ICPSR
A PARTNER IN
SOCIAL SCIENCE
RESEARCH

INTER-UNIVERSITY
CONSORTIUM FOR
POLITICAL AND
SOCIAL RESEARCH

UNIVERSITY OF MICHIGAN

Working Paper

Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files

Kristine M. Witkowski
Inter-university Consortium for Political and Social Research,
University of Michigan

ICPSR Working Paper Series
Working Paper No. 2

June 2008

Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files

Kristine M. Witkowski

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

Spatial information is essential for modern forms of analysis; and as a result, researchers have increasingly called for geographically-specific microdata. Contextualized microdata is one way to safely release this information without identifying the location of survey respondents. Analyzing an array of contextual data, I conduct reidentification experiments for 14,796 simulated datasets to measure the likelihood of pinpointing geographic locations under various distributional scenarios. Specifically, I investigate how disclosure risk for geographic units is affected by (1) spatial scale; (2) the identification of region, state, and population density; and (3) the number and coarseness of contextual variables provided in a dataset. Using the "datafile" as my unit of analysis, the proportion of easily-reidentified geographic units as the outcome of interest, and associated experimental traits, I detail the complexity of disclosure patterns that emerge when constructing public-use files that provide contextual data.

Key Words: confidentiality, data sources and archives

Acknowledgements: Research support from the National Institute of Child Health and Human Development (NICHD), Grant 5 P01 HD045753 as a supplement to the project Human Subject Protection and Disclosure Risk Analysis, is gratefully acknowledged. Special thanks are also given to Myron Gutmann, Sanders Korenman, and Trivellore Raghunathan for their thoughtful comments.

Contact Information: Please address all correspondence to Kristine M. Witkowski, Inter-University Consortium for Political and Social Research (ICSPR), Institute for Social Research (ISR), The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106-1248; Email: kwitkow@umich.edu; Telephone: 734-763-7102.

Draft: Please do not cite without author's permission.

Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files

Kristine M. Witkowski

*Inter-university Consortium for Political and Social Research,
University of Michigan*

Spatial information is essential for modern forms of analysis; and as a result, researchers have increasingly called for geographically-specific microdata. Contextualized microdata is one way to safely release this information without identifying the location of survey respondents. Analyzing an array of contextual data, I conduct reidentification experiments for 14,796 simulated datasets to measure the likelihood of pinpointing geographic locations under various distributional scenarios. Specifically, I investigate how disclosure risk for geographic units is affected by (1) spatial scale; (2) the identification of region, state, and population density; and (3) the number and coarseness of contextual variables provided in a dataset. Using the "datafile" as my unit of analysis, the proportion of easily-reidentified geographic units as the outcome of interest, and associated experimental traits, I detail the complexity of disclosure patterns that emerge when constructing public-use files that provide contextual data.

Key Words: confidentiality, dissemination

Acknowledgements: Research support from the National Institute of Child Health and Human Development (NICHD), Grant 5 P01 HD045753 as a supplement to the project Human Subject Protection and Disclosure Risk Analysis, is gratefully acknowledged. Special thanks are also given to Glenn Deane, Myron Gutmann, Sanders Korenman, and Trivellore Raghunathan for their thoughtful comments.

Contact Information: Please address all correspondence to Kristine M. Witkowski, Inter-University Consortium for Political and Social Research (ICSPR), Institute for Social Research (ISR), The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106-1248; Email: kwitkow@umich.edu; Telephone: 734-763-7102.

Draft: Please do not cite without author's permission.

1. Introduction

Many problems in contemporary social science lend themselves to an analysis in which the individuals under study are placed in their context, especially a context that can be defined spatially, as street, block, town, county, or some other spatial unit. Data producers have found two ways of providing this information, either identifying the spatial unit (so that the data user can link the appropriate contextual data herself), or merging the contextual data, effectively adding the characteristics of the spatial unit in which the subject lives. In this second case, the record for a given individual includes that person's characteristics (e.g., age of respondent) as well as those where they live (e.g., proportion of population in respondent's neighborhood that is poor).

One reason for providing the contextual data themselves, rather than the identity of the spatial unit, is that doing so makes it more difficult to identify the spatial unit in which the survey respondent lives (Armstrong, Rushton, and Zimmerman 1999). However, it is possible that the contextual data themselves constitute enough information to be a geographical unique. If that's the case – for example if the combination of contextual information about a given spatial unit is rare among spatial units of that type – then identification is more likely, rather than less (Saalfeld, Zayatz, and Hoel 1992). Care must then be taken to modify these data to maintain their confidentiality and their statistical properties, while at the same time ensuring that the data have the maximum analytic value for the broadest user group.

Two recent studies followed this practice of adding contextual data to their analytical files. In producing their public-use files for their Residential Energy Consumption Survey, the Energy Information Administration perturbed temperature data to mask the location of weather stations (Subcommittee on Disclosure Limitation Methodology 2005; Energy Information Administration 2001). And in a study of discrepancies between official votes and exit polls in the 2004 presidential election, official tallies of the proportion of Kerry votes were blurred for a sample of Ohio precincts; thereby concealing the identity of these controversial voter locations (Kyle et al 2007). Although they address confidentiality issues stemming from contextual data, these studies do not detail the likelihood of reidentifying these locations and associated determinants.[1]

Hence my goal for this study is to provide basic research informing the design of public-use microdata files containing contextual measures at three spatial scales: (1) census blockgroups, (2) census tracts, and (3) counties. I assess one crucial source of risk associated with these contextual data, that of pinpointing a sample of locations among the total population of geographic units. In doing so, I provide estimates for this risk component and illustrate how they vary across pertinent design elements. Given this information, I then identify distribution plans that are most likely to provide a specified level of confidentiality for sets of contextual variables.

