Preserving the Privacy of Health Records while Testing Hypotheses of Relationships between Health Outcomes and Point-Based Sources of Pollution

Gerard Rushton
University of Iowa
Department of Geography
316 Jessup Hall
Iowa City, IA 52242
Gerard-rushton@uiowa.edu
319 335-0162
319 335-2725 (FAX)

Soumya Mazumdar
University of Iowa
Department of Geography
316 Jessup Hall
Iowa City, IA 52242
Soumya-mazumdar@uiowa.edu
319 335-0150
319 335-2725 (FAX)

Abstract

For the purposes of preserving privacy and the confidentiality of records, health registries release spatial data to the public only after a process of "de-identification" There are different methods of accomplishing this but all involve loss of geographic detail on the locations of people. We describe a project in which the geographic detail of health records were preserved for the essential purpose of testing hypotheses about the relationship of health outcomes and exposures to sources of potential environmental pollution without putting analysts in the position where they could know or could infer the identity of people. The method involves detailed geocoding of residential addresses in one Iowa County, estimating environmental contaminant values at all residential locations in the county from one source of possible environmental pollution, and an algorithm that ensures that at least forty people in the county have contaminant values within the range of values in the adaptive, non-contiguous, spatial mask.

1.0 Introduction

The requirements of spatial analyses of health data often conflict with confidentiality requirements which holders of health data must meet to ensure the privacy of health records. Guidelines were issued in April 2003 for implementing the 1996 HIPAA law on the protection of health records. This law also applies to researchers who use individually identifiable health data (U.S. Health and Human Services, 2005; U.S. Health Resources Administration, 2005). Location, of course, is a means of identifying a person and anyone working in the area of GIS and health is likely to encounter the problem of acquiring and using location information about individuals as well as the problem of ensuring that results of their research do not convey the health information of individuals. The common method of protecting confidential information is to de-identify it which normally has meant to strip detailed location information from individual records and to either replace this location detail with some location entity covering a larger region, such as a census tract or a county, or to change the location information by some randomizing procedure. Both approaches can be described as a geographic masking process (Armstrong et al. 1999).

The context of our work is an environmental health tracking project supported by a grant to the Iowa Department of Public Health (IDPH) from the Centers for Disease Control and Prevention with a sub-contract from IDPH to The University of Iowa. The purpose of this project is to provide proof that it is possible to investigate relationships between the exposures of people to possible environmental contaminants and their health. The study area is Carroll County, Iowa, an approximately 24 by 24 mile area in West-Central Iowa. The overall design of the project is to identify all health problems presented to primary care doctors for patients living in the study area as recorded in the patient billing codes of the doctors, to identify and measure (estimate) the values of possible environmental contaminants at any location, and to analyze relationships, using statistical models, to determine whether the likelihood that a person presents a doctor with a particular health problem is related to their degree of exposure to a particular contaminant.

Individually identifiable health data is available to IDPH under an agreement between themselves and the doctors in the area. Under this agreement, only IDPH personnel have access to the individually identifiable health information and the information may only be used for purposes of the project. The approval of the University of Iowa (UI) Human Subjects Institutional Review Board for the University of Iowa component of this project explicitly states that no UI personnel may have access to individually identifiable health information. This approval was gained only after considerable discussion with University lawyers concerned about University liability issues under the new HIPAA guidelines described above. The interesting GIS problem this raised was one of complying with this agreement and yet ensuring the ability of the project to meet its goals of testing relationships between environmental exposures and the health of individuals.

