Privacy and Confidentiality in Health GIS

Gerard Rushton, PhD

Professor, Department of Geography,
Adjunct Professor, Department of Health Management & Policy
The University of Iowa
Iowa City, Iowa

ESRI Health GIS Conference, Scottsdale, AZ
October 10, 2007

Gerard-rushton@uiowa.edu

Organization

1. Conclusions and recommendations of NRC 2007 Report:
    1. "Putting People on the Map: Protecting confidentiality with linked social-spatial data."

2. Brief discussion of importance of access to linked social-spatial data in GIS and Public Health.

3. Methods of masking spatial data on individuals:
    1. technical methods
    2. Institutional methods

4. Conclusions

PUTTING PEOPLE ON THE MAP
PROTECTING CONFIDENTIALITY WITH LINKED SOCIAL-SPATIAL DATA

Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data

Myron P. Gutmann and Paul C. Stern, Editors

Committee on the Human Dimensions of Global Change

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL *OF THE NATIONAL ACADEMIES*
THE NATIONAL ACADEMIES PRESS (2007)

Washington, D.C.

The opportunities from linked social-spatial data

- “The linkage of spatial and social information, …has the potential to revolutionize social science and to significantly advance policy making.” (p.1)

- “The key issue for this study concerns the incremental risks of linking confidential social data to precise spatial information about research participants.” (p.25)

- The report addresses the importance of linked social-spatial data for research and does not address the importance of such data for day-to-day management of health programs. However, the same issues apply to both research and health program management.

- The report was funded by three Federal Agencies: The National Science Foundation, The National Institutes of Health and NASA.

The NRC Report has Four Conclusions

1. Recent advances in the availability of social-spatial data and the development of geographic information systems (GIS) and related techniques to manage and analyze those data give researchers important new ways to study important social, environmental, economic, and health policy issues and are worth further development.

2. The increasing use of linked social-spatial data has created significant uncertainties about the ability to protect the confidentiality promised to research participants. Knowledge is as yet inadequate concerning the conditions under which and the extent to which the availability of spatially explicit data about participants increases the risk of confidentiality breaches.

NRC Panel Conclusions contd.

3. Recent research on technical approaches for reducing the risk of identification and breach of confidentiality has demonstrated promise for future success. At this time, however, no known technical strategy or combination of technical strategies for managing linked spatial-social data adequately resolves conflicts among the objectives of data linkage, open access, data quality, and confidentiality protection across datasets and data uses.
4. Because technical strategies will not be sufficient in the foreseeable future for resolving the conflicting demands for data access, data quality, and confidentiality, institutional approaches will be required to balance those demands.

The NRC Report has Eight Recommendations

#1: Technical and Institutional Research

Federal agencies and other organizations that sponsor the collection and analysis of linked social-spatial data—or that support data that could provide added benefits with such **linkage—should sponsor research into techniques and procedures for disseminating** such data while protecting confidentiality and maintaining the usefulness of the data for social-spatial analysis.

This research **should include studies to adapt existing techniques from other fields**, to understand how the publication of linked social-spatial data might increase disclosure risk, and to explore institutional mechanisms for disseminating linked data while protecting confidentiality and maintaining the usefulness of the data.

2: Education and Training

Faculty, researchers, and organizations involved in the continuing professional development of researchers **should engage in the education of researchers in the ethical use** of spatial data.

Professional associations should participate by establishing and inculcating strong norms for the ethical use and sharing of linked social-spatial data.

3:Training in Ethical Issues

- Training in ethical considerations needs to accompany all methodological training in the acquisition and use of data that include geographically explicit information on research participants.

4: Outreach by Professional Societies and Other Organizations

- Research societies and other research organization that use linked social-spatial data and that have established traditions of protection of the confidentiality of human research participants **should engage in outreach to other research societies and organizations less conversant in research** with issues of human participant protection to increase attention to these issues in the context of the use of personal, identifiable data.

5: Research Design

- Primary researchers who intend to collect and use spatially explicit data **should design their studies in ways that not only take into account the obligation to share data and the disclosure risks posed, but also provide confidentiality protection for human participants** in the primary research as well as in secondary research use of the data. Although the reconciliation of these objectives is difficult, primary researchers should nevertheless assume a significant part of this burden.

6: Institutional Review Boards

- Institutional Review Boards and their organizational sponsors **should develop the expertise needed** to make well-informed decisions that balance the objectives of data access, confidentiality, and quality in research projects that will collect or analyze linked social-spatial data.

7: Data Enclaves

- Data enclaves deserve further development as a way to provide wider access to high-quality data while preserving confidentiality. This development **should focus on the establishment of expanded place-based enclaves, “virtual enclaves,”** and meaningful penalties for misuse of enclaved data.

8: Licensing

- Data stewards **should develop licensing agreements** to provide increased access to linked social-spatial datasets that include confidential information.

Reasons why individually geocoded data with linked social data are often required to answer important research questions

- Taken in the aggregate over many people, long-term large-scale population studies allow the discovery of statistical correlations between environmental factors and disease and are also used to help assess the efficacy of treatments, to determine the overall costs to particular kinds of treatment regimes, and to conduct epidemiological research that can generate insight into the genesis, development, and spread of disease." Waldo et al. 2007, p. 210.

- To put health data into other "regions"—to deal with the problem of spatial misalignment. For example: Health data in relation to a source of contamination:

A Heavy Toll From Disease Fuels Suspicion and Anger

New York Times, October 7, 2007

MIDDLEBOROUGH, Mass., Oct. 6 — The big news in this struggling southeastern Massachusetts community is a proposed $1 billion casino complex that many hope will bring financial salvation.

But for a small group of residents, the hope for economic revival is overshadowed by health concerns. They are awaiting a report later this year that could reveal whether the dozens of cases of Lou Gehrig’s disease centered around a downtown industrial area were caused by pollution.

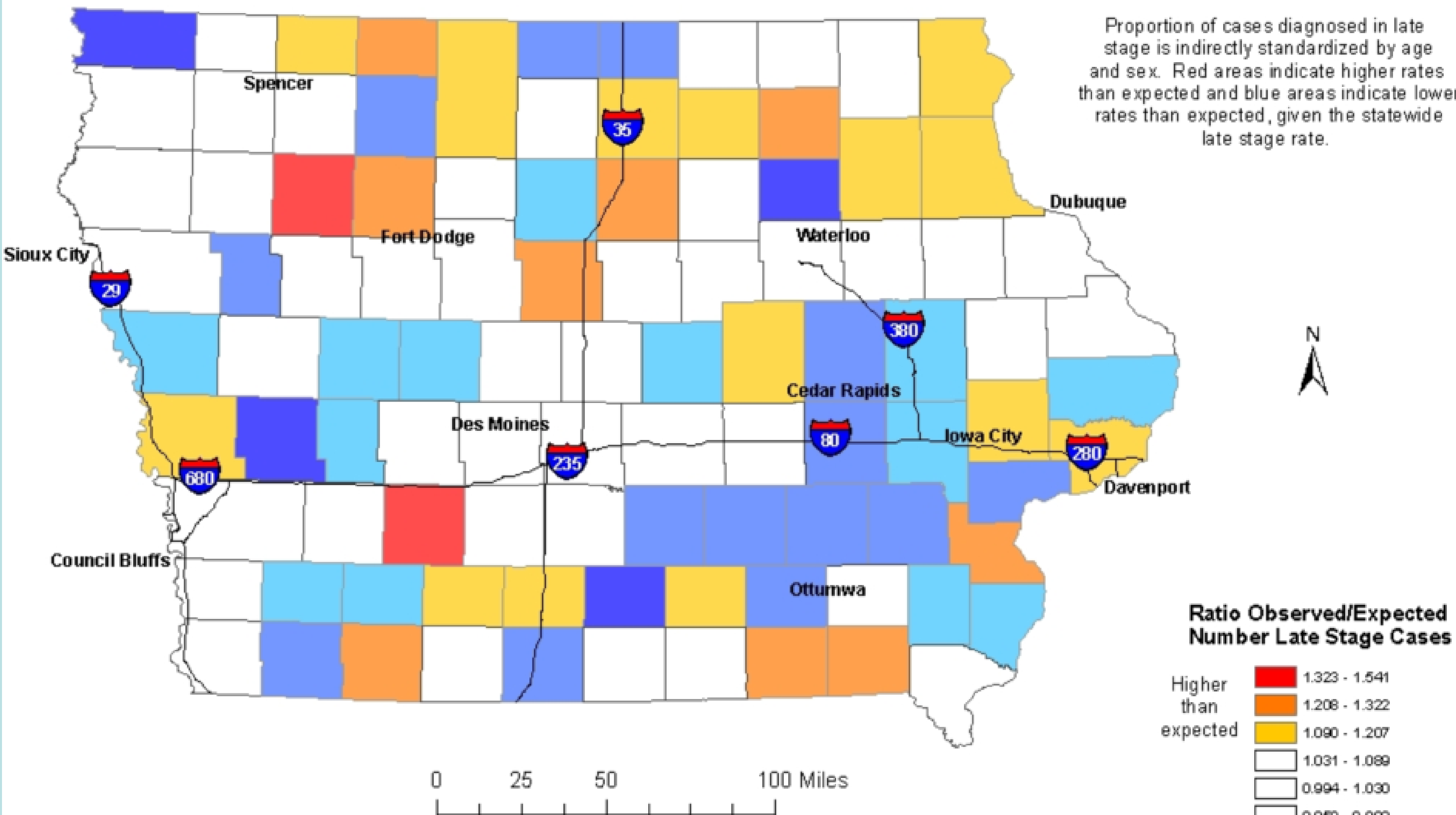
The cases, which both state and federal officials call a disease cluster, are located within a mile of Everett Square — a densely settled neighborhood adjacent to the town’s onetime factory row. It is now home to two Superfund sites.

The study, which was financed by the federal Agency for Toxic Substances and Disease Registry and conducted by state health scientists, will be followed by the creation of a statewide registry to track cases of the disease, formally known as amyotrophic lateral sclerosis, the cause of which is not fully understood.

Geocoded health data brings the opportunity to control the spatial basis of support in maps used in public health

- The tradition in mapping for public health has been to take tables of data, find shape files of area boundaries and make choropleth maps;
- The well-known problem of this approach is that statistics mapped have different degrees of reliability depending on the amount of support available for each of them (the small-number problem);
- The field of spatial epidemiology has developed advanced statistical methods to deal with this problem;
- But spatial epidemiologists have largely neglected the opportunity to use geospatial data and spatial analysis methods that permit them to control the spatial basis of support in maps for public health—see next example.

