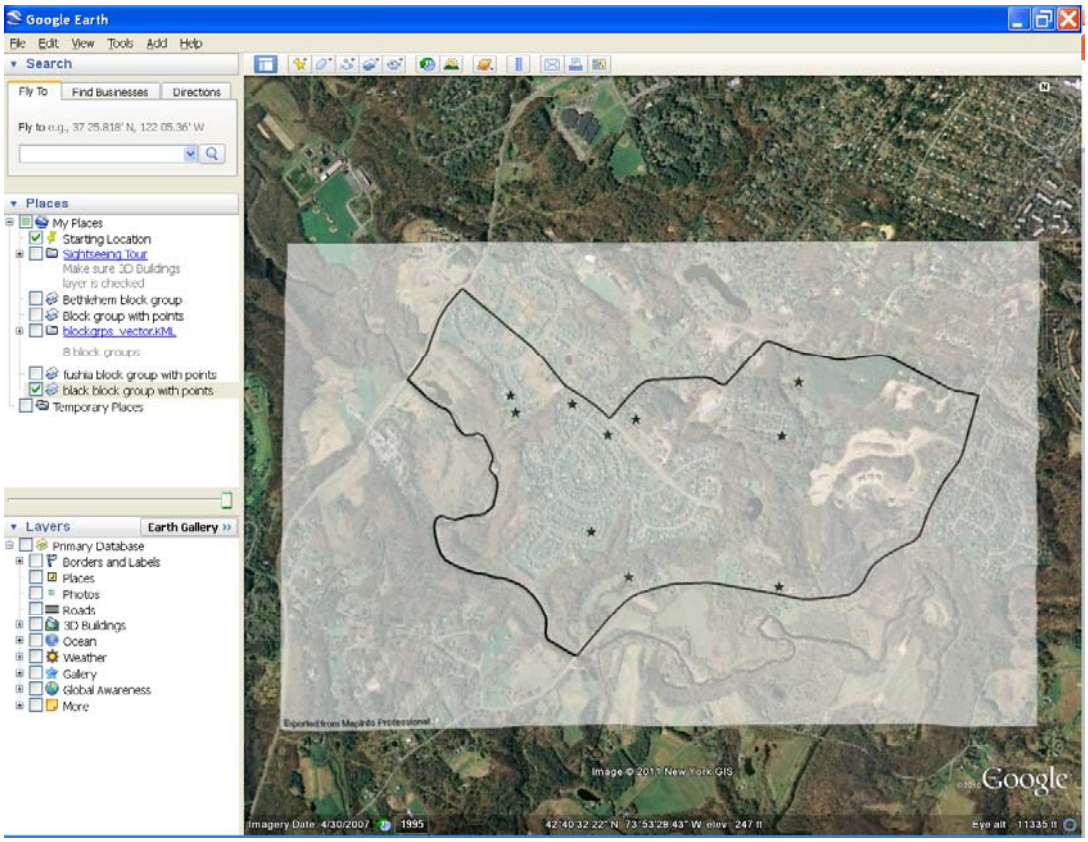
Masking

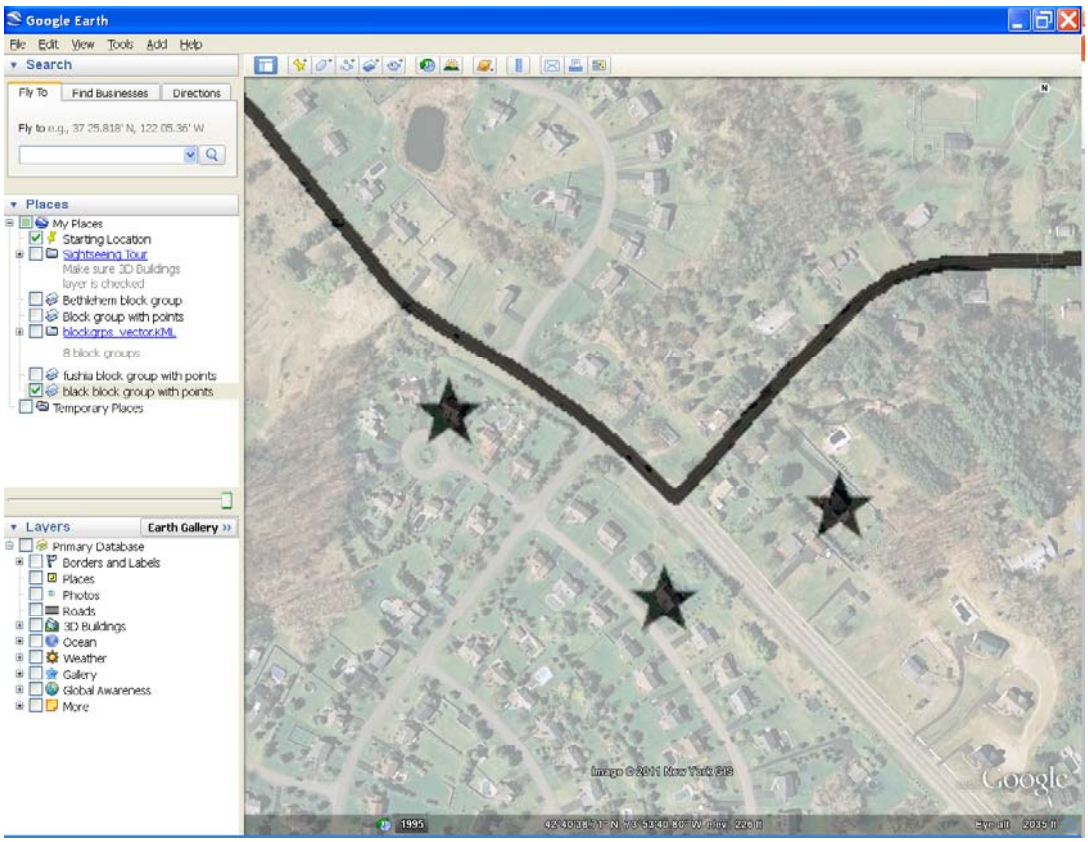
- Masking obscures the true location or the true attributes for a given geography.
- It is possible to reverse engineer data in tables and maps to reveal confidential information.
- Consider how your masking product will be used (Analysis, Visualization).

Maps can be registered to real-world coordinate systems

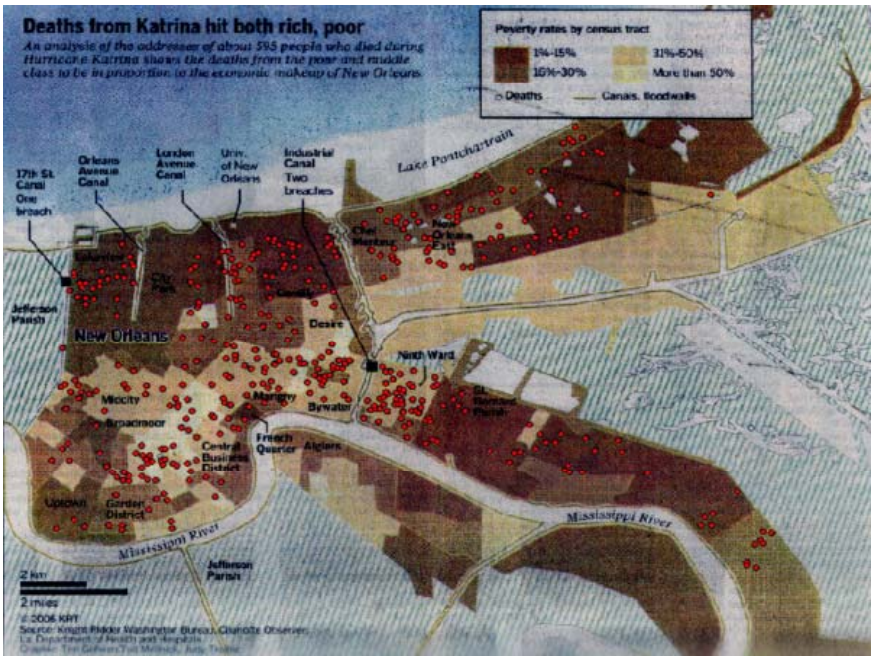





Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., 37 25.818' N, 122 05.36' W
Places
My Places
Starting Location
Sightseeing Tour
Make sure 3D Buildings layer is checked
Bethlehem block group
Block group with points
blockgrps_vector.KML
8 block groups
fushia block group with points
black block group with points
Temporary Places
Layers
Earth Gallery
Primary Database
Borders and Labels
Places
Photos
Roads
3D Buildings
Ocean
Weather
Gallery
Global Awareness
More
Image © 2011 New York GIS
Google
Imagery Date: 4/30/2007
1995
Eye alt 11335 ft