1.1 The Overall Strategy for Ensuring Privacy: Develop a Carroll County De-identified Health Encounter Data Set.

The overall strategy to accomplish this goal, described in further detail in later sections of this paper, is as follows. The Iowa Department of Public Health (IDPH), the custodians of the health encounter data, passes to UI a data file consisting of addresses for which they have health encounter data and an ID (ID #1). UI attaches the geocode of each address by matching the addresses provided by IDPH to the rural and urban addresses in the county for which it has found geocodes—see section 1.2 below. UI then attaches measures of exposures to possible environmental contaminants for the geocodes in question from its GIS data layers of estimates of environmental contaminants. IDPH applies a masking function to the exposure estimates provided by UI, uses their ID #1 to link their health encounter data to the UI data and then returns the entire health encounter data set to UI. The dataset returned to UI has a new set of IDPH IDs (ID #2); it no longer has ID #1; it has no geocodes; it does not have the address of the person, nor does it have any information about the individual that could lead to their identification. Date of birth, for example, is an example of information that could lead to a person's identification. The locational identifier at this point in the process is "this person lives in Carroll County, was seen by one of the participating primary health physicians, the presenting health problem was such and such (ICD Code #), and the masked environmental contaminant measures for the residence of this person have these values." ID #2 refers to a particular individual in a particular household. UI does not know the location of this person nor does it have any information from which it could identify them. However, because a masked value of UIs estimate of environmental contaminant exposure is attached to the record, it is possible, using this data set, to investigate relationships between environmental contaminants and health status of individuals in the area. Moreover, it is possible for UI personnel to collaborate with IDPH personnel in this process. IDPH personnel have access to the locations (geocodes) of individuals in the County; UI personnel do not. Using its knowledge of the geocodes of persons in its health encounter dataset, IDPH is able, in separate work in which UI would not be involved, to fulfill its constitutional obligations to protect the health of the public by following up results of these investigations with appropriate public health interventions or with further studies.

The remainder of this paper describes the geocoding process, the measurement of possible environmental contamination for one source of pollution, and the masking of the environmental pollution measure to protect the privacy of the health data.

1.2 Geocoding Addresses in the County

In 2004, Carroll County completed its implementation of an E911 addressing system. As part of this process it produced a geocode for each address in the county. The county's definition of the location of an address was to define a point which would most likely help a person responding to a telephone call asking for emergency service to find the person who had requested the service. This definition was implemented as the coordinate location of the place where a person would leave the public road and join the private road leading to the property from which the call was made. Figure 1 illustrates this definition.

It shows two locations for an address. The location at the road junction is the E911 location and the location at the site of the residence of the property identified at this address as defined for this project. In this case, the two geocodes are not far apart, though in a sample rural area of the county we found that the mean distance between the two geocodes was 500 feet. The difference in environmental exposure measures between locations 500 feet apart could be considerable; hence the need to define geocodes in relation to the locations of people rather than the locations that are useful for finding people in the case of emergencies. In the case of most addresses in this county, the residence property was unambiguously identifiable on the orthophoto map. In the few cases where several properties were located on the private road, tax assessor records with geocoded property lines were used to identify the residential structures that related to the addresses in question. All, approximately 3,500 rural addresses, received dual location codes (geocodes) through this process. To find the coordinates of these locations, the county GIS group used an orthophoto coverage, dated April 2001, which was available for the County at a spatial resolution of approximately 2.5 feet per pixel.

