

# **Procedures for Geomasking to Protect Patient Confidentiality**

**Dave Stinchcomb\***

**ESRI International Health GIS Conference  
October 19, 2004**

\* Based on work done at the Texas Department of Health.  
Now with the National Cancer Institute  
Email: [StinchcD@mail.nih.gov](mailto:StinchcD@mail.nih.gov)

Presented at the ESRI International Health GIS Conference held in Washington, DC, October 17-20, 2004.

## Overview

- ◆ **Motivations**
- ◆ **Survey of geomasking methods**
- ◆ **Random perturbation method:**
  - **General geomasking procedures**
  - **Issues and potential problems**
  - **Example of geomasking within a GIS environment**
- ◆ **Automation and extensions**
- ◆ **Conclusions**

Overview of the presentation:

- What is the problem – why geographically mask data?
- General masking methods.
- Explore the random perturbation method in detail including implementation methods and issues.
- Ways to automate and extend the geomasking process.
- Conclusions.

## Motivations

- ◆ **Confidentiality of a patient's identity**
  - In data: latitude and longitude
  - On a map: a specific point
  - A real problem?
    - Reverse geocoding tools available
    - Perception of the public
- ◆ **Two main purposes behind geomasking**
  - Release of data for subsequent analysis
    - Spatial relationships are important
  - Maps for public release
    - Equivalent visual patterns

The underlying purpose is to protect the identity of health subjects. Identity can be inferred from a patient's address. Hence address fields are treated as part of the patient's confidential information and removed prior to release of the data for research purposes. With geocoding, the address is converted to a latitude and longitude and these can be used to represent the subject on a dot map. So the latitude and longitude fields in data records and a specific point on a map should also be considered confidential information.

Is there a real risk of disclosure from latitude and longitude or a point on a map? Depending on the scale of a map, you can often estimate the latitude and longitude of a specific point. There are tools available that perform "reverse geocoding" - generating an approximate address based on a latitude and longitude.

In addition, public perceptions are important. If you present a dot map of disease cases in public and one of the subjects is in the audience, they may feel that their privacy has been compromised. We have a responsibility to make a good-faith effort to protect the identity of health subjects.

Geomasking is usually done for one of two types of data releases: (1) the release of tabular data for subsequent analysis and (2) the public presentation of a map. In the first case, the goal is to preserve spatial relationships so that the analysis results are not impacted. In the second case, the goal is to maintain an equivalent visual pattern after geomasking.

## **Survey of Geomasking Methods**

- ◆ **Excellent source: Armstrong et al. 1999.**
- ◆ **Transformations:**
  - **Translation: shift all points a fixed distance and direction**
  - **Scale: expand or contract all points by a scaling factor**
  - **Rotation: rotate all points by a fixed angle about a pivot point**
- ◆ **Random perturbation:**
  - **Random distributions – uniform or normal (uniform allows a min and max displacement)**
- ◆ **Aggregation**

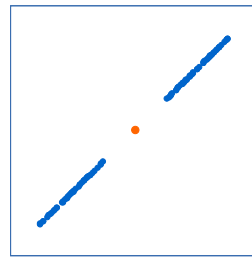
Armstrong et al. 1999 has an excellent summary of geomasking methods. The major categories include transformations, random perturbation, and aggregation.

Transformations are usually used for release of data for subsequent spatial analysis. Maintaining spatial relationships is the most important goal. One problem: this can shift subjects away from a potential pollution source etc.

Random perturbation is what most people think of when they hear the term “geomasking”. Good for publicly released dot maps. We focus primarily on random perturbation.