Compiling nearly 15,000 data files composed of geographic-unit records containing a variety of contextual data, I measure the likelihood of correctly reidentifying locations under alternative distribution scenarios, relating to: (1) spatial scale; (2) the identification of division, state, and MSA-status; and (3) the number and coarseness of contextual variables provided in a dataset. Using the “datafile” as my unit of analysis, the proportion of easily-reidentified geographic units as the outcome of interest, and associated experimental traits, I detail the complexity of reidentification patterns that emerge when constructing public-use files that provide contextual data.

Because microdata files typically consist of both individual and contextual measures, a full assessment of risk requires a hierarchical reidentification model that considers identifying characteristics of both survey respondents and their locations. By providing a set of baseline estimates for one key component of locational risk, this study helps lay the groundwork for such an evaluation.

2. Empirical Approach to Disclosure Analyses of Contextual Data

In this section, I outline the analytical steps involved in evaluating disclosure risk for contextual data (Armstrong, Rushton, and Zimmerman 1999; De Waal and Willenborg 1995, 1996; Interagency Confidentiality and Data Access Group 1999; Subcommittee on Disclosure Limitation Methodology 2005; United States General Accounting Office 2001; Zayatz 2005).

Widely-available summary files of census data identify all counties, tracts, and blockgroups in the United States and provide measures describing the characteristics of the population located within these geographies. Trying to safely meet user demands for geographically-rich information, a producer can attach these contextual data to their survey respondents’ locations, without directly identifying the geography (Armstrong, Rushton, and Zimmerman 1999). But before these microdata files can be released publicly, the producer must assess the likelihood that a survey respondent’s location can be correctly reidentified by an intruder using this contextual data.

An intruder may identify geographies by conducting an experiment that matches contextual information provided in the survey’s public-use file with data available from the original tabulation files for the full population of geographies. Searching within the survey file, the intruder identifies sampled locations that share a specific set of contextual characteristics. Using the same indicators, the intruder then identifies locations in the population file that have the same characteristics and compiles direct identifiers (i.e., county name, tract, and blockgroup identifiers) for these “population matches”.

A sampled location having a single match, defined as a “unique”, is unequivocally reidentified when the external database represents the full population. A location with a small number of matches, defined as a “rarity”, also faces a significant risk of being reidentified. Conversely, a location is said to be adequately obscured only when there is a sufficiently large number of population matches. In turn, the producer must decide upon a match-threshold that defines the upper bound of risk. If the number of matches falls below this threshold, then risk of reidentification is considered intolerable and therefore the contextual data is not safe for release. The producer considers information sensitivity and intruder behavior when defining this match-threshold, incorporating statistical inference arguments to justify their selection. After defining what constitutes anonymity for geographic units, the producer then assesses the amount of risk associated with a public-use file’s contextual information by calculating the proportion of locations that are easily reidentified.

Using this measure to gauge changes in risk, the producer can then modify the composition of their public-use file to meet their goal of maximizing the utility of contextual data while minimizing the chances that geographic units are reidentified. Data utility, as well as the ability to reidentify locations, is enhanced by releasing geographic identifiers and a large number of contextual variables. To effectively design their public-use dataset, the producer needs to set priorities regarding (1) the release of geographic identifiers, (2) the scope of geographic identifiers, (3) the scale of contextual variables, and (4) the number of contextual variables. The producer can then better select disclosure limitations methods (DLMs) that offset the risk associated with these contextual data.

To reduce disclosure risk in public-use microdata files, agencies often only apply nonperturbative methods in order to maintain the statistical properties of the original data, thus maximizing its utility for widely disparate and largely unknown applications (Interagency Confidentiality and Data Access Group 1999; Subcommittee on Disclosure Limitation Methodology 2005; Zayatz 2005). Consequently “global recoding” and “local suppression” are important techniques to be considered for statistical disclosure control (De Waal and Willenborg 1995, 1996). Aggregating continuous measures into various levels of coarseness decreases the likelihood that locations are reidentified (i.e., global recoding). But for geographic units that remain easily pinpointed, their contextual characteristics are not to be released on a microdata file (i.e., local suppression); where the proportion of sampled locations with suppressed contextual data are represented by the aggregate measure of disclosure risk. Because fewer measurement categories and high suppression rates increase the amount of information-loss, a producer needs to consider how data utility varies with these methods.

Using this empirical approach, I develop an experimental study that systematically varies the characteristics of contextual data that are associated with a single sample of locations. In doing so, I am able to assess how locational risk is shaped by different choices that a producer makes when designing a dataset that safely offers the most useful contextual data. Illustrating important aspects of the decision-making process, I discuss how disclosure risk is defined and issues associated with this definition in Section 3, along with providing an example of how decisions are informed by my study’s results in Section 5c.

3. Assumptions and Hypotheses

3. a. Matching Algorithm

For purposes of this study, I assess only one component of risk associated with contextual data, that of pinpointing a sample of locations among the total population of geographic units (i.e., A_j). The distribution of contextual variables characterizing these survey locations (i.e., $Y_{J|j}$) is determined by my experiment's sampling methodology. I define "Locational Disclosure Risk", my outcome of interest, as the fraction of locations associated with a survey that an intruder can confidently reidentify (i.e., LR).[2]

$$LR = \Sigma (Pr [R_J]) \tag{1}$$

$$Pr [R_J] = 1 – (Pr [Y_{J|j}] \times Pr [A_j]) \tag{2}$$

$$Pr [A_j] = f (S, G, K, M) \tag{3}$$

Where:

LR = Proportion of sampled geographic units that are easily reidentified

R_J = Sampled geographic unit (J) is easily reidentified, given "contextual" variables (j)

$Y_{J|j}$ = Sampled geographic unit (J) has "contextual" variables (j)

