A Novel, Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection

CA Cassa, MEng, SJ Grannis, MD MS, JM Overhage, MD PhD, KD Mandl, MD MPH
Children's Hospital Informatics Program, Children's Hospital Boston & The Regenstief Institute

Objective

The use of spatially-based methods and algorithms in epidemiology and surveillance presents privacy challenges for researchers and public health agencies. We describe a novel method for anonymizing individuals in public health datasets, by transposing their spatial locations through a process informed by the underlying population density. Further, we measure the impact of blurring patient locations on detection of spatial clustering as measured by the SaTScan purely-spatial Bernoulli scanning statistic.

Background

There is an inherent tension between needing precise patient locations to accurately detect an outbreak and the need to protect patient privacy. Case locations that are identified using home address or a portion of that address, such as the zip code or census tract, increase the risk of breaching patient confidentiality.

We describe a spatial anonymization method based on skewing precise geocoded case locations using knowledge of local population characteristics. Masking the identity of an individual in a densely populated urban area, for example, does not require as great a skew as one in a sparsely populated rural setting.

Methods

Cases were emergency department (ED) visits for respiratory illness. Baseline ED visit data were injected with artificially-created clusters ranging in magnitude, shape, and location. Degree of anonymization is defined in terms of *k*-anonymity[1] – where each patient is not identifiable among *k* other patients. To achieve a particular *k*-anonymity value in a given dataset including both high- and low-population density areas, the distance that each patient is moved should be inversely related to the local population density. Thus, patients in rural areas are moved a greater distance than those in cities. Additionally, age-based adjustments were integrated to compensate for spatial age-group population density variations.

Optimally, individual points will be skewed by a minimal distance to obscure identity, while preserving spatial information. Dataset blurring used a randomized Gaussian probability distribution function to allow most cases to be moved a small distance. Datasets were anonymized at ten different levels of spatial blurring. The sensitivity and specificity of cluster case detection using the SaTScan purely-spatial Bernoulli scanning statistic (p-value <= 0.05) were calculated.

Results

The anonymization algorithm produced skew of cases which resulted high values of dataset *k*-anonymity. De-identification that moves points an average distance of 0.25km lowers the spatial clustering detection sensitivity by less than 4%, and lowers the spatial cluster specificity less than 1%. As the average dataset distance from original point increases, the percentage of points that do not achieve a given k-anonymity value decreases. In this example, it is possible to calculate that a k-anonymity value of 20 has been reached in 99% of all patients in a sample dataset when the average distance to original point is 0.25km.

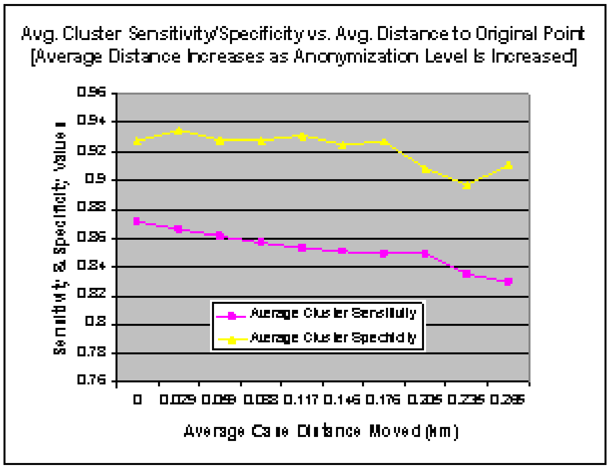

Figure 1 – Average Cluster Sensitivity/Specificity vs. Average Distance to Original Point [Average Distance Increases as Anonymization Level Increases]: The average sensitivity and specificity of spatial detection (using SaTScan Bernoulli Spatial Model with p-value <= 0.05) of artificially-injected clusters of patients is displayed with respect to the average distance that patients in a de-identified dataset are moved with respect to their original home addresses. Sensitivity and specificity are calculated using cases from the cluster and control data.

Conclusions

A population-density based Gaussian spatial blurring method markedly decreases the ability to identify individuals in a dataset while only slightly decreasing the performance of a standardly-used outbreak detection tool. These findings suggest new approaches to anonymizing data used in spatial epidemiology and surveillance.

References

[1] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Further Information:
Christopher Cassa, cassa@mit.edu
http://www.chip.org/