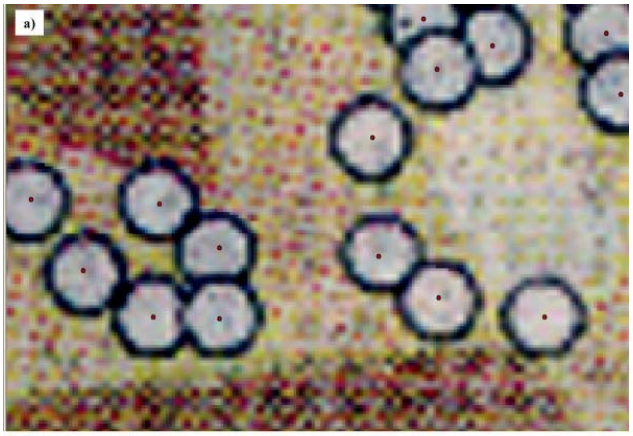

Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., 37 25.818' N, 122 05.36' W
Places
My Places
Starting Location
Sightseeing Tour
Make sure 3D Buildings layer is checked
Bethlehem block group
Block group with points
blockgrps_vector.KML
8 block groups
fushia block group with points
black block group with points
Temporary Places
Layers
Earth Gallery
Primary Database
Borders and Labels
Places
Photos
Roads
3D Buildings
Ocean
Weather
Gallery
Global Awareness
More
Image © 2011 New York GIS
Google
1995

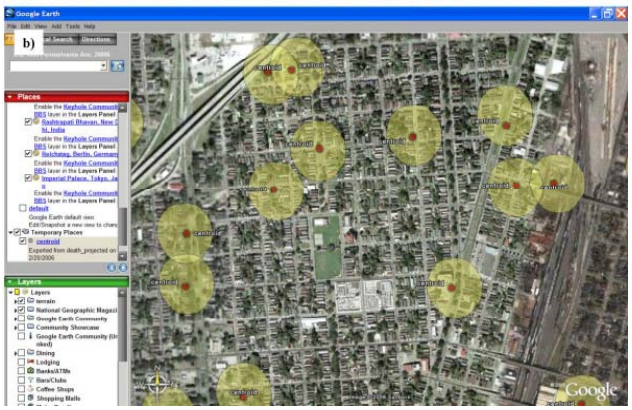
Hurricane Katrina example.



From: Curtis, AJ. Mills, J W. Leitner, M. , 2006

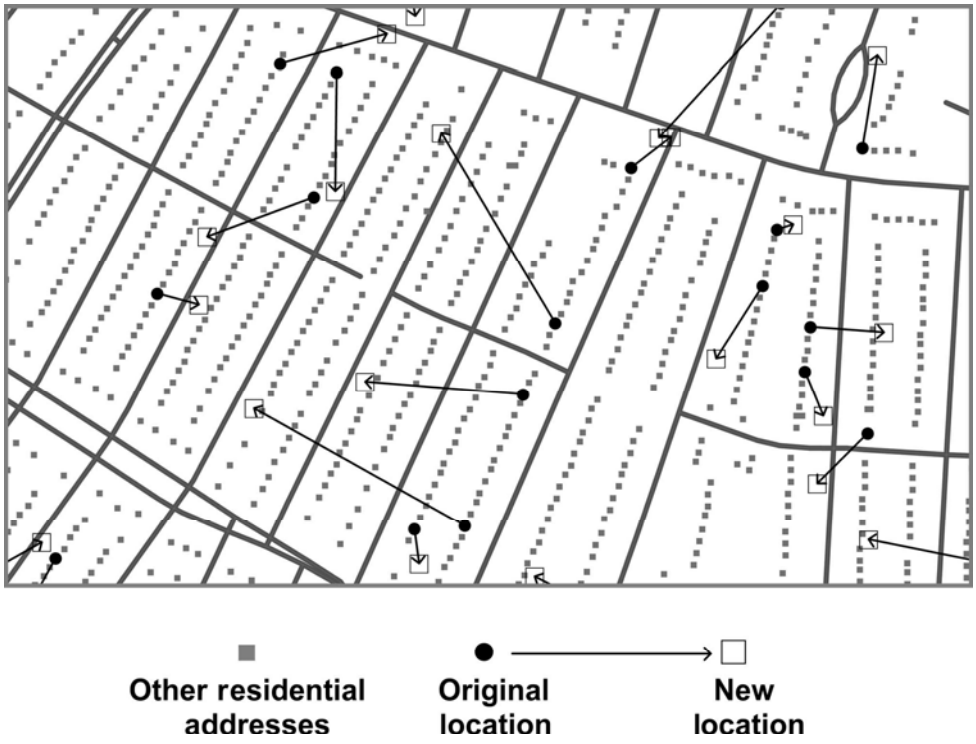
Hurricane Katrina example.