A_j = Combination of values of "contextual" variables (j) is considered safe
if combination occurs at least T_1 times among geographic unit population

S = Spatial scale of contextual data

G = Identified geography

K = Number of "contextual" keys

M = Masking technique

Assumption: $T_1 = 20$, "Geography Population Unique" Threshold

In turn, I assess how easy it is to reidentify a geographic unit given perfectly accurate information about its contextual characteristics (Lambert 1993; Duncan and Lambert 1989). Using simple combinations of contextual variables, the probability of a unit being reidentified depends on the number of matches found in the population; where "matches" are those units sharing the same set of characteristics (i.e., A_j). A set of geographic units having the same characteristics are considered anonymous when there are at least T_1 in the set. Identifying matches within the population depends on four characteristics of a dataset: the spatial scale of contexts (S), geographic identifiers (G), the number of contextual variables (K), and the coarseness of their measures (M).[3]

3. b. Statistical Properties and Expected Outcomes

There are dramatic differences in the number of units in the populations of counties, tracts, and blockgroups. Compared to the number of counties, there are 20.7 times and 66.3 times as many tracts and blockgroups (3,140 counties or county-equivalents; 65,133 tracts and 208,235 blockgroups; excluding Washington DC). These differences result from the methodology underlying the construction of these administrative units which places a cap on their population size. With population sizes ranging between 67 to 9,519,338 people (i.e., Loving County, Texas and Los Angeles County, California), counties are entities that have been defined legally; that is, they are created by State law or some other administrative action. In contrast, census tracts and blockgroups have been defined specifically for data collection purposes. Census tracts designate areas that are relatively uniform in their population characteristics, economic status, and living conditions, with as many as 1,500 to 8,000 people (U.S. Census Bureau 2000). With as many as 600 to 3,000 people, blockgroups further subdivide tracts into areas bounded by visible and legal features (e.g., streets, property lines) (U.S. Census Bureau 2000).

Laws of probability predict that the likelihood of identifying a unique is negatively associated with the number of units within the total population. Consequently, disclosure risk rises with the spatial scale of a geographic unit because of declining population size and the possibility of locating matches. In other words, small geographic units – that are large in number – offer more opportunities to locate matches, thereby ensuring confidentiality. Hence risk should be lowest for blockgroups, given the high probability of finding multiple matches. Applying the same logic to counties, risk of disclosure should be highest for these large-scale geographic units.

However the anonymity benefits of a relatively large number of potential matches is offset by another statistical artifact. Because space is further delineated into units that are relatively small in area, large in number, and heterogeneous, blockgroup-level data exhibit considerably more variation in contextual characteristics. Consequently, there is an increased chance of identifying unique contexts among blockgroups.

Besides these scale factors, the ability to reidentify geographic units varies with the scope of the study. Directly identifying state/regional location and MSA-status limits the size of the population that is matched upon by confining the disclosure assessment to units within these areas, thereby increasing the likelihood of locating uniques. In turn, risk should generally be higher with the release of state, regional, and population density variables. Furthermore risk of reidentification should be highest when state-location is known because they are precise geographies covering the smallest land area.

The ability to identify population uniques is further enhanced when more information is provided about a given location. Consequently geographic units are generally more easily reidentified in datasets with relatively large numbers of contextual measures. The amount of risk resulting from these keys depends on the coarseness of their measurement.

Expected Disclosure Risk Associated with Spatial Scale of Contextual Data, Identified Geography, and the Number and Coarseness of Contextual Variables

Spatial Scale (S)	Risk	Identified Geography (G)	Risk	Number & Coarseness of Contextual Variables (K, M)	Risk
Counties	+++	MSA-Status	+	1-Key or Top, Bottom-25% Categories	+
Tracts	++	Region	++	3-Keys or 10%-Categories	++
Blockgroups	+	State	+++	5-Keys or 1%-Categories	+++

3. c. Specific Considerations

To inform the design of the most typical survey datasets, I construct a moderately-sized, cross-sectional sample of locations that is randomly drawn to reflect each state's population distribution. In selecting my reidentification threshold (i.e., T_1), I take into account these survey characteristics as well as my project's need to concurrently study a variety of contextual data – whereby I consistently apply this parameter across all datasets.

Using statistical inference, survey respondents can refute an intruder's assertion that he/she has correctly reidentified their location. But to provide convincing evidence, researchers must certify that the chances of an intruder's match being correct is no better than if the intruder chose it at random from the other matches in the set. Hence, the task at hand is to avoid type I errors where we have mistakenly rejected the null hypothesis (H_0) that the intruder's locational-match is correct. Assuming that the significance level (or p-value) of 0.05 provides a sufficient test of this hypothesis, there is a 5-in-100 chance of wrongly rejecting the intruder's claim, if it is in fact true. If we are unable to reject the null hypothesis, it only suggests that there is not sufficient evidence against the intruder even though the match may still be incorrect. By correctly rejecting the null hypothesis, we can assuredly deny the intruder's claim. This is discussed in greater detail in VanWey, et al. (2005).

Although the p-value of 0.05 has been established as the standard for correctly rejecting null hypotheses (i.e., avoiding type I errors), this significance-level may be inadequate when applied to the disclosure of geographic units. Intruder costs – associated with verifying reidentified locations – are significantly affected by a dataset's scope which concentrates the geographic dispersion of matches. The question then becomes: Given the close proximity of matches, do intruders gather additional information that increases our chances of incorrectly rejecting their claim? If this is true, researchers may want to lower their significance level to offset reduced intruder costs associated with a dataset's release of geographic identifiers; whereby smaller p-values indicate stronger evidence for rejecting the null hypothesis. Further research is needed to inform this decision.