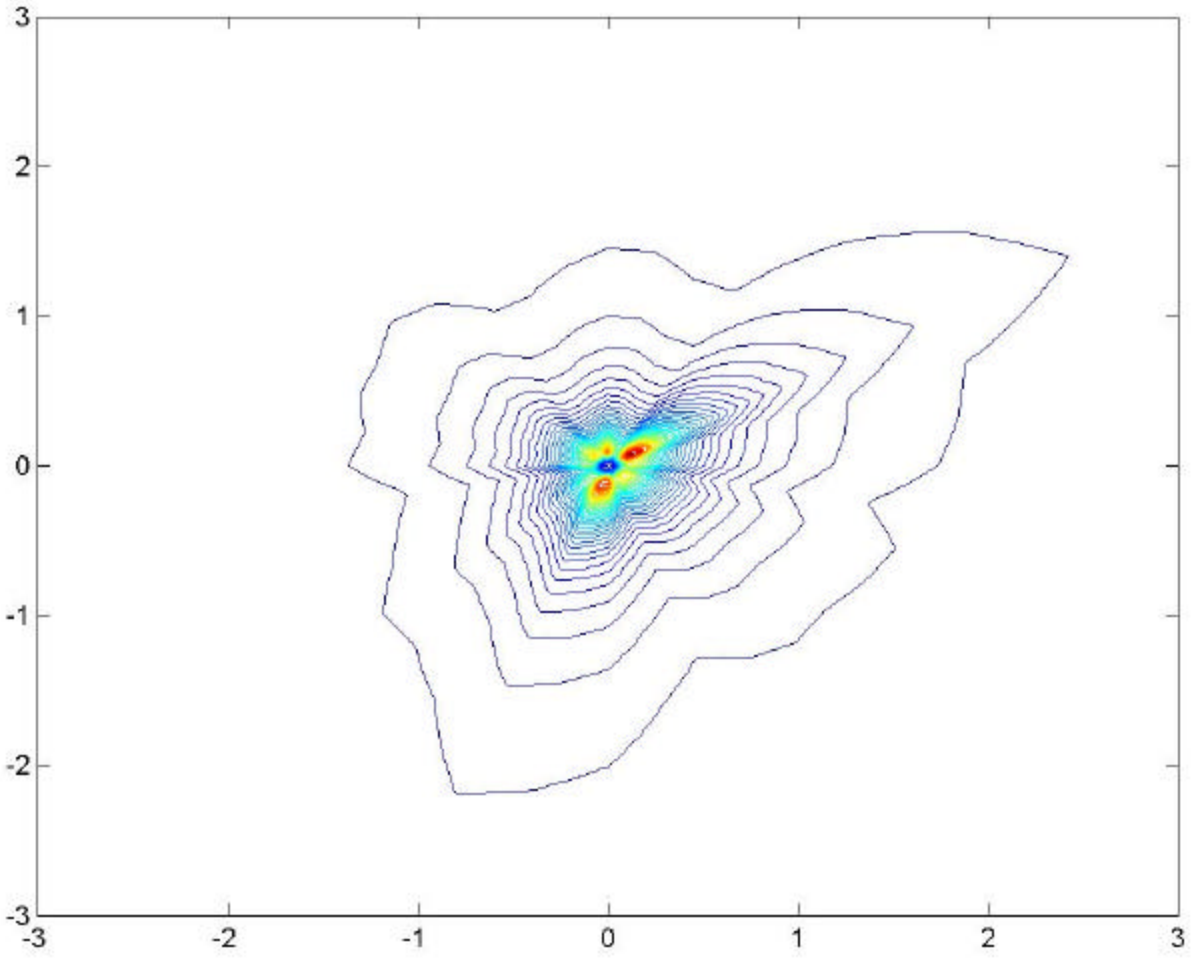
Figure 1. Dual encoding of rural address locations used in the project

1.3 Estimating Possible Contaminant Values at all Locations

We illustrate our approach to measuring possible airborne contamination with the case of releases into the air of particulates from confined animal feedlot operations (CAFOs) that