## Random Perturbation Methods

- ♦ **Uniform random distribution**
  - Usually  $\geq 0$  and  $< 1$
  - Can shift and scale to include negative values
- ♦ **Minimum and maximum distance**
- ♦ **Common pitfall:**
  - Generate a single random distance, add it to both X and Y



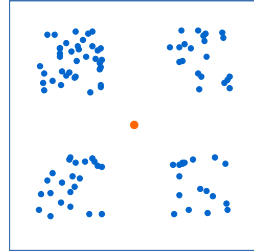
For random perturbation, a uniform random distribution is convenient since the minimum and maximum can be controlled (a normal distribution has infinite tails). Most random number generators return a value between zero and one. These can be scaled and shifted as needed.

The charts on this slide and the next slide show 100 different random perturbations from the point in the center.

Adding a single random number to both the longitude and latitude (X and Y) will force all randomized points to be on a line from the SW to the NE of the original. This is rarely what is intended. The example includes negative values and a minimum.

## Random Perturbation Specifics

- ◆ Two random distances, add one to X, one to Y

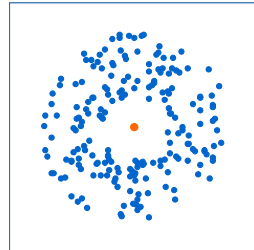


- ◆ A random distance and a random angle

With a bit of trigonometry:

$$x' = x_0 + \text{RandDist} * \cos(\text{RandAngle})$$

$$y' = y_0 + \text{RandDist} * \sin(\text{RandAngle})$$



Another approach is to generate two random numbers, one for the longitude and one for the latitude. The resulting pattern is in a square shape – again rarely what was intended. With negative values and a minimum, you get four squares as shown in the example.

Normally, the goal is to move a point within a circular region around the original point. This can be done by generating a random distance and a random angle. The adjustments to the longitude and latitude (X and Y) involve some simple trigonometry to break the distance into its X and Y components.

## Choosing Min and Max Distances

- ◆ **Goals**
  - **Minimum distance must be enough to make true location uncertain**
    - For example, 10 feet is not enough
  - **If maximum is too large, spatial patterns will be lost**
- ◆ **Proposed basis for decision: the average distance between subjects**
  - **Estimate by:  $\sqrt{1/\text{PopDensity}}$**
  - **Minimum: 1 to 2 times the average distance**
  - **Maximum: 3 to 5 times the average distance**

How do you decide how far to move a point? You need to move it far enough to mask the true location. However, if you move all of the points too far, the resultant map will not reflect the true spatial distribution.

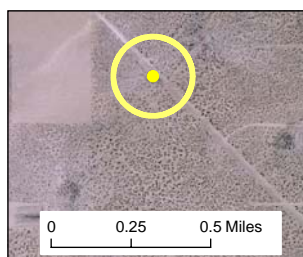
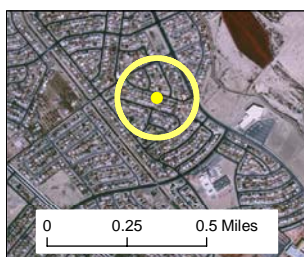
Suggestion: base the decision on an estimate of the average distance between subjects. This can be derived from the population density:

If population density is people per square mile,  
then  $1/\text{PopDen}$  (the inverse of population density) is square miles per person,  
and the square root of  $1/\text{PopDen}$  is miles per person (average distance  
between people)

Setting the minimum and maximum is still a subjective decision but should be guided by the average distance between people.

## Varying Population Density

- ♦ What if the population density varies within the study area?
  - The distance needed to protect confidentiality is larger in less densely populated areas



- Vary the perturbation distance based on the population density (Armstrong et al. 1999)

This presentation was the result of geomasking work done for a study of childhood blood lead levels in El Paso County, Texas. As is often the case, the study area contained a wide variety of population densities. We knew that masking the points in urban areas required only a small distance but that this same distance would not hide anything in the rural areas. This was particularly worrisome because we were working with children ages 0 to 6 and, in the rural areas, there might only be one child for miles around a given location.