When deciding what constitutes adequate protection, researchers must also contemplate how the utility of their data will be affected. Given the small number of large geographic units, counties are generally easier to reidentify – based solely on the number of potential matches, without regards to the number and coarseness of contextual measures. Suppressing data for "at-risk" geographic units, contextual data would primarily be released for an extremely homogeneous set of counties; thereby reducing variation in these measures to the point where they become analytically useless. This suppression bias is exacerbated with increased threshold-levels (i.e., T_1).

Finally when choosing a threshold, researchers need to consider the spatial scale of contexts and whether the dataset will provide identifying geographic information. Reducing the number of matches, the ability to provide convincing evidence is significantly lowered when a dataset increases its contextual scale or limits its scope of study.

While I am unable to address variability in intruder costs, I do incorporate a consistent definition of risk that maximizes data utility across all test datasets. In turn, I choose the reidentification confidence level of $p>0.05$, where "at-risk" geographic units are those with 1 to 19 matches. This reidentification confidence level conversely translates into a threshold value of 20, where contextual data are considered safe for release when geographic units have twenty or more matches (i.e., $T_1=20$).

4. Assessing Disclosure Risk of Masked Contextual Data

As the empirical basis of my study, I conduct experiments to assess the amount of disclosure risk associated with a dataset's contextual data (Domingo-Ferrer and Torra, 2001a). Using the test dataset as my unit of analysis, the amount of locational disclosure risk as my outcome of interest, and associated experimental traits, I produce descriptive, multivariate analyses to test my hypotheses. In doing so, I followed eight methodological steps:

- select contextual measures and identify population of base variable sets;
- draw a sample of base variable sets;
- construct test datasets that vary in spatial scale of contextual measures, identified geography, number of keys, and masking method, holding constant sample of base variable sets;
- identify a set of geographic units associated with a single synthetic sample of survey respondents;
- construct microdata files composed of sampled locations, attaching test datasets of contextual data to these geographic-level records;
- reidentify a set of sampled locations, using available geographic identifiers and contextual data for counties, tracts, and blockgroups;
- calculate aggregate disclosure risk for each test microdata file as proportion of geographic units that are easily reidentified; and

- estimate aggregate disclosure risk for all possible datasets (i.e., full population of base variable sets).

4. a. Sources, Measures, and Sampled Datasets of Contextual Data

My sources of contextual data are summary files tabulated from the 2000 U.S. Census of Population and Housing (U.S. Department of Commerce 2000a, 2000b, 2000c). These summary files are prominent public-use databases within the social sciences, providing a diversity of measures and a range of geographic detail. Contextual data are compiled from published tabulations for all blockgroups, tracts, and counties in the United States.

To identify which measures would be of most interest to researchers, I draw on the sociological literature on stratification, residential segregation and mobility, and labor markets. I limit my test datasets to those having subsets of the seventeen concepts (i.e., base variables) listed below.

- Race/Ethnic Composition
  - % Persons, Non-Hispanic White
  - % Persons, Non-Hispanic African-American
  - % Persons, Non-Hispanic Asian or Pacific Islander
  - % Persons, Non-Hispanic Other Race
  - % Persons, Hispanic
  - % Persons, Foreign-Born
  - % Foreign-Born, Naturalized Citizens
  - % Households, Linguistically Isolated

- Socioeconomic Status
  - % Persons, In-Poverty
  - % Households, With Wage Income
  - % Households, Receiving Public Assistance
  - % Persons Age 25+, College Degree

- Social Context
  - % Families, Female-Headed
  - % Persons Age 16-19, Neither Enrolled nor Graduated from High School
  - % Housing Units, Owner-Occupied

- Labor Market
  - % Persons Age 16+, Civilian Labor Force
  - % Civilian Labor Force, Unemployed

The conceptual content of datasets varies with (1) the number of base variables (i.e., k = 1, 2, 3, 4, 5) and (2) which base variables are included in the sets. All possible combinations of base variables (i.e., base variables sets) are constructed for sets containing one, two, three, four, and five concepts (N_k = 17; 136; 680; 2,380; 6,188; respectively). I compile datasets – holding constant a specific base variable set – by varying its spatial scale, identified geography, and masking technique. Consequently, I can better clarify risk factors associated with spatial scale and geographic relationships and avoid confounding my results with varying sub-domains.

Relationships between variables within a dataset have implications for reidentification and vary with the geographic scale of the measures. Large amounts of variation within measures (i.e., wide range of values) increase the likelihood of identifying uniques and, therefore, disclosure risk. However large amounts of collinearity among measures may allow producers to release more variables within a dataset without drastically increasing disclosure risk. To assure an unbiased selection of measures that are representative of the degrees of variance and collinearity among all possible datasets, I sample 137 sets of base variables composed of one to five concepts. All seventeen base variable sets with a single concept are included in my study. Thirty datasets are randomly sampled from each stratum of the multiple-concept base variable sets. An assessment of the effectiveness of this sampling approach is presented in a working paper that is available upon request.

4. b. Experimental Traits

Given a finite number of sampled sets of contextual concepts (n=137), datasets ($C_{b,s,g,m}$) are varied along the [B x S x G x M] matrix of: (1) base variable sets (b = *1 to 137, described above*); (2) spatial scales of contextual data (s = *1, 2, 3 = counties, tracts, blockgroups*); (3) identified geographies (g = *1, 2, 3, 4, 5, 6 = none, population density, division, state, population density and division, population density and state*); and (4) masking techniques (m = *1, 2, 3, 4, 5, 6 = 1%, 5%, 10%, 15%, 20%, and Top and Bottom-25% categories*). Consequently, 14,796 datasets (= 137 x 3 x 6 x 6) are compiled and assessed.