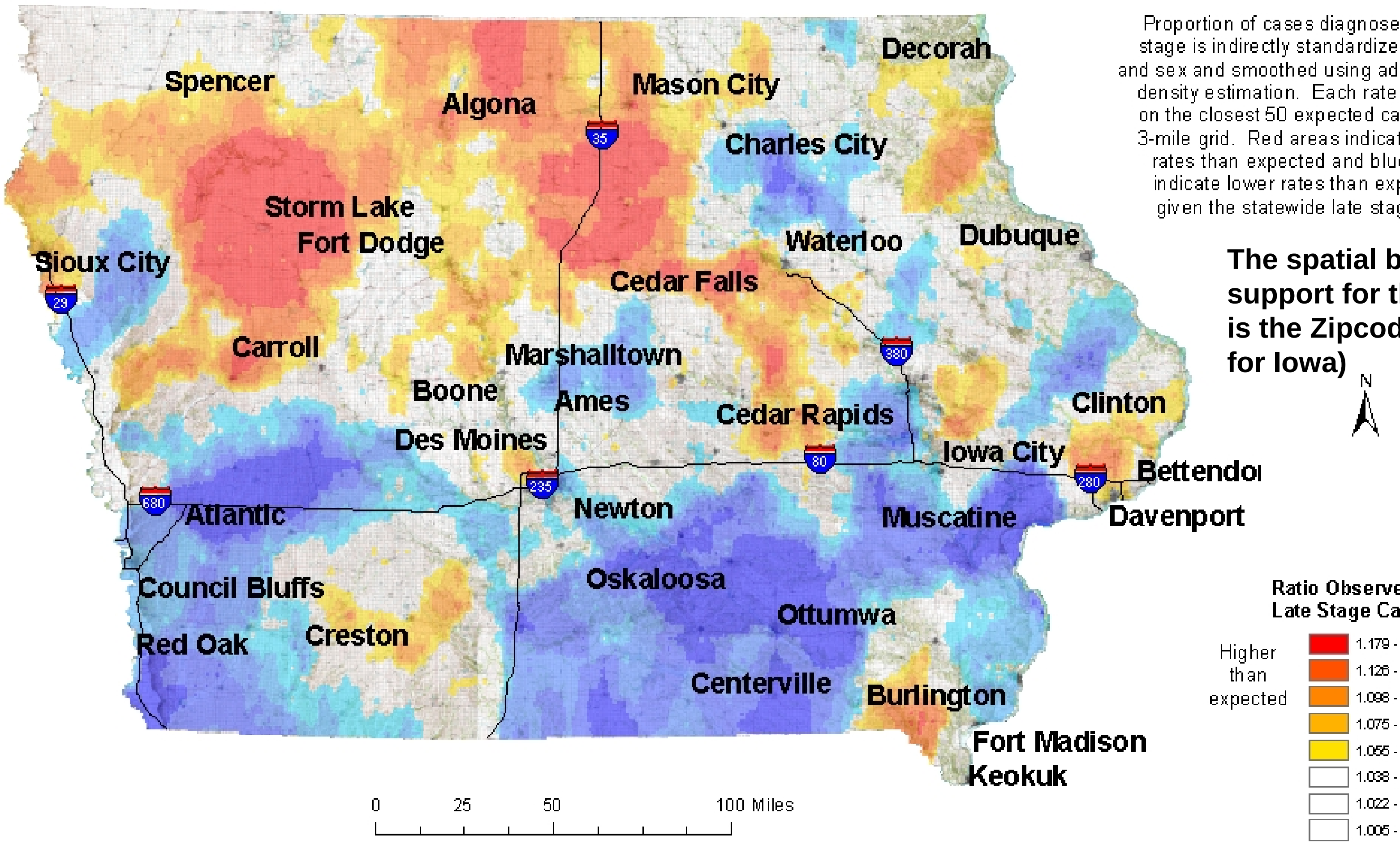
Colorectal Cancer Late Stage Rate, 1998-2003



Data Sources: Cancer data from the Iowa Cancer Registry; County border data from the Natural Resources GIS Library hosted by the Iowa Department of Natural Resources and the US Geological Survey.

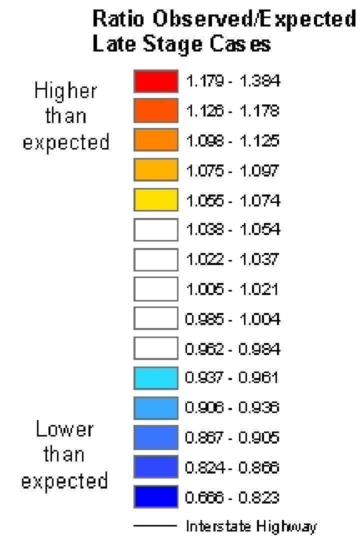
Created By: Kirsten Beyer, Dept. of Geography, University of Iowa, June 2006

Colorectal Cancer Late Stage Rate, 1998-2003



Proportion of cases diagnosed in late stage is indirectly standardized by age and sex and smoothed using adaptive filter density estimation. Each rate is based on the closest 50 expected cases on a 3-mile grid. Red areas indicate higher rates than expected and blue areas indicate lower rates than expected, given the statewide late stage rate.

The spatial basis of support for this map is the Zipcode (940 for Iowa)



Data Sources: Cancer incidence data from the Iowa Cancer Registry; 1:100,000 Digital Raster Graphic from the Iowa Geographic Image Map Server hosted by Iowa State University; Incorporated area locations from the Natural Resources GIS Library hosted by the Iowa Department of Natural Resources and the US Geological Survey.

Created By: Kirsten Beyer, Dept. of Geography, University of Iowa, June 2006

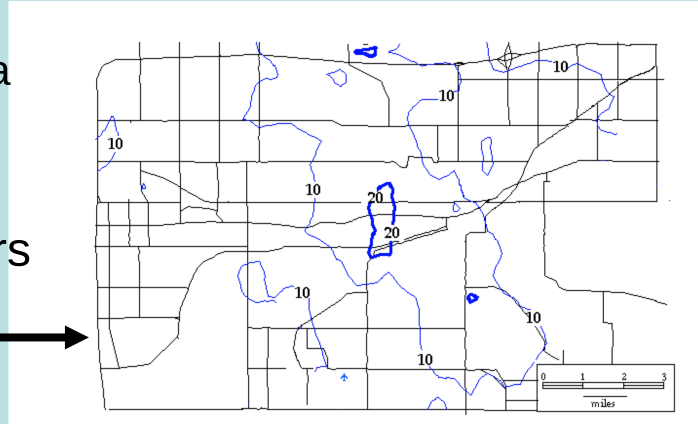
Infant Mortality Rates at Three Different Spatial Scales and Their Approximate Counterparts Using Available Census Administrative Areas