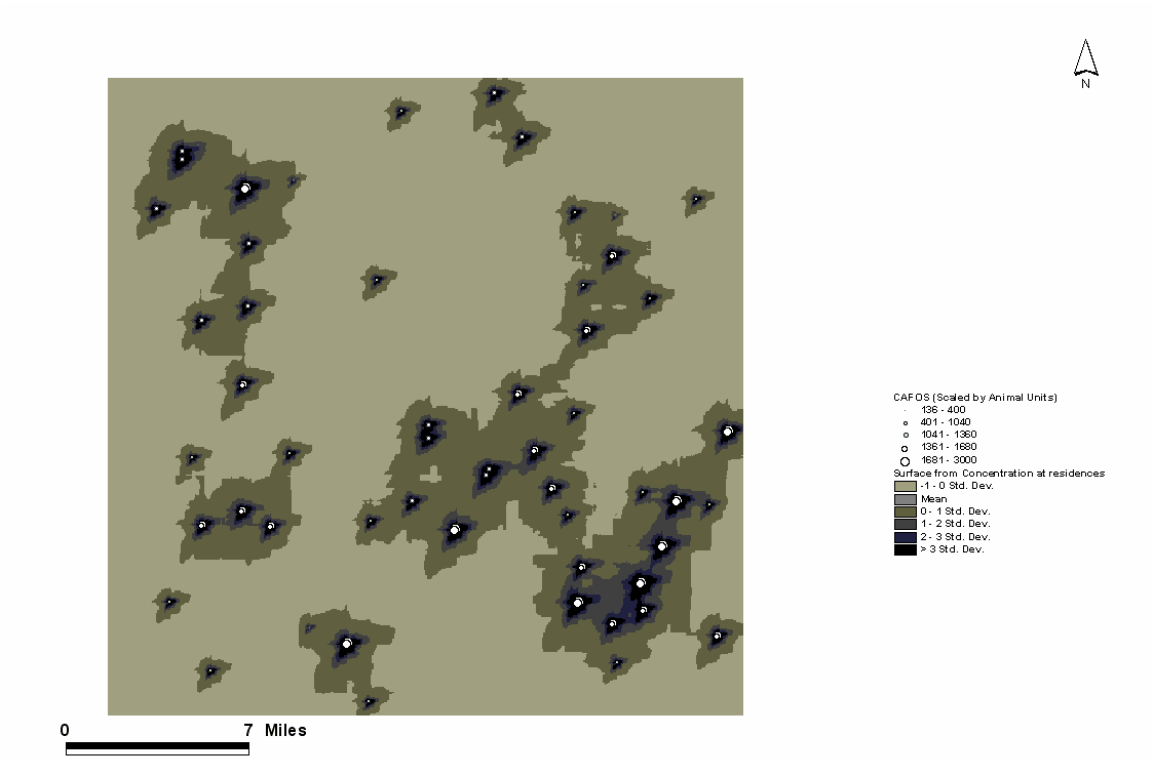
are regulated in Iowa for feedlots above a statutory determined size. All such feedlots in Carroll County have been identified and geocoded using the same orthophoto coverage as described above. The inputs to the airborne model are number of animal units in each CAFO and the dispersion of particulates is for a particulate unit dispersed. Hence the contaminant levels on the map we produce are all relative to the unknown amount released by one unit of feedlot production. The Gaussian plume model used is distorted from the symmetrical circular norm through the use of dispersion equations with parameters that describe typical prevailing wind directions and strength as estimated at each CAFO site location. The output of the model is the levels of exposure at locations at any given distance and direction from the base location. This pattern, illustrated in Figure 2 became a floating grid which we moved sequentially through each of the fifty-three CAFO sites.

Figure 2 Spatial pattern of plume output values for the unit norm CAFO release



The outputs from the application of the floating grid were integrated with the GIS model to produce a topographic-like map of Carroll County with modeled cumulative airborne exposures at all residential locations in the County. This pattern is illustrated in Figure 3.

Figure 3 The spatial pattern of estimated contaminant exposures from CAFOs in Carroll County



1.4 Masking Environmental Exposure Data to Protect Privacy

The mask on the environmental exposure data is determined as a range of data values of sufficient size to ensure that the person cannot be identified. The range is determined by calculation. It is an adaptive value calculated from a detailed population map, described below, and the estimated contaminant values shown in Figure 3. Because any measure of modeled exposure will always have uncertainty, the effect on the accuracy of the results should be in the range of negligible to none. We are currently conducting simulation studies to shed more light on this important question.

