# DISCLOSURE RISK OF CONTEXTUAL DATA: THE ROLE OF IDENTIFIED GEOGRAPHY, SPATIAL SCALE, AND NESTING OF INFORMATION IN PUBLIC-USE FILES

Kristine M. Witkowski

Inter-University Consortium for Political and Social Research
University of Michigan

This project investigates the ways that contextual data associated with individual-level data in a dataset relate to disclosure risk. This problem is significant because contextual information is both essential to modern forms of analysis and have the potential (depending on how the data are constructed) either to increase or decrease the risk of disclosure. Analyzing an array of contextual data at different spatial scales, I simulate models to measure the likelihood of pinpointing geographic location under various distributional scenarios. Specifically, I investigate how disclosure risk for geographic units is affected by (1) spatial-scale; (2) providing contextual data at multiple geographic-levels; and (3) identifying region and state.

Laws of probability predict that the likelihood of identifying a unique is negatively associated with the number of units within the total population. Since small geographic units tend to be more numerous, disclosure risk declines with the areal size, or spatial scale, of geographic units. However, the ability to identify population uniques is enhanced when more information (i.e., keys) is provided about a given location. Consequently geographic units are generally more easily reidentified when contextual measures are provided at two or more scales. Furthermore geographic units are also more easily reidentified when we directly identify state/regional location; thereby limiting the scope of the disclosure assessment to populations within these areas.

To assess these effects, experiments were conducted to estimate the amount of disclosure risk associated with the characteristics of the test data collection and its masking method (Domingo-Ferrer and Torra, 2001a and 2001b)[1]. Using individual geographic units as well as the "test data collection" as my units of analysis, the amount of disclosure risk as the outcome of interest, and associated experimental traits, descriptive statistics and maps have been produced; with multivariate analyses forthcoming.

Disclosure risk is conceptualized as the probability of correctly identifying a specific geographic-unit, where its p-value equals one divided by the number of other "potential matches" (i.e., p<.05 represents 20 or more matches per unit). For each scenario, aggregate risk is calculated as the average risk among all spatial units.

Representing my test data collections, a finite set of five contextual concepts are measured at the county-, tract-, and blockgroups-levels for a sample of geographic units and are held constant across all my simulations: percent of the population who are either (1) male, (2) white, (3) age 0 to 11, (4) age 65 and over, and (5) have family income of $75,000 or more.

My sources of contextual data are summary files tabulated from the 2000 U.S. Census of Population and Housing (ICPSR Studies #13402, 13566, and 13576). A stratified sample of blocks was drawn to reflect the areal distribution of the U.S. population within each state. Tabulations from counties, tracts, and blockgroups, which overlap with our sampled blocks, are included in my study as contextual data.

Given the continuous nature of my contextual data, masking is inherently necessary prior to any meaningful risk assessment. As a simple exposition, I present results only for a series of non-perturbative masks,

[1] Domingo-Ferrer, Josep and Vicenc Torra. 2001a. "A Quantitative Comparison of Disclosure Control Methods for Microdata." Pp. 111-133 in *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, edited by Pat Doyle, Julia J. Lane, J.M. Jules, and Laura M. Zayatz. Elsevier Science.

Domingo-Ferrer, Josep and Vicenc Torra. 2001b. "Disclosure Control Methods and Information Loss for Microdata." Pp. 91-110 in *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, edited by Pat Doyle, Julia J. Lane, J.M. Jules, and Laura M. Zayatz. Elsevier Science.

consisting of my continuous measures collapsed into a wide-range of categorical systems (e.g., 5%, 10%, 20%, and 25% categories).

Given its flexibility to incorporate both unperturbed and perturbed data, *distance-based record linkage* is used to assess disclosure risk in test data collections. The contextual characteristics of each sampled record is compared with a master contextual file containing the same measures – unperturbed but recategorized – for the full population of geographic units and their identifying information. Test datasets that directly identify either region or state-location will be matched against regional or state-specific master contextual files.

Presented in Table 1, results from aggregate scenario data indicate that counties are more easily reidentified when a collection contains direct location identifiers. When only contextual data are available (masked into 25% categories), counties have a probability risk of .012. This risk increases to .053 and .148 if the collection identifies states or divisions, respectively. As expected, disclosure risk increases when the geographic scope of a study is constrained to a sub-national level.