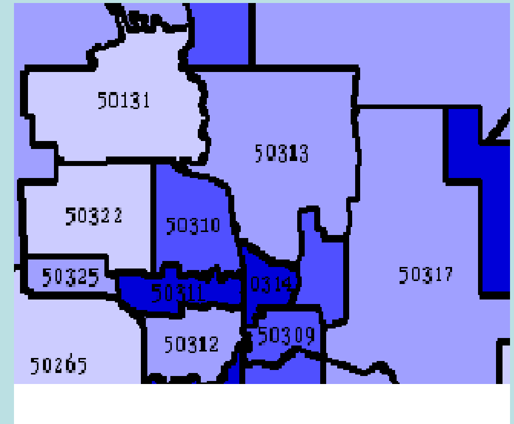
Des Moines, Iowa

1989 - 1992

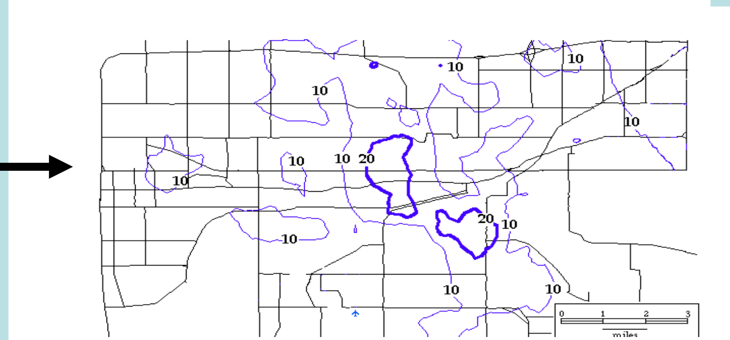
Spatial Filters

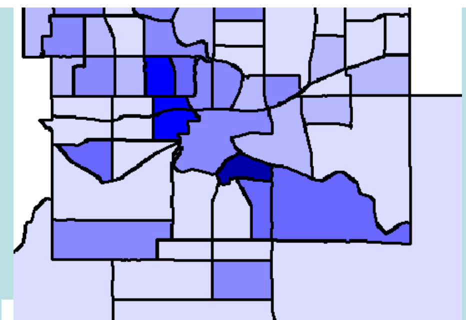
1.2 miles



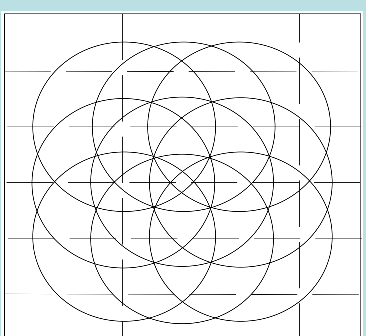


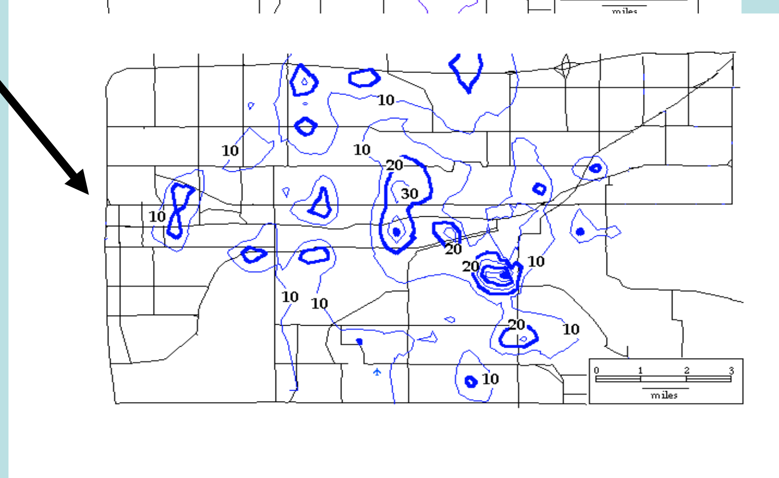
0.8 miles

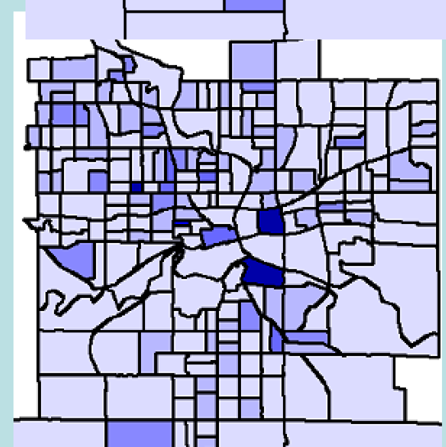




0.4 miles





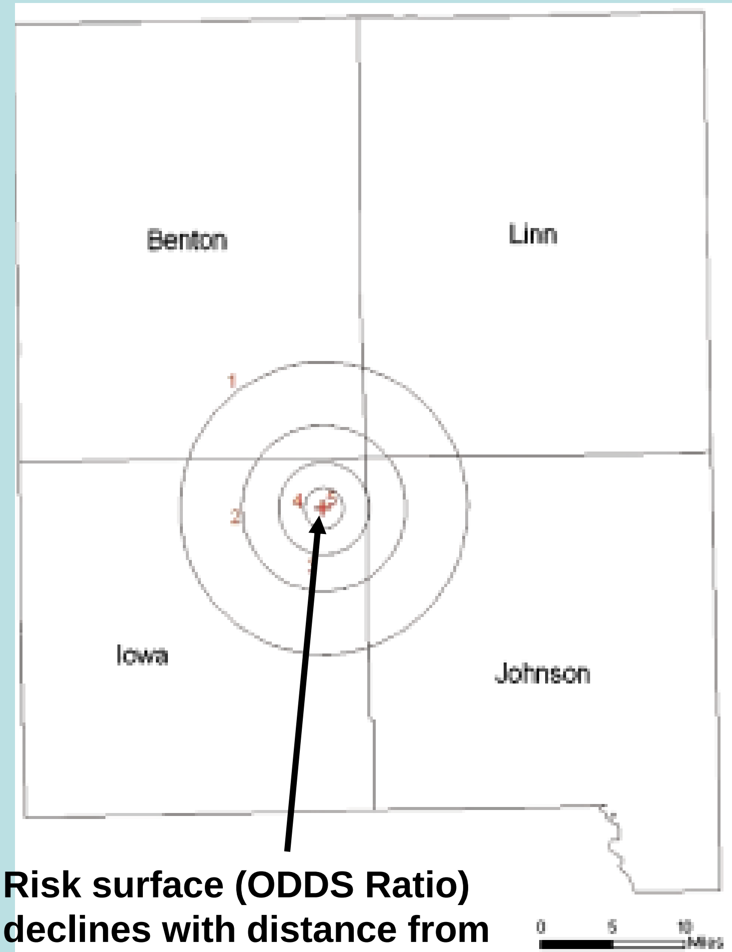


Tests for clustering are weakened when data is aggregated for regions—i.e. when location is treated as a categorical variable.

- Space-time statistics can incorporate covariate information contained in individual records (see Buckeridge et al. 2005, p.105)

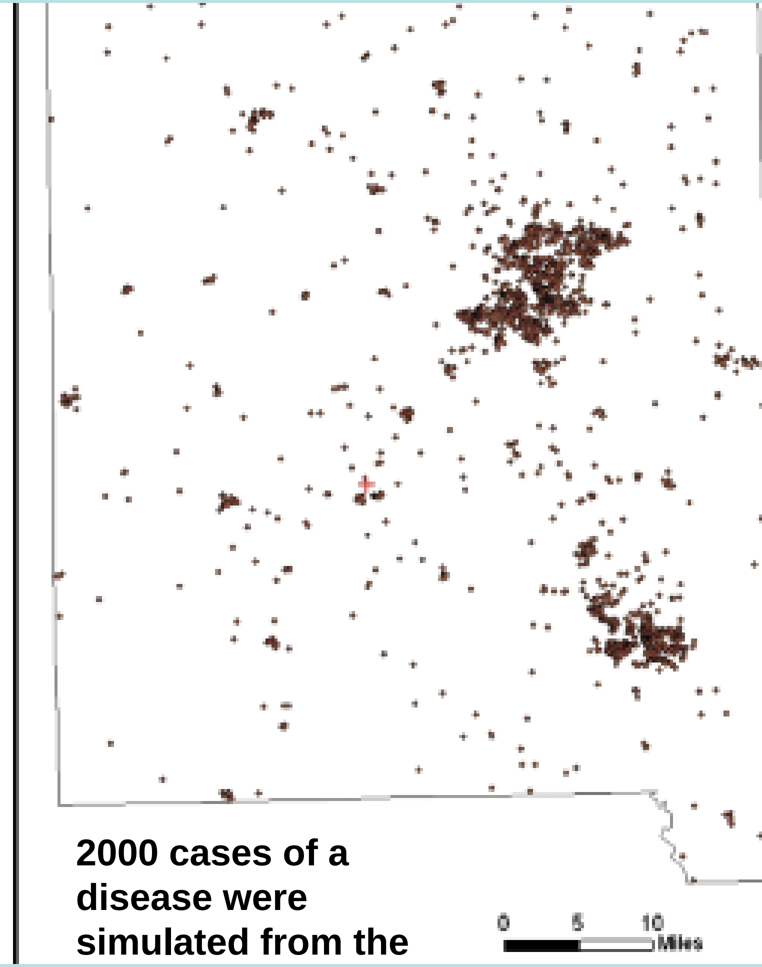
Estimating relative risk from individual case data and disaggregated population data

Hypothetical relative risk of disease



Risk surface (ODDS Ratio) declines with distance from the center

simulated cases of disease

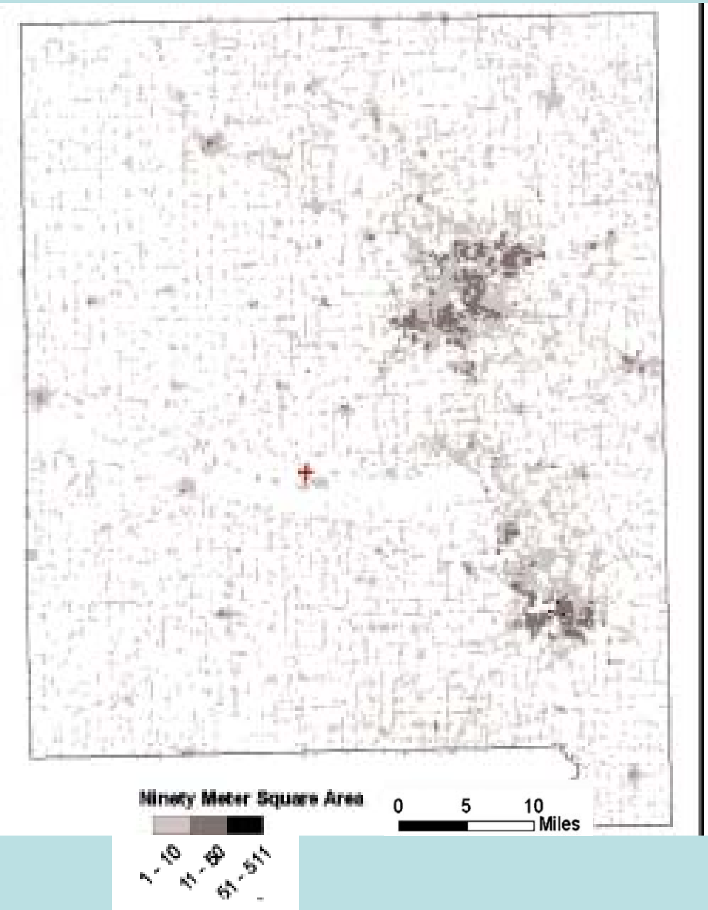


2000 cases of a disease were simulated from the risk surface on the left.

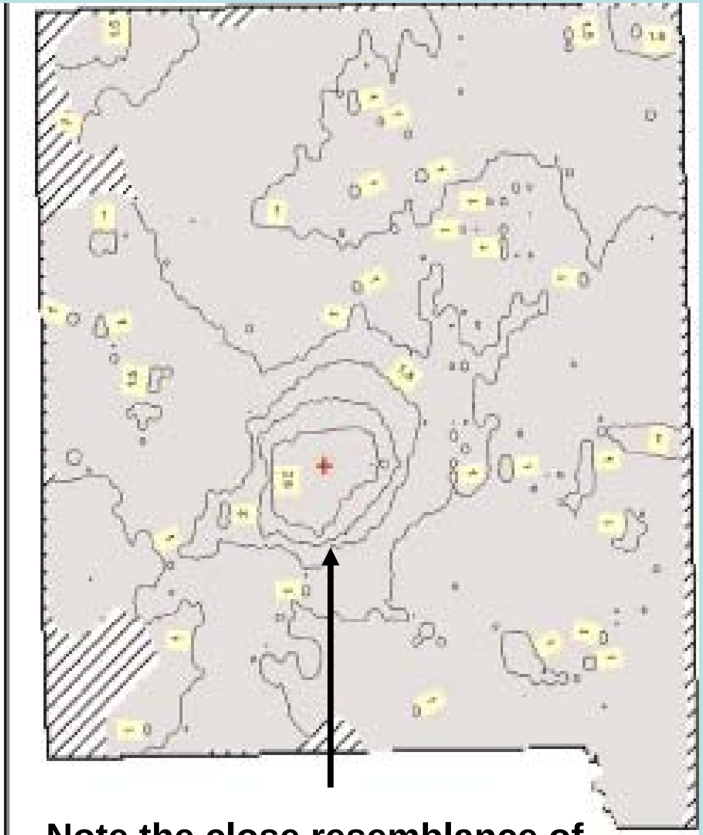
Source: Boulos et al. 2006, p.162.