Examples are shown from the El Paso study area comparing urban to rural areas. These aerial photos are true color – El Paso a desert environment. In the rural area photo, the houses are in the darker areas (where there is more vegetation due to irrigation). The example on the right above includes three houses, one under the dot and two below it on either side. A one-eighth-mile buffer may be adequate in the urban area but not in the rural area.

A solution suggested by Armstrong: use a different distance for each point based on the local population density.

## Implementation

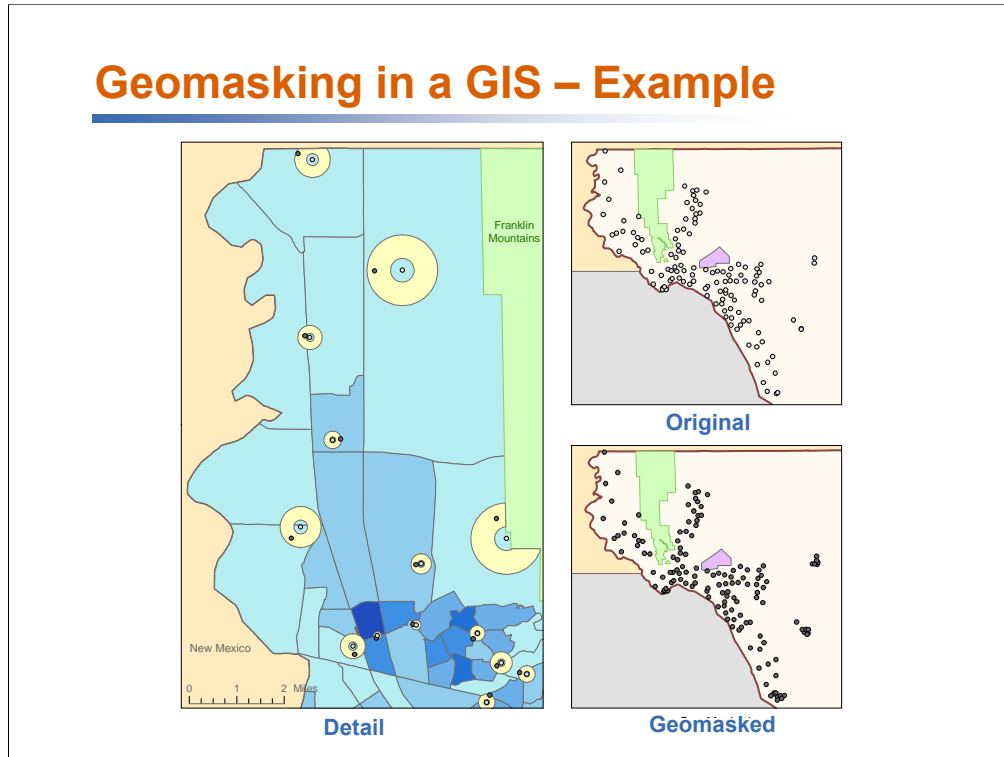
- ◆ How to implement random perturbation?
  - Outside a GIS (spreadsheet or database)
    - Can perform arithmetic on X and Y (latitude and longitude) fields
    - Can set min and max based on population density if known (relatively fixed or available in table)
  - Inside a GIS
    - Random point within a circular buffer
    - Can easily add population density layer
      - ◆ Set different min and max for each point
    - Can avoid placing points “out-of-bounds”
      - ◆ In the water
      - ◆ Outside an administrative boundary

Random perturbation can be done outside a GIS in a spreadsheet or database application. But there are limits to what can be done in this environment.

Working within a GIS has several advantages:

- The randomization can be done by selecting a point in a polygon. This avoids the arithmetic issues adjusting latitude and longitude values described earlier.
- It is easy to add a population density layer and set a different min and max value as the population density varies.
- You can also avoid placing masked points where they clearly don't belong.

## Geomasking in a GIS – Example



An example using public school locations in El Paso County.