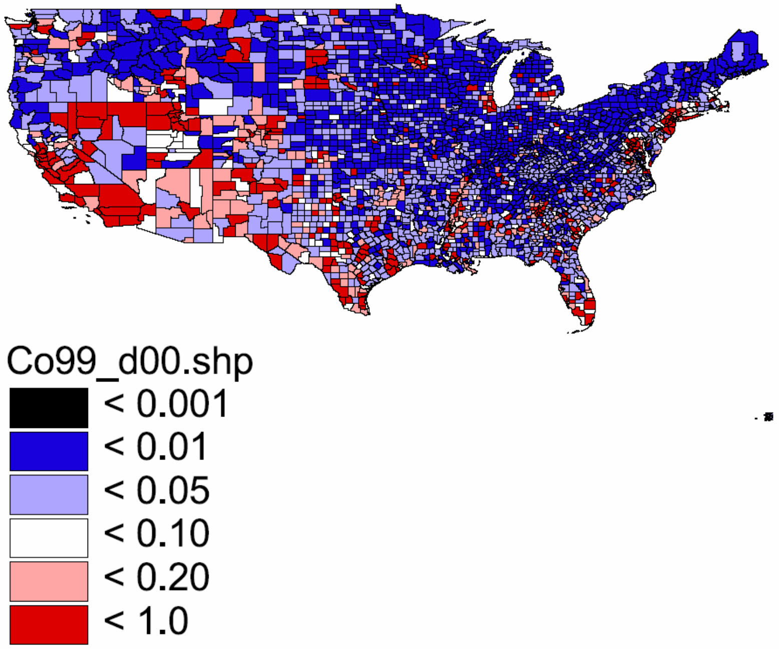
The spatial scale of contextual data is also related to disclosure risk. With 5% categorical measures and the identification of state-location, counties are most easily reidentified – compared to either tracts or blockgroups. Risk is further compounded when contextual data is provided at more than one scale. There is a 41 to 46% increase in risk when nested contextual data are made available.

Presented in the two maps below, spatial patterns of disclosure risk for counties are illustrated when data are collapsed into either 10% or 20% categories. Reidentification is most likely among the more heterogeneous counties along coastal and highly urbanized areas. These “disclosure hot-spots” are indicative of the concentration of two populations: (1) low-income, ethnically diverse persons, and (2) high-income, predominately white, and older persons. As expected, it is more difficult to reidentify counties when contextual data are collapsed into large categories. The less detailed the information, the less risk of disclosure.

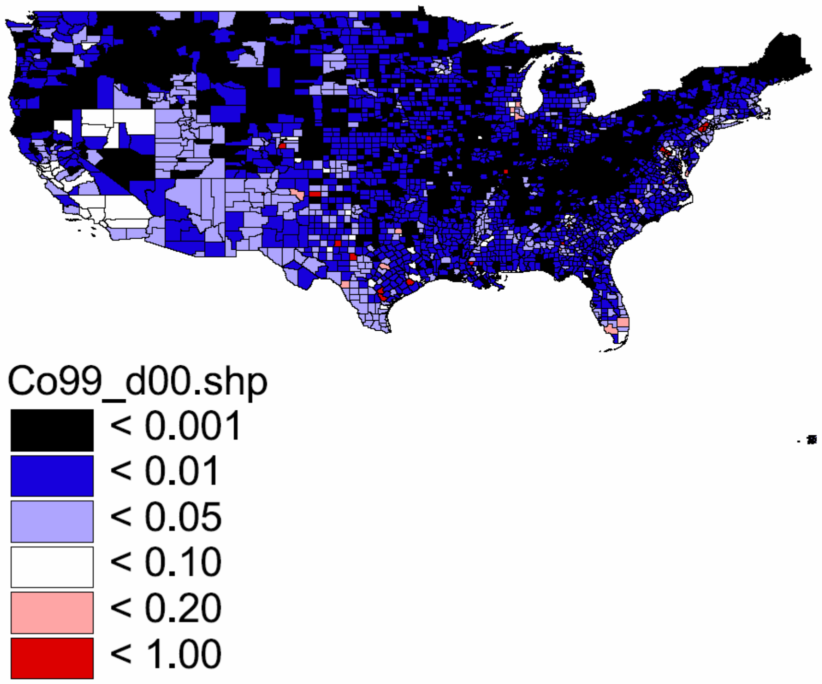
**Table 1: Disclosure Risk Associated with Identified Geography, Spatial Scale and Nesting of Contextual Data**

| Identified Geography, County Geo-Units (25% Categories) | Hyp. | Avg. Risk | Spatial Scale, With State Identified (5% Categories) | Hyp. | Avg. Risk | Nested Data, With State Identified (5% Categories) | Hyp. | Avg. Risk |
|---|---|---|---|---|---|---|---|---|
| Nation | + | .012 | Counties | +++ | .718 | County-Tract Dyads | +++ | .871 |
| Division | ++ | .053 | Tracts | ++ | .617 | County-Blockgroup Dyads | +++ | .905 |
| State | +++ | .148 | Blockgroups | + | .656 | Tract-Blockgroup Dyads | ++ | .960 |

## Disclosure Risk P-Value: 10% Categories



## Disclosure Risk P-Value: 20% Categories


Co99_d00.shp
< 0.001
< 0.01
< 0.05
< 0.10
< 0.20
< 1.00