Estimating relative risk from individual case data and disaggregated population data

Population for 90 meter squares



estimated relative risk—kernel density

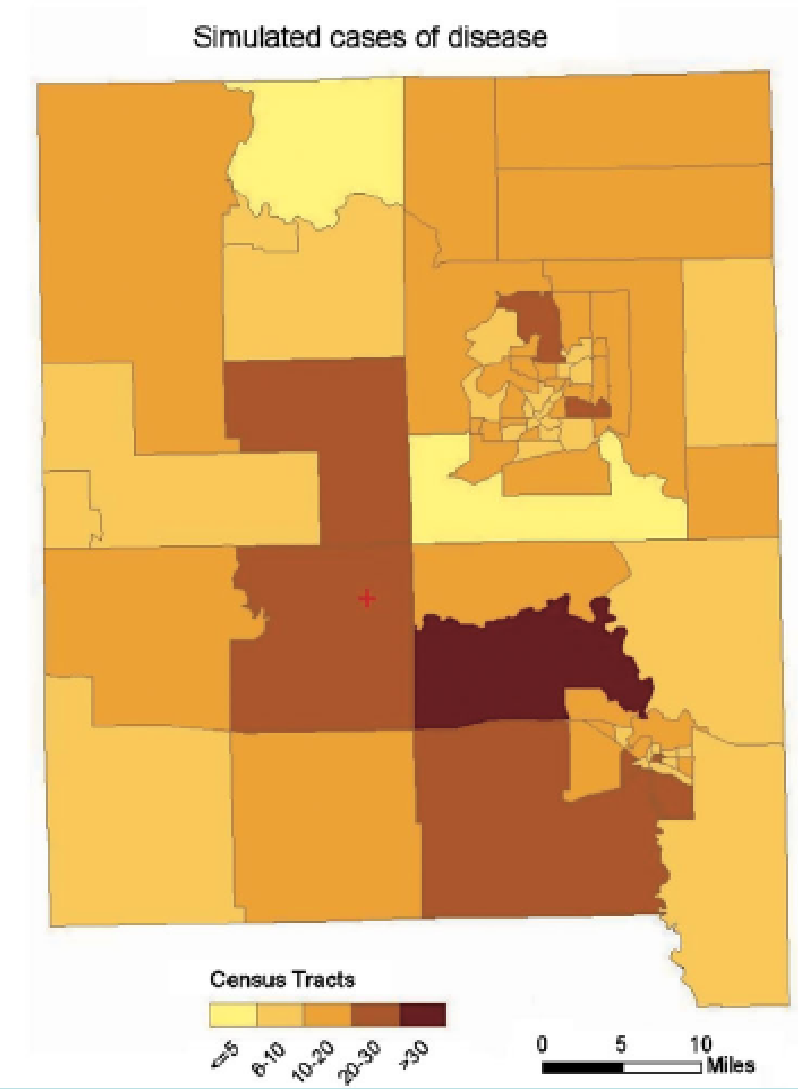


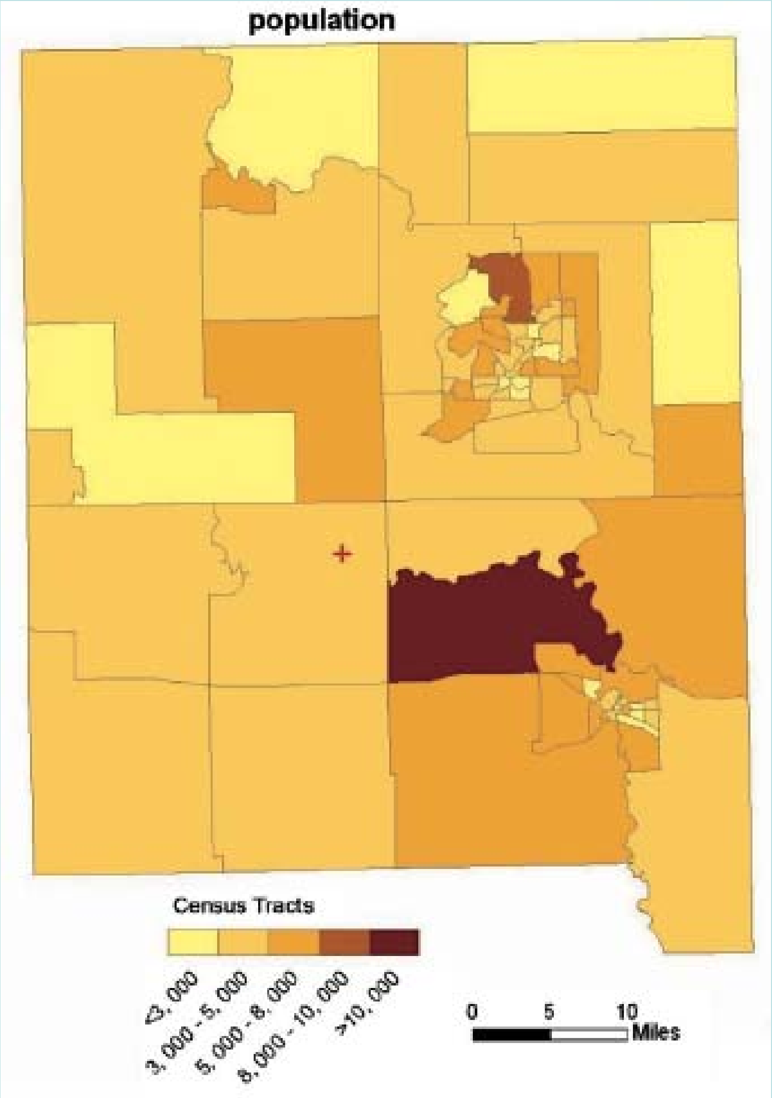
Note the close resemblance of the spatial pattern of recovered estimates of risk with the hypothetical pattern on previous slide

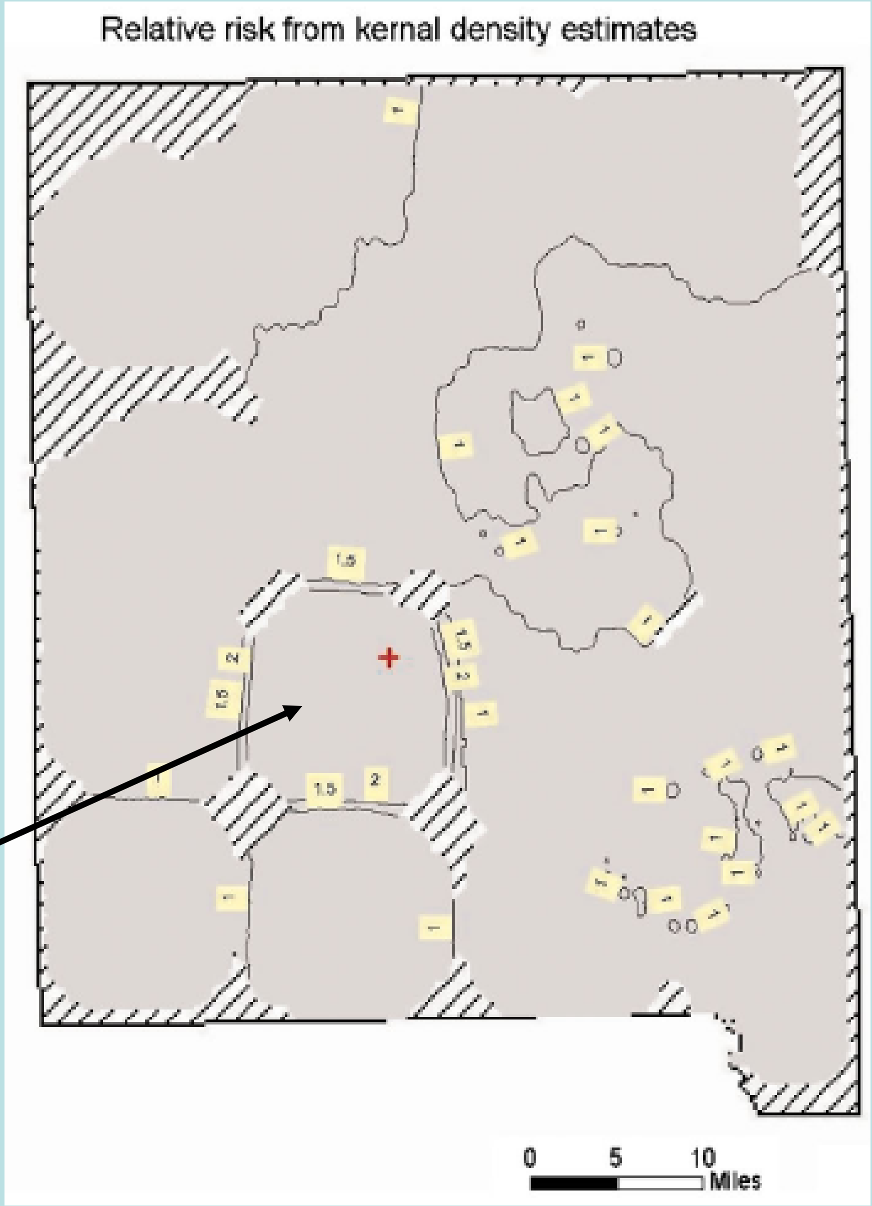
Population data source is LandSCAN USA (Oakridge National Labs. 2005)

Source: Boulos et al. 2006, p.162.

Estimating relative risk from spatially aggregated case (tract) data and census tract population data.







Note that although estimated risks are highest in the vicinity of the area of highest disease risk, the inability to produce the correct spatial pattern of risk is a consequence of the spatial aggregation of the two data sources

The risks to privacy

- “to the extent that data are spatially precise, there is a corresponding increase in the risk of identification of the people or organizations to which the data apply.” (p.1)
- A user “could also discover additional information about the research participant, without asking for it, by linking to geographically coded information from other sources.” (p.2)

Informed Consent

- “Informed consent for the collection or use of personally identifiable information should be obtained “whenever feasible,”

- When not feasible..should be reviewed by some “formal, authoritative, and publicly accountable process.”

Use limitation p.217 waldo

The Health Insurance Portability and Accountability Act of 1996—implemented in 2003

Under this act individual level data may be released if it is “deidentified.”

“Under the statistical deidentification method, a properly qualified statistician using accepted techniques must conclude that “the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information.”

Under the safe harbor method, a covered entity must remove any of 18 individual identifiers from the information, and the covered entity must not have “actual knowledge that the remaining information could be used alone or in combination with other data to identify an individual who is subject of the information.”” see Gittler, 2007, p. 210.

One of the 18 individual identifiers is “all geographic subdivisions smaller than a state, including county, city, street address, precinct, zip code, and their equivalent geocodes.”

For the full list see Box 2 “Individual identifiers under the Privacy rule” at http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

“The rule also allows covered entities to release limited data sets, with more identifiers than deidentified data sets, for public health research, or health care operations.

- But, the rule requires that a covered entity must enter into a data use agreement with the recipient, and that this agreement must enumerate how the data will be used and disclosed and how the data will be protected against impermissible use and disclosure”
- Limited data sets may include “town or city, state, and zip code…”

Gittler, J. 2007. “Cancer registry data and geocoding: privacy, confidentiality, and security issues.” pp.210-211.

Example of a Limited Data Set under the Privacy HIPAA Rule

The Data Access Center at the UCLA Center for Health Policy Research

The Data Access Center, based at the UCLA Center for Health Policy Research, provides researchers with access to confidential data files in a secure, controlled environment that protects the confidentiality of respondents.