The mask ensures privacy by showing that a given number of people live in the study area within the range of exposure values from which the random exposure value was drawn. This can be determined for any person if each person's exposure value can be ranked relative to the exposure value of other people in the area. Since, for each address

in the area, an estimated exposure value is known, by ranking addresses by their exposure values we can determine a cumulative exposure curve for all addresses in the area. However, it is the number of people with an exposure value, rather than the number of addresses with an exposure value, that protects privacy. Therefore we estimate the number of people with a given level of exposure by computing the number of people with exposure values from a detailed, high spatial resolution, population map of the area. For this we use the population map of Iowa developed by the Oak Ridge National Laboratory GIS Technology group (Bhaduri et al. 2004) to which they have given the title LandScan USA. This very high-resolution (3 arc seconds or approximately 90 meter x 90 meter cell size) population distribution data is illustrated in Figure 4. The dataset is projected in Lat Long (State Plane Geographic Coordinates using the NAD -1983 Datum as the Geoid).

The centroids of the above polygons were used and the contamination value at each of these centroids was calculated using the contamination calculation program described above. Thus, we now have an estimate of the number of people with a given exposure value. This methodology ensures that for each physical address in Carroll County, Iowa, there will be at least forty other people in the area whose exposure values are within the range of values attached to it. We have written the environmental masking program that uses the tabular version of Figure 5 that enables the Iowa Department of Public Health to select an exposure value randomly within the range attached to each address that it will then substitute for the computed exposure value that UI originally attached to the address. Following attaching the random exposure value to the address and deleting the original exposure value, IDPH will add the health encounter data to the record and send to the UI project group. At this point the UI CARHI team will be unable to identify the person or the location of the health data but will be able, using statistical analysis methods not described in this paper, to analyze the association between exposure measures and health effects. In as yet unpublished work, the project team (see Acknowledgments below) has validated the method using Monte Carlo simulation studies.

Figure 4 The ORNL population data grid for a small sample area of Carroll County (see Bhaduri et al. 2004)

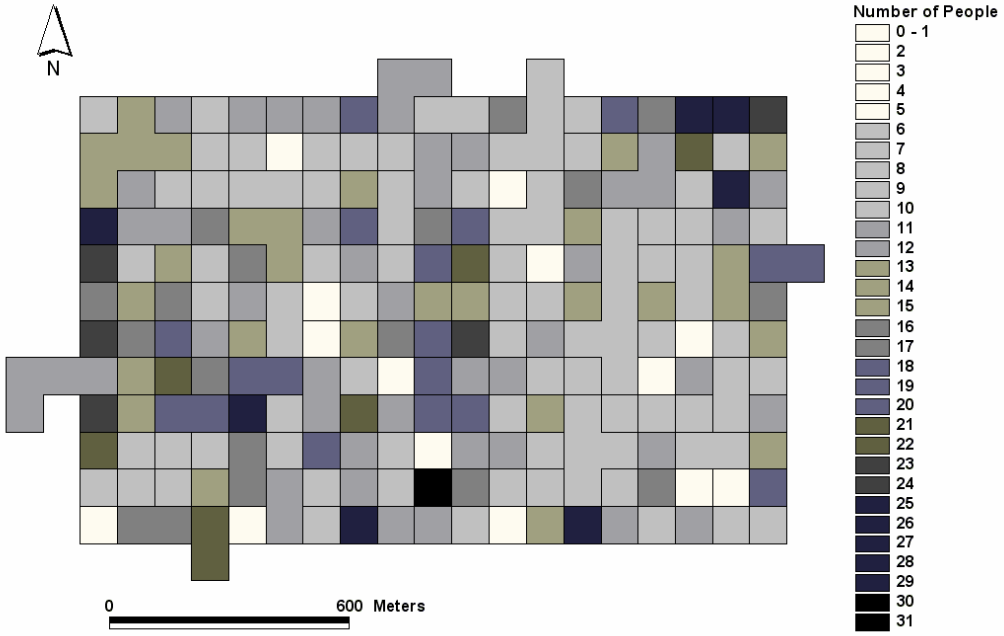
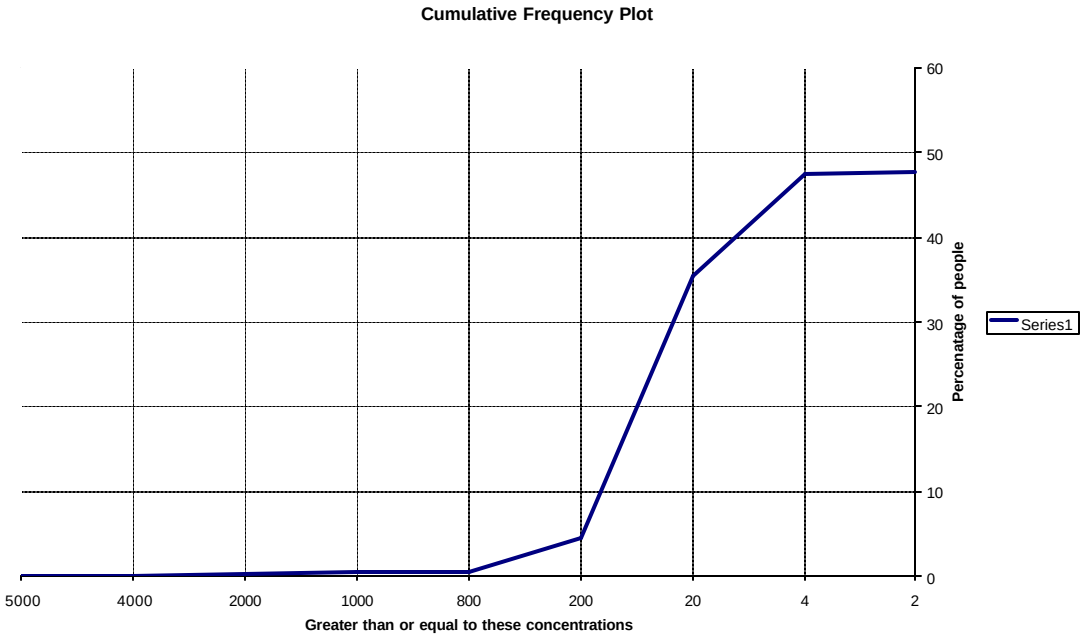