Illustrating this nomenclature, two datasets are compiled using one of the sampled set of concepts consisting of five base variables (b=137): (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) % Households, Receiving Public Assistance; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed. Test dataset ($C_{137,1,1,1}$) contains these five contextual variables measured at the county-level (s=1) without any geographic identifiers (g=1), masked into 1% categories (m=1). In comparison, test dataset ($C_{137,3,3,3}$) contains these five contextual variables measured at the blockgroup-level (s=3) along with the identification of division (g=3), masked into 5% categories (m=3).

Presented in Appendix Table A-1, equal proportions of datasets across categories of spatial scale, geographic identifiers, and masking techniques reflect my experimental design. Taking a base variable set of concepts, I compile datasets that systematically vary across a matrix of these experimental traits. However, the proportions of datasets across categories of the number and conceptual composition of contextual variables reflect the random sampling of base variables sets, stratified by key-sets (i.e., including all 17 sets with 1 key; including 30 sets each with 2, 3, 4, and 5 keys). Every conceptual domain (i.e., base variable) is represented in my test datasets. Fifteen of the seventeen concepts were included in 15 to 24% of the datasets. However, “% Persons, Non-Hispanic Asian or Pacific Islander” was least likely to be represented (7% of datasets); while “% Persons, Non-Hispanic White” was most often represented (31% of datasets). These inconsistencies are strictly random artifacts.

Given these base variable sets, I then constrain my matching process by geographic identifiers released in the dataset. A dataset can directly identify the state, division, and population density of respondent location. U.S. Census geographic divisions categorize states into seven regional groups of (1) New England, (2) Middle Atlantic, (3) East North Central, (4) West North Central, (5) South Atlantic, (6) East South Central, (7) West South Central, (8) Mountain, and (9) Pacific. Population density is defined by three categories of MSA-status: (1) MSA 1-million or more, (2) MSA less than 1-million, and (3) Non-MSA (Sources: U.S. Census Bureau, 2002, 2006a, 2006b). Measured at the county-level, these data are also used to characterize the MSA-status of tracts and blockgroups.

Finally I systematically vary the amount of measurement detail across my experimental datasets; thereby assessing how rates of local suppression fluctuate with global recoding schema. After top-coding and bottom-coding my continuous variables to conceal outliers, I recode contextual measures into six grades of coarseness (i.e., 1%, 5%, 10%, 15%, 20%, and Top and Bottom-25% categories). Outliers were identified as those within the top and bottom 0.5% of each variables distribution (Zayatz, 2005), given geographically-specific distributions defined by each dataset’s identified geography. Contextual variables are recoded into aggregated categories based on their absolute values (i.e., absolute recoding). For example, let us consider a county having 72% of its population that is non-Hispanic White. Coarsening the measure into 10%-categories, the county would be characterized as having an absolute value that falls between 70% and 80%.[4]

Global Recoding of Contextual Variables					
Coarseness of Contextual Variables	Metric Spaces	Absolute Values	Coarseness of Contextual Variables	Metric Spaces	Absolute Values
1%-Categories	100	0%, 1%, 2% . . . 98%, 99%, 100%	15%-Categories	7	0 - 14%, 15 - 29%, . . . 75 - 89%, 90 -100%
5%-Categories	20	0 - 4%, 5 - 9%, . . . 90 - 94%, 95 - 100%	20%-Categories	5	0 - 19%, 20 - 39%, . . . 60 - 79%, 80 - 100%
10%-Categories	10	0 - 9%, 10 - 19%, . . . 80 - 89%, 90 - 100%	Top, Bottom-25% Categories	3	Top-25%, Bottom-25%, Other

4. c. Sampled Locations

Represented in the 2000 U.S. Census of Population and Housing (U.S. Department of Commerce 2000a), a stratified sample of blocks is drawn to reflect the areal distribution of the U.S. population across states. The block is chosen as my sampling unit because it most closely approximates the residential location of our theoretically ideal sample of individual survey respondents (i.e., persons). Being the foundational spatial unit from which all geographies are built upon, blocks also pinpoint various contexts to a single location. In turn, tabulations from identified counties, tracts, and blockgroups, which overlap with my sampled blocks, are included in my study as contextual data. These contextual data are then represented in a dataset of location-records.

Fifty-one state-specific block samples (including the District of Columbia) are drawn with probability-proportional-to-size without replacement (PPS). Each block within a state has a probability of selection that is proportional to its population density, defined as the total number of persons per square meter of block area. Presented in Appendix Table A-2, 11,562 blocks are sampled, representing 11,562 synthetic persons dispersed across approximately 5% of all blockgroups, 14% of all tracts, and 57% of all counties in the U.S. Further details about my sampling methodology and construction of weights are available upon request.

4. d. Locational Disclosure Risk Associated with Test Datasets

The foundation of my component of locational disclosure risk (i.e., A_j) is the confidence-level of correctly identifying a sampled location among a population of geographic units (i.e., T_1). Given the p-value of .05, I assume that I can strongly refute an intruder's claim of correctly reidentifying a location (H_0) when there is a 5-in-100 (or less) chance of wrongly rejecting this hypothesis (VanWey, et. al, 2005). Taking the inverse of this confidence-level, a geographic unit is considered easily reidentified when it has fewer than 20 matches (i.e., at_risk=1).

To ascertain the number of matches, I compare the contextual characteristics associated with a sample of locations – with a master contextual file containing the same measures for the full population of geographic units and their identifying information. Data in the master file are top and bottom coded and collapsed into intervals as defined by the test dataset. In that my test datasets also directly identify population density, division, and state-location, I further refine my matching process by utilizing geographically-specific master contextual files. Because contextual data for a sample of locations are originally drawn from this master file and have not been perturbed, the identification of matches is exact. Consequently, counting the number of matches simply requires that I tabulate the number of geographic units in the population file having a specified set of contextual characteristics, coinciding with those found in my experimental survey (Winkler 2004).