The CHIS files available in the Data Access Center contain detailed geographic identifiers and full demographic descriptions for the survey respondents. The files also include highly sensitive information (e.g., sexual behaviors) that has been specifically excluded from the freely available CHIS Public Use Data Files.

Researchers can analyze data remotely by using the programming services of the DAC staff or can write their own programming code and e-mail it to the DAC programmers. Researchers can also schedule time to work on site as guest researchers, where they have access to statistical, programming, and consulting services offered by the DAC.

If you are a researcher who would like to utilize the DAC, please submit an application for review by the CHIS Data Disclosure Review Committee and by the CHIS Principal Investigator.

http://www.chis.ucla.edu/main/default.asp?page=dac

Exceptions are permitted for public health purposes

- P. 218 “In addition, the guidelines recognize that legal requirements from law enforcement or public health agencies sometimes require the release of personally identifiable information without the consent of the individual.”
- “But the exceptions for access in accordance with the law reflects the history of public health in this country, where laws have been passed that recognize the need to violate the privacy of the individual in cases where the health of the general public is put at risk.

Geographic Masking for preserving privacy and confidentiality protection

- Masking methods are methods for concealing locations, or other data associated with locations to protect personal privacy and assure confidentiality.
- Geographic masking is defined as any transformation of spatial data designed to protect the privacy of individuals by concealing their identity from anyone with access to geocoded information about them.
- Interest in geographical masking has increased recently with the widespread availability of fine-grained spatial data that can be used to link geographic coordinates to demographic, social and economic variables in a geographic information system.

Questions asked of masked data

- How effective is the mask at protecting the true locations from discovery by others?
- The degree to which results of spatial analyses on the masked data are the same as results of analyses on the unmasked data.

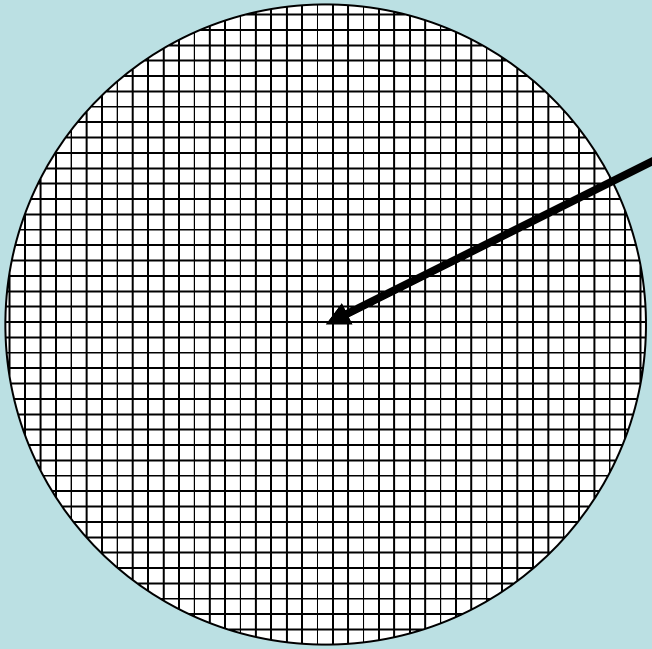
Protecting privacy: spatial masks

1. Spatial data aggregation
2. Coarsened cartographic display
3. Scale, translation and rotation (affine transformations)
4. Random perturbation
5. Spatially adaptive perturbation
6. Attribute perturbation

Random Re-location of Individuals As a Geographic Mask

- This re-location can be controlled by key parameters.
- Knowing these parameters can allow users of geographically masked data to compute the reliability of results of their analyses.
- Re-location parameters can include the constraint that the given total of individuals in a geographic area before masking is the same after masking.
- “Sham” data can be prepared for analysis in protected computer environments.

A Spatial Displacement Geographic Mask

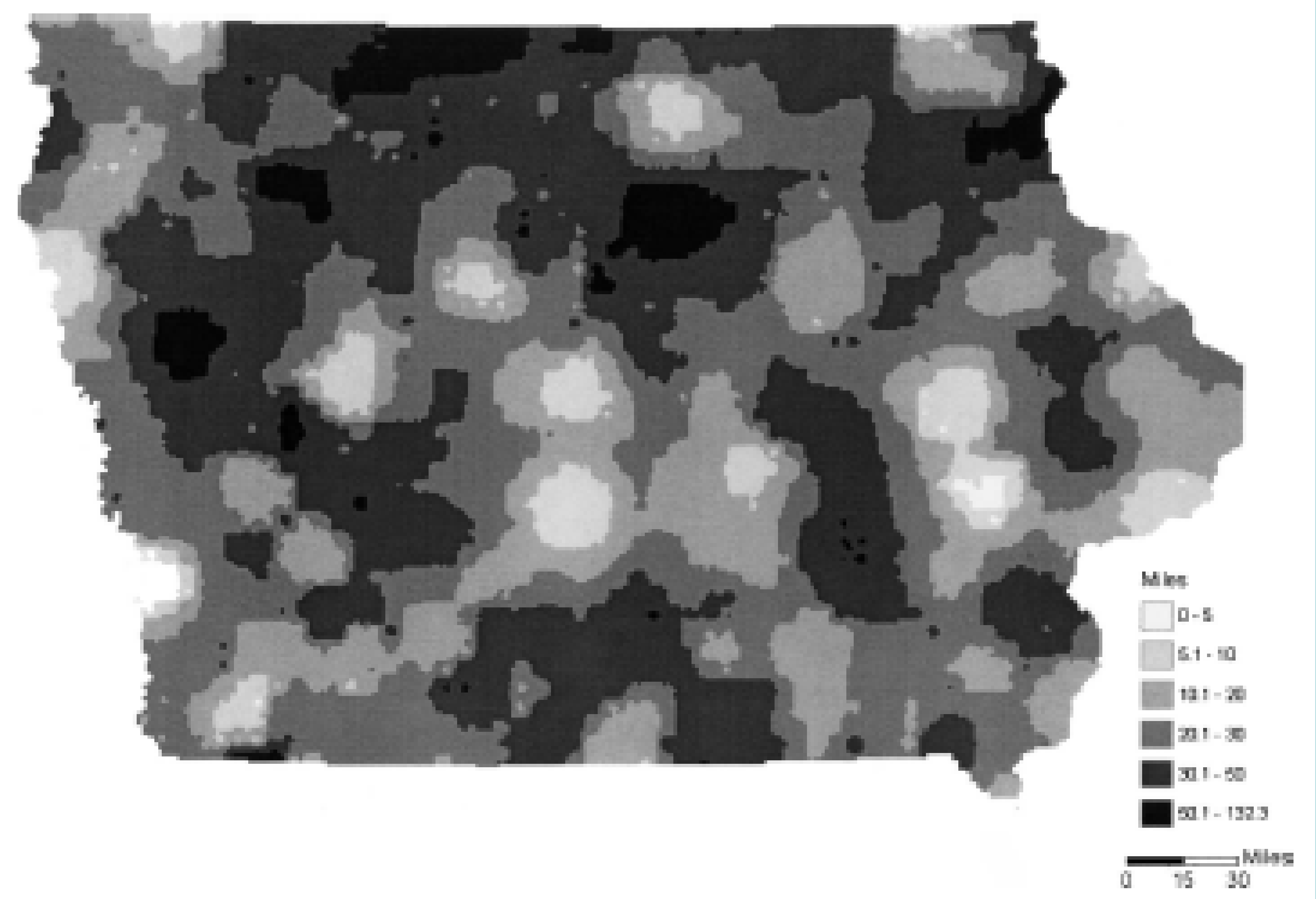


All locations within the circle are equally likely to be selected as the “masked” location for the true location in the center

Spider map: shows place of first diagnosis for colorectal cancer patients in Iowa, 1993-1997

- This map is not shown here.
- The map was withdrawn from publication in *The Journal of Medical Systems* in the proofreading stage.
- See Rushton et al. 2004.
- The locations of the patients were masked by random re-location in their local area but their place of diagnosis was explicit. This violated the promise of confidentiality to most hospitals in Iowa made by the Iowa Cancer Registry since many Iowa hospitals are the only hospital in their town.
- Instead, the following map replaced the spider map.

A different view of the same cancer data: average distance to place of diagnosis for colorectal cancer cases in Iowa, 1993-1997.



Rushton et al. 2004.