From: Curtis, AJ. Mills, J W. Leitner, M. , 2006

Randomly Moving Points



Example: to move a point within a 500 meter square in Excel

New Latitude =

(RAND()-0.5)*500/111000 + original latitude

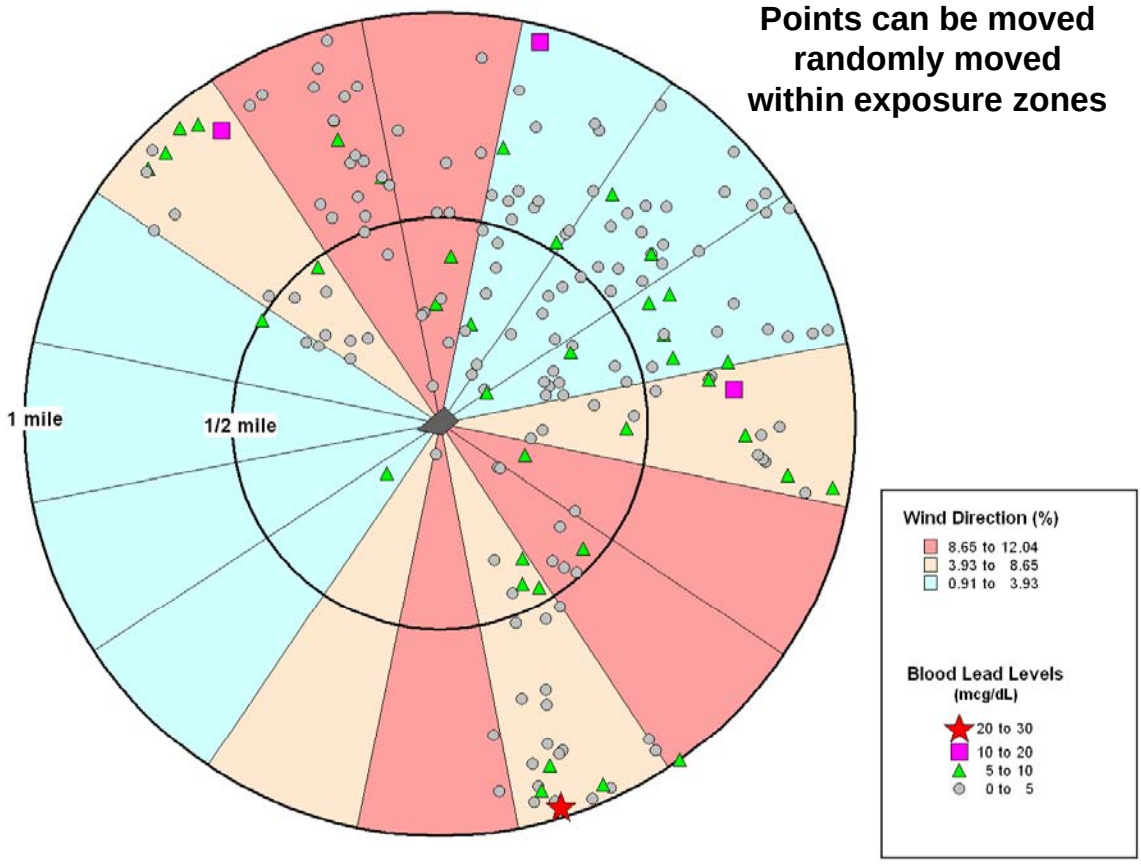
New Longitude =

(RAND()-0.5)*500/(COS(RADIANS(original latitude))*111321)+original longitude

RADIANS() a function which converts degrees to radians

1 degree= π/180 radians

RAND() a function that generates a random number from 0 to 1



US Census Bureau (partial list)

Noise infusion: Insert randomly-generated noise. Distorts values for 'risky' items.

Data swapping: For example, rank-based proximity swapping. Sort values/swap values so that exchanges can occur within a prescribed limit.

Synthetic data: Generate synthetic data according to a model. The synthesized data is 'similar to' and has 'relationships consistent with' the real information.

Masking References

- Geospatial Modeling Environment for ArcGIS10: www.spatialecology.com
- Talbot, TO.; Kumar, S.; Babcock, GD.; Haley, Valerie B.; Forand, SP.; Hwang, S. *Development of an Interactive Environmental Public Health Tracking System for Data Analysis, Visualization, and Reporting*. Journal of Public Health Management & Practice. 14(6):526-532, November/December 2008
- Winkler, WE. *Examples of Easy-to-implement Widely Used Methods of Masking for which Analytic Properties are not Justified.* Research Report Series *(Statistics #2007-21).* US Census Bureau
- http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf
- Curtis, AJ. Mills, J W. Leitner, M. *Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina*. International Journal of Health Geographics. 2006, 5:44
- http://www.ij-healthgeographics.com/content/5/1/44
- Matthews, G.; Harel, O. *Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy.* Statistics Surveys. Vol. 5:1-29. 2011

Some areas don't need to be smoothed!

Questions
&
Discussion