After assessing whether a sampled geographic unit is considered easily reidentified, I measure the amount of aggregate risk associated with a dataset by calculating the proportion of locations that are considered “at-risk” of being reidentified because of the release of contextual data (i.e., at_risk=1). This estimate measures risk for a single sample of locations associated with a survey of respondents, instead of drawing separate samples of counties, tracts, and blockgroups.

4. e. Generalized Estimates

Using the above measure, I calculate the average amount of disclosure risk associated with contextual data at three spatial scales as a function of geographic identifiers, number of contextual variables, and masking techniques. Analyzing metadata characterizing my sample of 14,796 datasets (see Appendix Table A-3 for further details), I produce estimates that are generalized to all possible datasets. In doing so, I provide point and interval estimates of average aggregate risk for 540 dataset typologies defined by my study’s experimental traits (i.e., S x G x K x M = 3 x 6 x 5 x 6).

Confidence intervals for extreme values of risk tend to be narrower than those for moderate values. Since estimates are adjusted to account for the complex survey design of sampled variables sets, this pattern does not reflect bias introduced from heteroskedasticity; rather it arises from a confluence of matching inefficiencies in my reidentification algorithm. It is easy to predict that nearly 100% of respondents will be reidentified when we have a large number of fine-grained contextual data that characterize a relatively finite geographic area. But it becomes more difficult to predict aggregate risk (with as much precision) when reidentification depends upon fewer, coarsely-grained measures that characterize a larger population of potential matches. Consequently, confidence intervals tend to be the narrowest at the extremes and widen across more moderate levels of risk. This variation should be considered when using this study’s results for designing datasets.

5. Presentation of Results

In the following section I describe how aggregate disclosure risk of geographic units varies with the experimental traits of datasets. When I refer to “aggregate disclosure risk”, I am specifically talking about the probability of confidently reidentifying a geographic unit among its population because of its contextual characteristics. As a way to encapsulate the complexity of these results, I produce summary statistics and provide a specific example of how this information may be integrated into the design of datasets.

5. a. Tool for Creating Public-Use Microdata Files with Contextual Data

In Table 1, I present the average amount of aggregate risk for each dataset typology. I also present the upper bound of this estimate in that the highest, probable amount of risk is also of concern when designing public-use datasets. This table is organized so that the reader can easily assess patterns of risk. Dividing the table into three pages, each page shows estimates for datasets with different spatial scales of contextual data. Each page is further divided into six panels, where each panel displays information for datasets with varying geographic identifiers. Within each panel, predicted values are presented for datasets with one to five contextual variables (across columns) that vary in the coarseness of their measurement (within rows). Hence, within each page, aggregate risk tends to increase as one reads from left-to-right and from top-to-bottom.

[Table 1 Here]

5. b. Descriptive Statistics

Presenting Table 1’s estimates averaged across all masking typologies, Figures A, B, and C illustrate the degree to which disclosure risk is heightened with the release of identified geography and additional contextual keys.

Looking at Figure A, the risk posed by county-level contextual data dramatically rises when we constrain the geographic scope of the dataset. Considering datasets where only one contextual variable is released, only 1% of sampled locations are at-risk when the study is national in scope. While for sub-national studies, 4 to 9% of locations are at-risk with identifying either population density or division; 20 to 29% of locations with identifying both division and population density or state alone; and 52% of locations with identifying both state and population density.

For each scope of study, disclosure risk also increases with the addition of contextual variables. The largest increase in risk occurs with the release of a second contextual variable. National datasets and those identifying both state and population density experience a 11 to 12 percentage-point jump; slightly less than the 14 to 17 percentage-point increase experienced by other geographically-specific datasets.

When datasets release 3 or more keys, there generally is a 6 to 9 percentage-point increase in the marginal proportions of locations at-risk. However, there is one exception when datasets identify both states and population densities. Since four keys already allow for 77% of locations to be easily reidentified, it is expected that a fifth key will have a relatively lesser impact on risk – with marginal rate of only 3 percentage-points.

[Figures A, B, C Here]

Similar patterns are found for datasets with tract- and blockgroup-level contextual data (Figures B and C). However, the rise in risk due to constraints in geographic scope and the addition of keys are less pronounced. For single-variable datasets with smaller-scaled contexts, the proportions of locations at-risk increase by only 12 and 5 percentage-points (for tract and blockgroups, respectively) when identifying both state and population density (compared to national datasets).

Once again disclosure risk increases the most with the addition of a second contextual variable, but only for datasets identifying division, state, and their population densities (i.e., 12 to 19 percentage-point increase for tracts; 11 to 17 percentage-point increase for blockgroups). However national datasets and those identifying only population density experience the highest jump in risk with the addition of a third variables (i.e., 8 to 10 percentage-point increase).

Furthermore marginal changes in risk tend to rise when the scope of study is constrained. Adding a second contextual key, risk increases by only 2 to 4 percentage-points for small-scale, national datasets; while experiencing a 17 to 19 percentage-point jump for datasets identifying both state and population density. With the addition of the third, fourth, and fifth contextual key, risk grows at a similar (or proportional) rate of 7 to 11 percentage-points for tract and blockgroup-level datasets.

[Figures D, E, F Here]

Presenting Table 1's estimates averaged across all key typologies, Figures D, E, and F further illustrate how risk is reduced by coarsening contextual measures. For nearly all datasets risk drops the most when data are recoded into 5%-categories; and most often risk declines considerably more with additional coarsening into 10%-categories. This is not surprising considering data are now presented across 20 and 10 metric spaces (instead of 100 metric spaces).

In datasets known for their extreme levels of risk, the application of 15%-categories (7 metric spaces) also offers a sufficient drop in aggregate risk. In fact, datasets identifying divisions, states, and their population densities experience an additional 7 to 13 percentage-point decrease in risk with this level of coarseness.