Late stage colorectal cancer, Iowa, 1993-1997

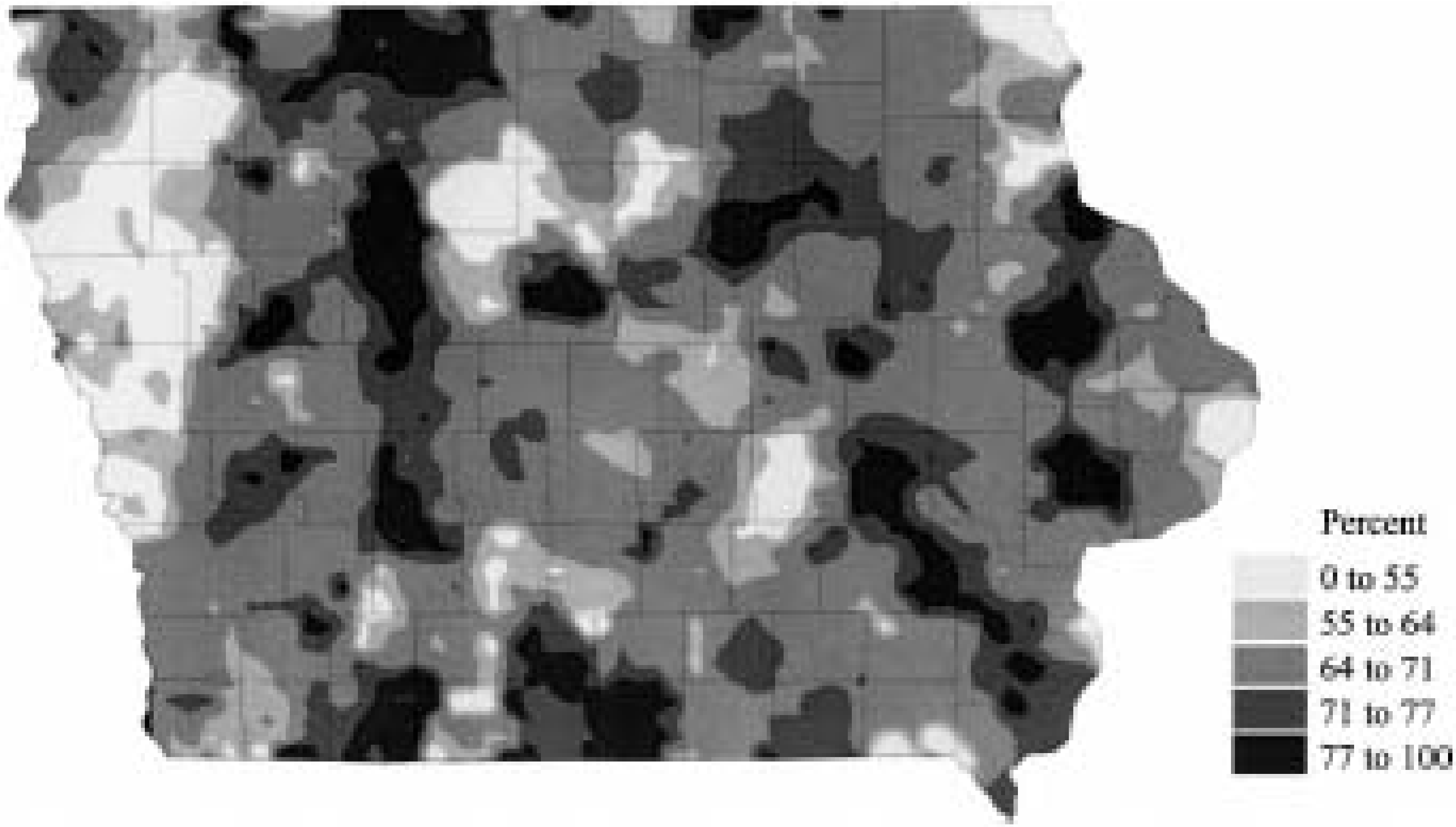
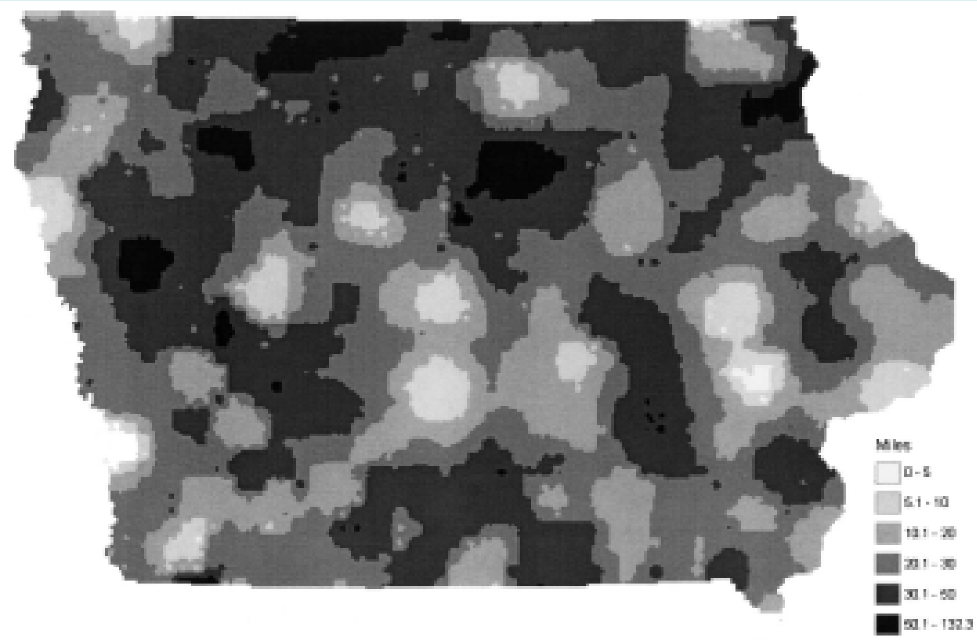


Fig. 3. Late-stage colorectal cancer rates, kernel density estimates, 12.5 mile filters, 1993–1997.

Source: Rushton et al. 2004.



Average distance to place of diagnosis for colorectal cancer cases in Iowa, 1993-1997.

Late stage colorectal cancer, Iowa, 1993-1997

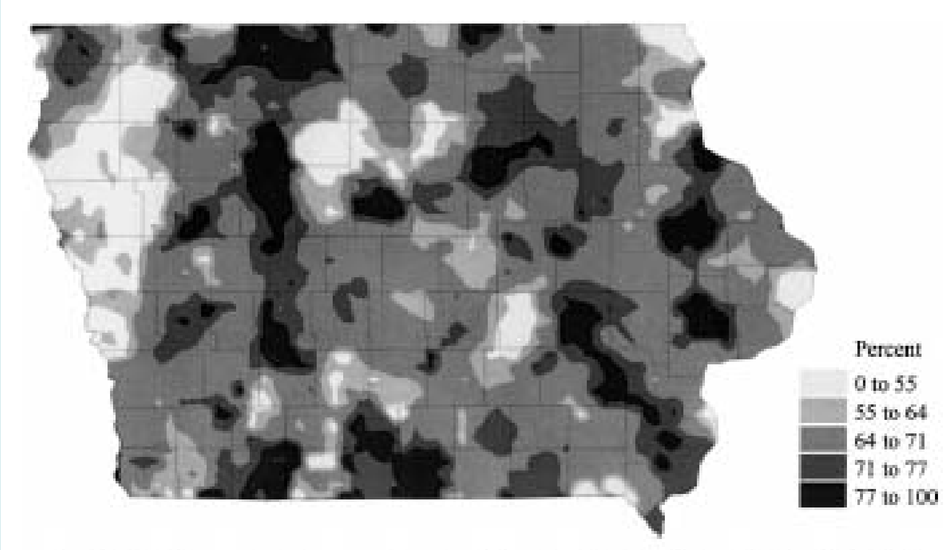


Fig. 3. Late-stage colorectal cancer rates, kernel density estimates, 12.5 mile filters, 1993–1997.

Source: Rushton et al. 2004.