In the figure on the left, subject points are in the center of each donut. The blue classed layer is population density by block group – dark blue is more dense, light blue is less dense. Even within a class, population density varies. The size of each donut is a function of the population density of the census area at that point (with a minimum and maximum). The donuts have been clipped by an in-bounds area layer to avoid placing points outside of El Paso County or in the Franklin Mountains State Park. For each subject, a random point has been generated within the donut polygon.

The two figures on the right are dot maps before and after randomization. Note that the visual impression of the overall spatial pattern is not affected by the randomization.

## **Geomasking in a GIS – Steps (1 of 2)**

- ♦ **Add/build population density layer**
  - As detailed as possible (e.g., block groups)
  - Match study population (e.g., children ages 0-6)
- ♦ **Spatial join subjects to population density**
- ♦ **Add MinDist and MaxDist fields and calculate**
  - Example:  $\text{MinDist} = 1 \times \sqrt{1/\text{PopDensity}}$
  - Convert distance units to match the projection
- ♦ **Create buffer donut polygons**
  - Buffer subject points at MinDist and MaxDist
  - Subtract

Here are detailed steps for random perturbation geomasking inside a GIS.

Some notes:

- Use as detailed population density layer as possible. The population should be based on your study population (for example, children ages 0-6).
- It may be necessary to convert the distance units for the MinDist and MaxDist to match the distance units used in the projection.

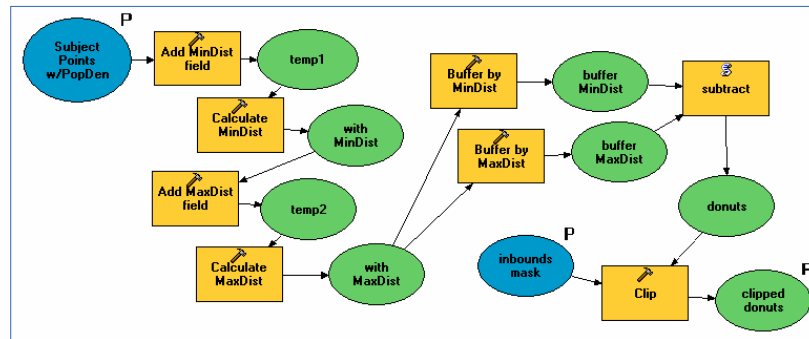
## **Geomasking in a GIS – Steps (2 of 2)**

- ♦ **Build an “in-bounds” mask**
- ♦ **Clip buffer donuts with the in-bounds mask**
- ♦ **Generate a random point in each donut**
  - **Many tools available to generate random points in polygons**
  - **One example: Hawth’s Generate Random Points tool – versions available for ArcGIS 8.x and 9.x**
- ♦ **May want to join with original attribute data**
  - **Attribute values needed for final map**

Detailed steps for random perturbation geomasking inside a GIS – continued.

## Automating the Process

- ◆ Much of the process can be easily automated
- ◆ For example, automating the creation of clipped donuts:

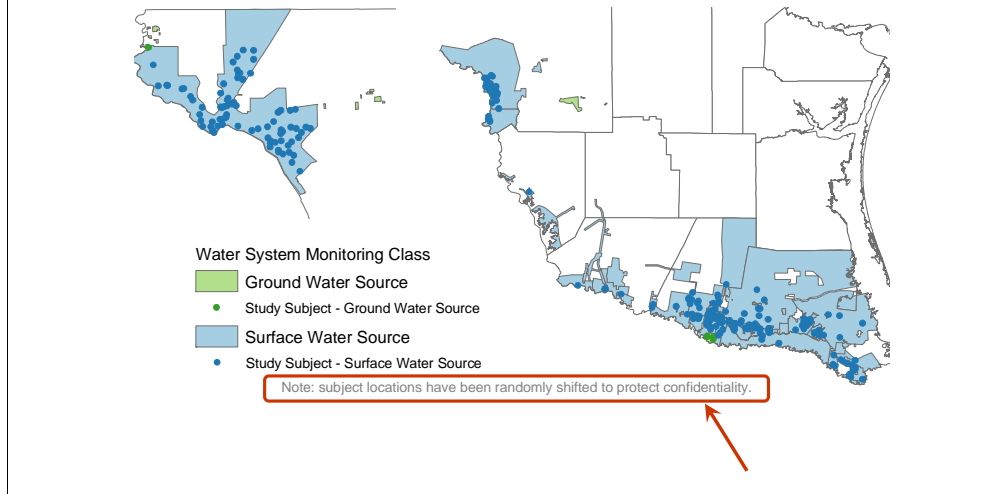


These steps can be automated. An ArcMap Model Builder example is shown. There are two inputs: the subject points (with population density attribute values), and an in-bounds mask. The output is a layer containing the clipped polygons. These can be fed into a random-points-in-polygon tool.

The process could also be built into a script and each point treated individually. This would avoid the generation of the many intermediate layers shown above.

## Final Map

- ◆ Include a note that locations have been randomly shifted to protect confidentiality



An important final step: be sure to include a note on the final map indicating that the locations have been masked.

The example shown is from an evaluation of public water sources for study subjects in the US/Mexico border region.

## **Extending the Process**

- ♦ **Collocated points – subjects at the same location**
  - **Can randomize separately or together**
- ♦ **Other geomasking methods in a GIS**
  - **Transformations**
    - **Can shift and rotate using standard GIS editing capabilities**
    - **Can avoid moving points out-of-bounds**
  - **Aggregation – easy in a GIS**
  - **Can add k-nearest neighbor information**

Many extensions are possible.

For collocated points – cases in the same family or the same apartment building – you may want to keep them together or spread them out for visibility. Can randomize either way – as cases or as case-locations.

Many of the other geomasking techniques can also be implemented in a GIS environment.

## Conclusions

- ♦ **Choose geomasking method based on goals**
  - Transformations, random perturbation, aggregation
- ♦ **For random perturbation:**
  - **Make sure randomization does what you expect**
    - Circular area: use a random distance and random angle
  - **Choose perturbation distance based on population density**
  - **In a study area with varying population density, scale the perturbation distance**
- ♦ **Using a GIS has several advantages:**
  - **Random point within a buffer polygon**
  - **Easy to get population density at each point**
  - **Can avoid placing points out-of-bounds**

### Summary:

- Base the choice of method on the goals of the project.
- Make sure the randomization does what you expect. If you are adjusting latitude and longitude values, use a random distance and angle method. If you are working within a GIS and using a random-point-in-polygon tool, be sure the tool generates a uniform distribution of points (you can experiment using a test file of 100 points all at the same location).
- Choose distances based on population density and consider varying the distance if the population density varies over the study area.
- Working with a GIS has several key advantages.

## References

- ♦ Armstrong MP, Rushton G, Zimmerman DL. 1999. Geographically Masking Health Data to Preserve Confidentiality. *Statistics in Medicine* 18:497-525.
- ♦ Meador M and Ruggles AJ. 2000. Steps Involved in Randomizing the Coordinates of Address-Matched Locations. *Public Health GIS News and Information* 35:11-14.
- ♦ Beyer, Hawthorn. Hawth's Tools – Generate Random Points Tool. Available at: <http://www.spatial ecology.com/htools/rndpnts.php>

The first reference provides an excellent survey of geomasking methods.

The second reference includes detailed steps for geomasking by adjusting latitude and longitude values that are not included here. Two cautions: (1) the spreadsheet example adds two random numbers to the latitude and longitude resulting in a square pattern rather than using the distance and angle method recommended here and (2) the GIS example (which adjusts latitude and longitude in the attribute table and imports the new table) adds a single random number to both the latitude and longitude, resulting in a straight-line SW-NE pattern of adjusted points.

The third reference is a pointer to one of several random-point-in-polygon tools available.