These patterns of marginal change illustrate the trade-offs between measurement detail and risk. But let us consider whether a recoding scheme typically offers a "considerable" amount of confidentiality; that of obscuring the identity of 85% or more of locations (i.e., aggregate risk-level less than 15%, averaged across the number of keys).

For national datasets or those identifying either population density or division alone, recoding tract- and blockgroup-level data into 10%-categories seems to offer a considerable amount of confidentiality. For county-level data, only national datasets or those identifying population density alone receive this level of protection with 10%-categories. Instead 15%-categories may be better suited for large-scale datasets that identify division.

For datasets that identify division, state, and their population densities, collapsing tract- and blockgroup-level measures into 15% or 20%-categories achieves this target risk-level. Geographically-constrained counties are easily reidentified regardless of the coarseness of their contextual measures.

5. c. Example of Decision-Making Process

When designing a public-use dataset, the producer seeks to release as much information as possible while ensuring that all locations are kept confidential. In doing so, the producer needs to consider the trade-offs between data utility and access to analytical files when defining acceptable levels of risk and contextual content. After choosing the mode of distributing datafiles to their user community, the producer must define what constitutes anonymity for survey respondents and their locations by selecting values of thresholds that underlie the definition of risk. Completing their reidentification experiments, the producer distinguishes respondents whose locations are pinpointed with their contextual data. This geographic information must then be deleted or modified to ensure the respondent's identity is kept confidential.

Rates of local suppression are reflected by the level of aggregate risk, expressed as the proportion of respondents whose contextual data may not be safely released. Perturbative methods, such as swapping, can then be used to construct confidential information that replaces these missing values. The producer must consider how these ascribed data may distort analyses and whether a group of geographic units is particularly affected by these aberrations.

Establishing an acceptable rate of data perturbation, the producer can achieve the coinciding aggregate risk-level by coarsening the contextual measures. However, if the necessary amount of coarsening results in analytically useless data, the producer may decide to release fewer contextual variables or less specific geographic identifiers. Finally the producer may adjust their definition of disclosure risk such that more and higher quality data can be released through more restricted modes of data access.

Given these decision-making factors and the above broad patterns of risk, I now offer an example of how the results of my study can be used to inform dataset design. In this exercise, I focus on developing a microdata file that has tract-level contextual data. Keeping in mind the general trends presented in Figures A through F, I start my review by appraising datasets with two and three contextual variables. Ideally I would prefer to also release a geographic identifier; therefore I consider the trade-offs in risk with identifying population density, division, and state. Furthermore I investigate how risk will be offset by aggregating my measures into 10%-, 15%- and 20%- categories. I do not consider national or single-key datasets seeing they pose very little threat to confidentiality.

[Figure G Here]

For datasets of 10%-categorical measures (blue bars, solid and hashed, in Figure G), the rise in risk with the addition of a third contextual variable is heightened when the scope of study is constrained. While datasets identifying population density experience a 3 percentage-point increase, risk jumps by 7 and 15 percentage-points for those identifying divisions and states (respectively).

If data are collapsed into 15%-categories (red bars, solid and hashed), datasets identifying population density and division no longer experience such increases, with less than 4% of their locations at-risk. If data are collapsed into 20%-categories (green bars, solid and hashed), datasets releasing two contextual variables and state-geography are able to achieve comparably low levels of aggregate risk, with about 4% of their locations at-risk.

If the maximum rate of local suppression is set to 5% (upper bound), three optimal designs are then indicated: (1) population density, 3-keys, 10%-categories; (2) division, 3-keys, 15%-categories; and (3) state, 2-keys, 20%-categories. The final choice among these designs ultimately hinges on the analytical value of the measures and the scientific relevance of the contextual content.

6. Conclusions

The results of these analyses clearly illustrate two well-known reidentification mechanisms (as hypothesized), whereby the amount of risk is a function of: (1) the number of potential matches and (2) the quantity of information used in the matching process. Decreasing the chances of locating numerous matches, shrinking populations generally increase disclosure risk. Increasing the chances of identifying uniques, additional information about the characteristics of geographic areas – determined by the release of more contextual variables – also tends to increase disclosure risk. But the utility of this information is dampened by the coarsening of measures to reduce reidentification.

While these broad patterns offer some insight, most interesting is the degree to which each factor helps intruders pinpoint the location of survey locations and how coarsening offsets this risk. Predicted values of aggregate disclosure risk provide broad guidelines for the production of context-linked microdata files, helping to improve the utility and confidentiality of public-use data. Increasingly required by funding agencies, data distribution plans – in their design and evaluation – are also informed by these a priori disclosure risk estimates of contextual data linkage. Finally this research holds the promise of wide scale adaptation and application to content- and user-specific products.

The current study applies a <u>single threshold</u> when analyzing a sample of base variables, where all keys within a set share the <u>same level of coarseness</u>. While these simplifications were necessary for the initial stages of this work, further research needs to be conducted to (1) inform the selection of the reidentification threshold; (2) identify more optimal recoding and suppression schemes; and (3) closely study how specific contextual measures (e.g. % persons, in-poverty) shape disclosure risk.

Furthermore I plan to quantify the utility of masked contextual data by measuring the amount of information lost (e.g., Domingo-Ferrer and Torra, 2001a and 2001b; Raghunathan, et. al. 2003; Winkler 2004), thereby better illustrating the risk-utility tradeoffs when safely constructing these data. I am particularly interested in documenting how the suppression of contextual data distorts the original sampling framework and whether sub-domains of contextual measures are differentially affected.

7. References

Armstrong, Marc P., Gerard Rushton, and Dale L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18: 497-525.