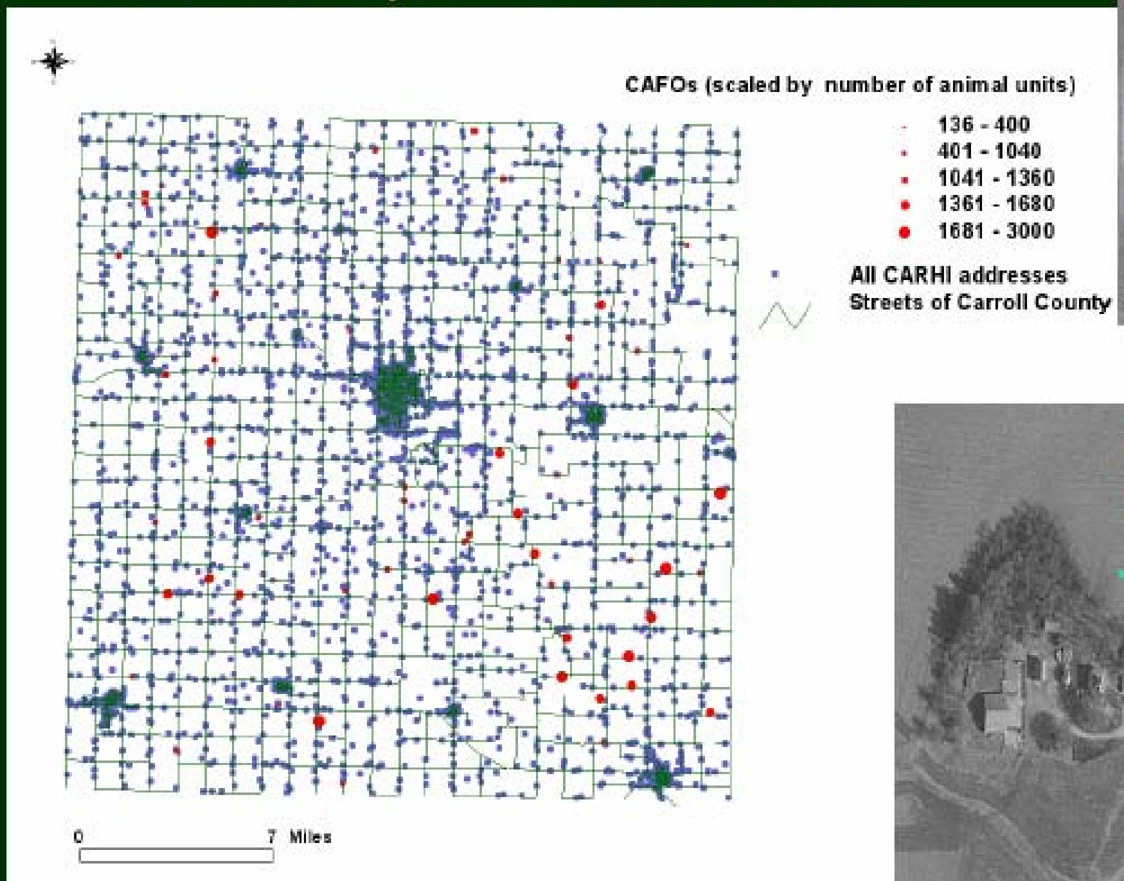
Attribute Masking

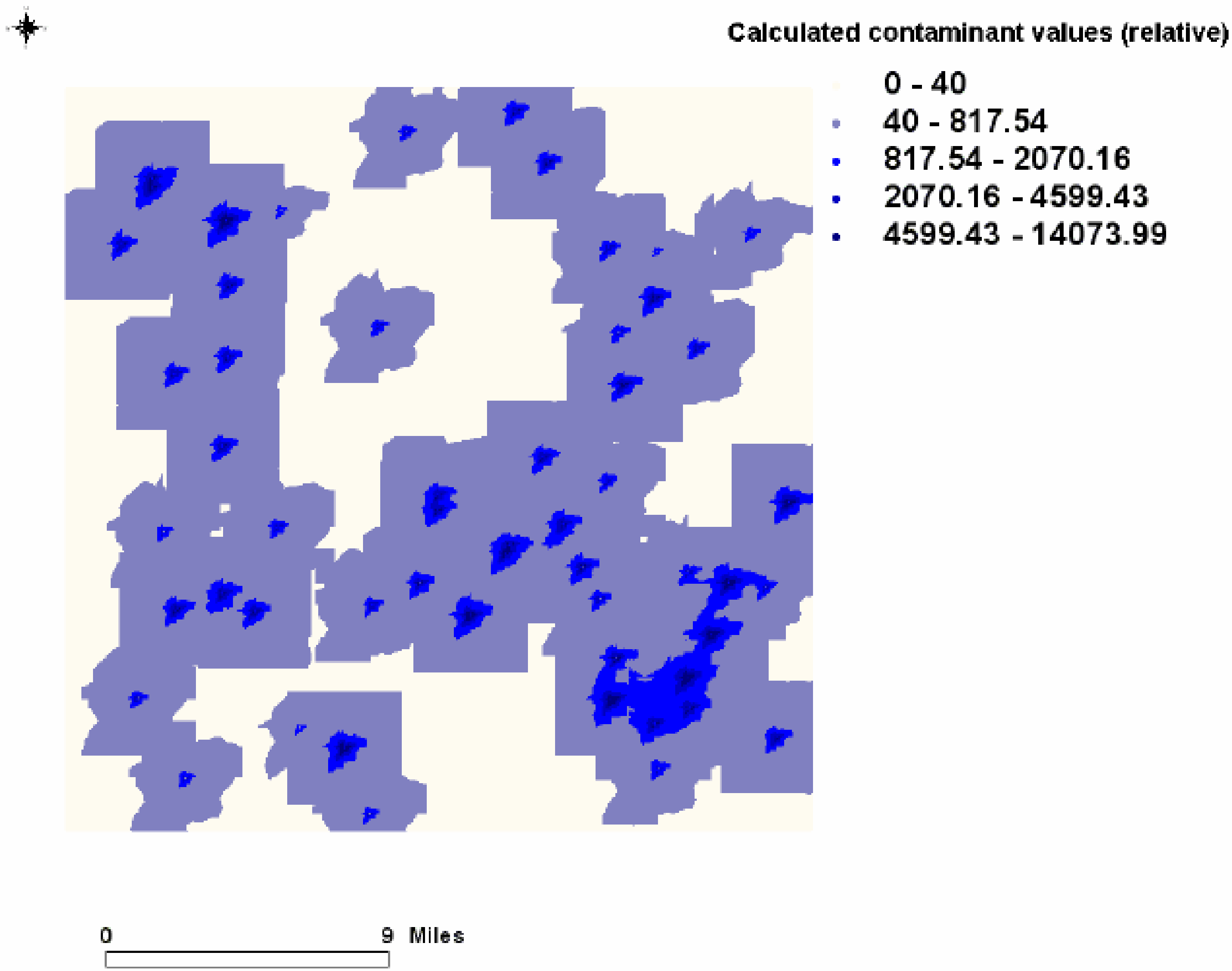
- Knowing the value of an attribute can reveal location in some cases
- This information can then be linked to access other types of personal-level information
- Attributes may require masking as a consequence

2. UI calculates environmental contamination at each addresses.

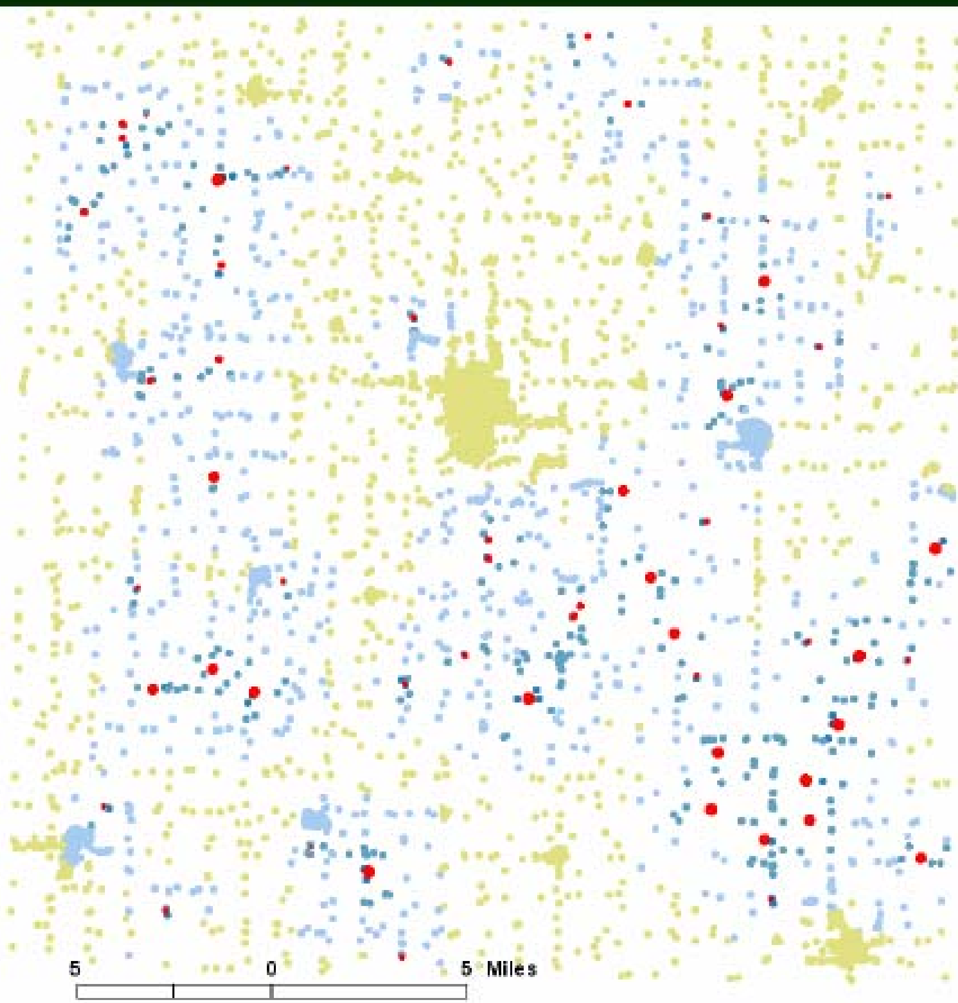
A. UI has 9520 E911 Addresses geocoded to residence locations.

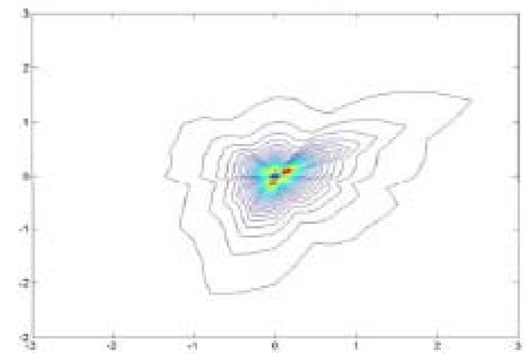
B. UI has 55 permitted CAFO locations.

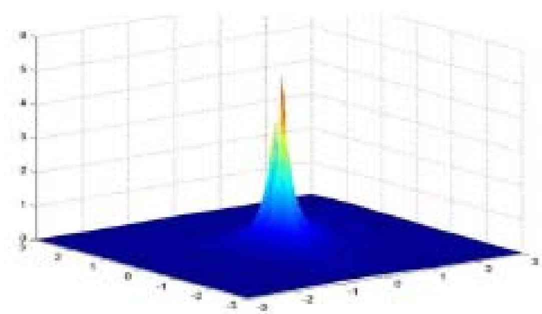



Calculated contaminant values (relative)
0 - 40
40 - 817.54
817.54 - 2070.16
2070.16 - 4599.43
4599.43 - 14073.99
0
9 Miles

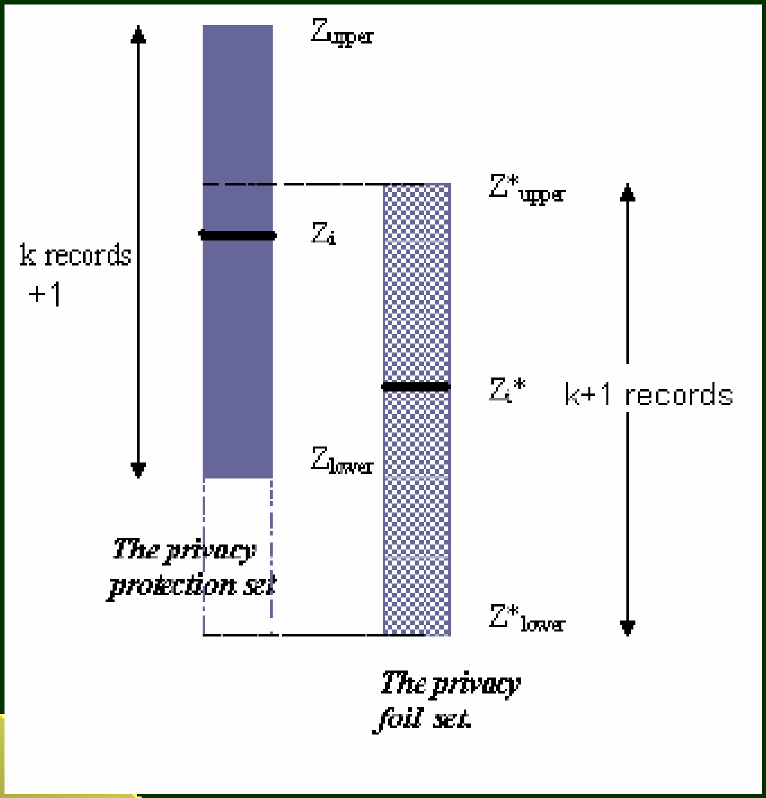
2a. **UI** uses a plume dispersal model and a Spreadsheet program to evaluate contamination at ***<u>each 9520 E-911 addresses</u>***







Masking the i'th record :



90	23.37285798	204 W 14TH ST
91	23.41566403	1700 PIKE
92	23.42193538	1515 N MAIN ST
93	23.47834792	1609 PIKE AVE
94	23.54193972	1521 N MAIN
95	23.54691033	1502 N ADAMS
96	23.60499445	1508 N ADAMS
97	23.60884482	1417 NORTH ADAM
98	23.6509593	1621 PIKE AVE
99	23.72118269	1801 NORTH MAIN
100	23.72948171	209 WEST 15TH ST
101	23.72948171	208 E 16TH ST
102	23.76547126	1705 PIKE
103	23.76547126	1704 PIKE AVE
104	23.78563435	1516 N ADAMS
105	23.78825538	1505 N ADAMS
106	23.82821962	1701 PIKE AVE
107	23.82821962	1701 PIKE
108	23.84904901	1513 NORTH ADAM
109	23.84904901	1507 N ADAMS
110	23.85005133	221 W 15TH

Masking the i'th record :

Z_i = 23.72946171
Z_i*= masked contamination
k = 11 – 1
τ = A random number between 0 and 1.