Figure 5 Cumulative percent of people in Carroll County with estimated airborne concentrations



Acknowledgments

The project described here is a GIS component of the "Comprehensive Assessment of Rural Health in Iowa" project which is funded by CDC with a subcontract from the Iowa Department of Public Health to the Iowa Center for Agricultural Safety and Health at The University of Iowa. The Project Director is Professor Kelley Donham. Other personnel are Jaysi Butler, Eileen Fisher, Michael Humann, Paul James, Soumya Mazumdar, Gerard Rushton, Peter Weyer. Ishwari Sivagnanum also participated. Patrick O'Shaughnessy developed the plume model for particulate dispersion. At the Iowa Department of Public Health, Kenneth Sharp and Thomas Newton direct the project. We thank Mr. Carl Wilburn, GIS Coordinator for Carroll County, Iowa, for assistance with GIS coverages of Carroll County. The Geographic Information Science & Technology Research Group at Oak Ridge National Laboratory Provided the high resolution population distribution data under a subcontract to the University of Iowa. We acknowledge the support of Project NCI-PC-25014-20 from the National Cancer Institute and The Centers for Disease Control and Prevention.

References

Armstrong MP, Rushton G, Zimmerman DL. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, **18:**497-525.

Bhaduri, B, Bright, E., Coleman, P. 2004. Development of High Resolution Population Distribution Data to Enhance Cancer Prevention and Control Research. Oak Ridge National Laboratory, Report. http://www.uiowa.edu/~gishlth/UIORNL/2_SEER_reportl_ornl_0304.pdf, accessed Jan. 15 2005.

U.S. Health Resources Administration. 2005. HRSA's HIPAA Web Site: http://www.hrsa.gov/website.htm, accessed January 14, 2005.

U.S. Department of Health and Human Services. 2005. National Institutes of Health, HIPAA Privacy Rule for Researchers: http://privacyruleandresearch.nih.gov/, accessed January 14, 2005.