Z_{upper}
Z_{lower}

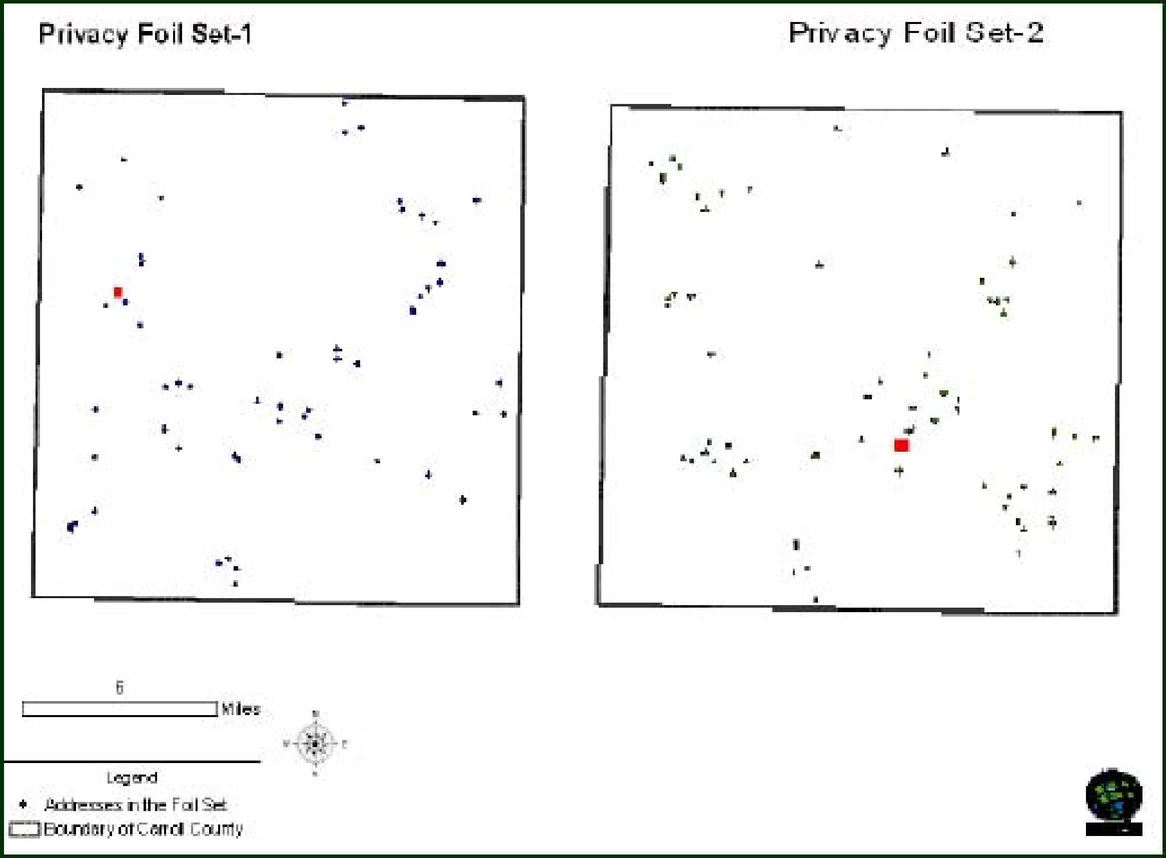
90	23.37285798	204 W 14TH ST
91	23.41566403	1700 PIKE
92	23.42193538	1515 N MAIN ST
93	23.47834792	1609 PIKE AVE
94	23.54193972	1521 N MAIN
95	23.54691033	1502 N ADAMS
96	23.60499445	1508 N ADAMS
97	23.60884482	1417 NORTH ADAM
98	23.6509593	1621 PIKE AVE
99	23.72118269	1801 NORTH MAIN
100	23.72948171	209 WEST 15TH ST
101	23.72946171	208 E 15TH ST
102	23.76547126	1705 PIKE
103	23.76547126	1704 PIKE AVE
104	23.78563435	1516 N ADAMS
105	23.78825538	1505 N ADAMS
106	23.82821962	1701 PIKE AVE
107	23.82821962	1701 PIKE
108	23.84904901	1513 NORTH ADAM
109	23.84904901	1507 N ADAMS
110	23.85005133	221 W 15TH

- Spruill, (1982) suggests a measure of disclosure risk based on a criterion involving the relationship of the masked value to the original value. Spruill's measure of disclosure risk is the proportion of records in the masked dataset that are closer to the parent records than any other records in the original dataset

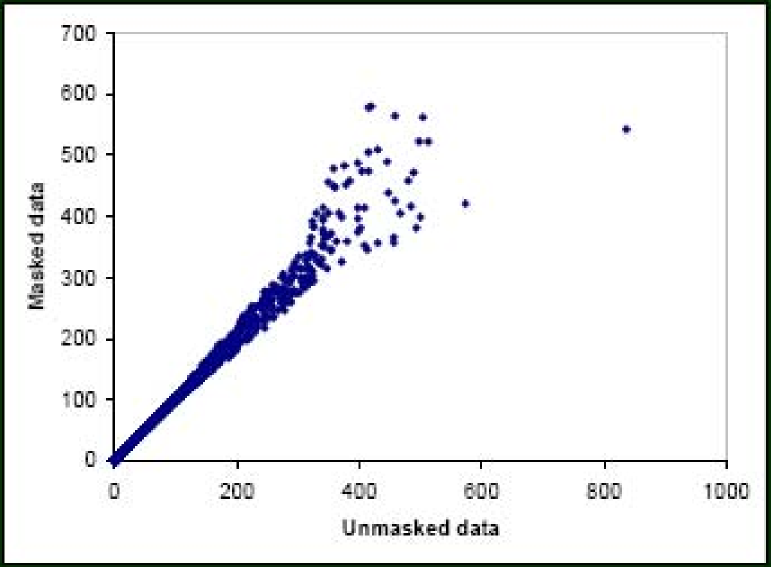
Example privacy Foil Sets (mask size=81)

Foil set 1 :
For a contamination
Value of 235.23

Foil set 2:
For a contamination
Value of 579.18



Performance of the masked data



Mask performance is not too desirable at the boundaries, if the boundary records are excluded, the correlation coefficient increases to 0.993.

Mask Metadata: The pros and cons of concealing details of the mask

- The more detail revealed about the mask, the more a person dedicated to identifying individuals may succeed in defeating its purpose.
- However, knowing details of the mask permits sensitivity analyses that will inform a researcher whether their conclusions using the masked data are likely to be valid.

Alternatives to geographical masking: Institutional arrangements

1. Data enclaves where access and use of individual data is monitored and only selected individuals are allowed entry.
2. Data sharing agreements which are contractual agreements between willing partners where users promise not to engage in data linkages that might lead to the identification of records.

    Example of a data use agreement at:

    http://researchcompliance.uc.edu/hipaa/DataUseAgreement.pdf

Alternative # 3: Agent-based computer access to selected spatial analyses

- The holder of individual information (data steward) makes available, usually on a website, a secure server which is capable of providing selected spatial analyses of their data.
- The individual-level data resides at the server-side (on the computer that hosts the data), rather than on the client side (your personal computer)
- Their server is organized not to allow any original data to be returned to the enquirer; instead, it returns results of analyses which do not include any individually identifiable information either directly or by inference by connecting the results to any other information in the possession of the enquirer; (see Boulos et al. 2006).

Conclusions

- More knowledge is needed about the effectiveness of masking techniques to protect confidentiality and to provide valid analysis results.
- For many purposes, small area spatial data can support the same conclusions as when using individual data.
- Agent-based access to individual data on secure servers is likely to prove an effective and efficient solution to providing access to individual data and protecting confidentiality of data.
- Education and training in ethical use of spatially precise health data linked to social data.

References

Armstrong, M.P., G. Rushton, and D.L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. Statistics in Medicine 18:497-525.

Boulos, K.M.N., Q. Cai, J.A. Padget, and G. Rushton. 2006. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. Jn. of Biomedical Informatics. 39:160-170.

Brownstein, J.S., C.A. Cassa and K.D. Mandl (2006) No place to hide—reverse identification of patients from published maps. New England Jn. of Medicine 355(16):1741-1742.

Buckeridge, D.L. et al. 2005. Algorithms for rapid outbreak detection: a research synthesis. Jn. of Biomedical Informatics 38:99-113.

Centers for Disease Control. 2003. HIPAA Privacy Rule and Public Health: Guidance from CDC and the U.S. Department of Health and Human Services. MMWR 52:1-12 (April 11, 2003).

Gittler, J. 2007.(in press) Cancer Registry Data and Geocoding: privacy, confidentiality, and security issues. Ch. 12 (pp.195-223) in G. Rushton et al. Geocoding Health Data: The use of geographic codes in cancer prevention and control, research and practice. Boca Raton, Fla. CRC Press.

Kwan, M.-P., I. Casas, and B.C. Schmitz. 2004. Protection of geoprivacy and accuracy of spatial inforamtion: how effective are geographical masks? Cartographica 39:15-28.

Leitner, M., A. Curtis. 2004. Cartographic guidelines for geographically masking the locations of confidential point data. Cartographic Perspectives 49:22-39.

National Research Council. 2007. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, M.P. Gutmann and P.C. Stern, Eds. Washington, D.C.: The National Academies Press.

Rogerson, P.A. 1997. Surveillance systems for monitoring the development of spatial patterns. Statistics in Medicine 16:2081-2093.

Rogerson, P.A. 2001. Monitoring point patterns for the development of space-time clusters. Jn. Royal Statistical Society Ser. A 164:87-96.

Rushton, G., Peleg, I., Banerjee, A. Smith, G., West, M. 2004. Analyzing geographic patterns of disease incidence: rates of late-stage colorectal cancer in Iowa. Journal of Medical Systems 28:223-236.

Rushton, G. et al. 2006. Geocoding in cancer research: a review. American Jn. Preventive Med. 30(2S):S16-S24.

Spruill, N.L. 1983. The confidentiality and analytic usefulness of masked business micro-data. Proceedings of the Section on Survey Research Methods, 602-607.

Waldo, J., Lin, H.S. and Millett, L.I. 2007. Engaging Privacy and Information technology in a Digital Age. National Research Council, The National Academies Press, Washington, DC.

Zimmerman, D.L., M.P. Armstrong and G. Rushton. 2007 (in press). Alternative techniques for masking geographic detail to protect privacy. Chapter 7 (pp. 127-138) in G. Rushton et al. Geocoding Health Data: The use of geographic codes in cancer prevention and control, research and practice. Boca Raton, Fla. CRC Press.

Selected References

[Applied Spatial Statistics for Public Health Data] by Lance A. Waller
[Spatial Epidemiology: Methods and Applications] by P. Elliott

GIS and Public Health

by [Ellen K. Cromley], [Sara L. McLafferty]

[Introduction to Geographic Information Systems for Public Health]
by Alan L., M.D. Melnick

Acknowledgments

- University of Iowa Research Assistants: Kirsten Beyer, Qiang Cai, Zunqiu Chen, Soumya Mazumdar, Chetan Tiwari.
- University of Iowa Colleagues: Marc Armstrong, Josy Gittler, Dale Zimmerman.

Thank You!

This slide presentation can be viewed at:

www.uiowa.edu/~gishlth/ScottsdaleESRI/

After October 12, 2007