DeWaal, A.G., and L.C.R.J. Willenborg. 1995. Global recodings and local suppressions in microdata sets. *Proceedings of Statistics Canada* 95: 121-132.

DeWaal, A.G., and L.C.R.J. Willenborg. 1996. A view of statistical disclosure control for microdata. *Survey Methodology* 22: 95-103.

Domingo-Ferrer, Josep, and Vicenc Torra. 2001a. A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, disclosure, and data access: Theory and practical application for* *statistical agencies*, edited by Pat Doyle, Julia I. Lane, J.M. Theeuwes, and Laura V. Zayatz, 111-133. North-Holland: Amsterdam.

Domingo-Ferrer, Josep, and Vicenc Torra. 2001b. “Disclosure control methods and information loss for microdata.” In *Confidentiality, disclosure, and data access: Theory and practical application for statistical agencies*, edited by Pat Doyle, Julia I. Lane, J.M. Theeuwes, and Laura V. Zayatz, 91-110. North-Holland: Amsterdam.

Duke-Williams, Oliver, and Philip Rees. 1998. Can census offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure. *International Journal of Geographical Information Science* 12: 579-605.

Duncan, George, and Diane Lambert. 1989. The risk of disclosure for microdata. *Journal of Business and Economic Statistics* 7: 207-217.

Energy Information Administration. 2001. *Residential Energy Consumption Survey.* http://www.eia.doe.gov/emeu/recs/recs2001/codebook82001.txt (accessed December 27, 2007).

ESRI. 2006. *GIS Dictionary*. http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway (accessed December 27, 2007).

Interagency Confidentiality and Data Access Group, Statistical Policy Office, Office of Information and Regulatory Affairs. 1999. Checklist on disclosure potential of proposed data releases. Washington, DC: Office of Management and Budget.

Kyle, Susan, Douglas A. Samuelson, Fritz Scheuren, and Nicole Vicinanze. 2007. Explaining discrepancies between official votes and exit polls in the 2004 presidential election. *Chance* 20: 36-47.

Lambert, Diane. 1993. Measures of disclosure risk and harm. *Journal of Official Statistics* 9: 313-331.

Raghunathan, T.E., J.P. Reiter, and D.R. Rubin. 2003. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19: 1-16.

Saalfeld, Alan, Laura Zayatz, and Erik Hoel. 1992. Contextual variables via geographic sorting: A moving averages approach. In *Proceedings of the Section on Survey Research Methods*, 691-696. Alexandria, VA: American Statistical Association.

Subcommittee on Disclosure Limitation Methodology, Confidentiality and Data Access Committee, Federal Committee on Statistical Methodology. 2005. Statistical policy working paper 22: Report on statistical disclosure limitation methodology, GAO-010126SP. Washington, DC: Office of Management and Budget.

United States General Accounting Office. 2001. Record linkage and privacy: Issues in creating new federal and statistical information, GAO-01-126SP. Washington DC: United States General Accounting Office.

U.S. Census Bureau. 2000. *Census 2000 Geographic Terms and Concepts.* http://www.census.gov/geo/www/tiger/glossry2.pdf (accessed December 27, 2007).

VanWey, Leah K., Ronald R. Rindfuss, Myron P. Gutmann, Barbara Entwisle, and Deborah L. Balk. 2005. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15337-15342.

Winkler, William. 2004. Masking and reidentification methods for public-use microdata: Overview and research problems. *Research Report Series* (Statistics #2004-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

Zayatz, Laura. 2005. Disclosure avoidance practices and research at the U.S. Census Bureau: An update. *Research Report Series* (Statistics #2005-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

8. Data Sources

U.S. Census Bureau, Population Division. 2002. Census 2000 PHC-T-3. Ranking tables for metropolitan areas: 1990 and 2000 (Table 3: Metropolitan areas ranked by population). http://www.census.gov/population/cen2000/phc-t3/tab03.xls (accessed December 27, 2008).

U.S. Census Bureau, Geography Division, Cartographic Products Management Branch. 2005. Cartographic boundary files. http://www.census.gov/geo/www/cob/index.html (accessed December 27, 2008).

U.S. Census Bureau, Population Division. 2006a. Geographic relationship files: 1999 MA to 2003 CBSA. http://www.census.gov/population/www/estimates/CBSA03_MSA99.xls (accessed December 27, 2008).

U.S. Census Bureau. 2006b. 2000 Census of population and housing, summary file 1 (matrices P1). http://factfinder.census.gov (accessed November 6, 2006).

U.S. Department of Commerce, Bureau of the Census. CENSUS OF POPULATION AND HOUSING, 2000a [UNITED STATES]: SUMMARY FILE 1 SUPPLEMENT, STATES [Computer file]. ICPSR release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 2003. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, [distributor], 2003.

U.S. Department of Commerce, Bureau of the Census, and Inter-university Consortium for Political and Social Research. CENSUS OF POPULATION AND HOUSING, 2000b [UNITED STATES]: BLOCK GROUP SUBSET FROM SUMMARY FILE 3 [Computer file]. ICPSR ed. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

U.S. Department of Commerce, Bureau of the Census, and Inter-university Consortium for Political and Social Research. CENSUS OF POPULATION AND HOUSING, 2000c [UNITED STATES]: SELECTED SUBSETS FROM SUMMARY FILE 3 [Computer file]. 2nd ICPSR ed. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

9. Endnotes

[1] A complimentary yet different approach to studying disclosure risk associated with multiple scales is Duke-Williams and Rees' (1998) analysis of geographic "sliver polygons". Sliver polygons result when boundaries of units at different scales overlap but are not mutually exclusive (ESRI 2006). Risk is then assessed in terms of person-counts within these redefined small areal units. Falling below acceptable population-levels, unique slivers were found to increase the chances of pinpointing respondents who

resided within these extemporaneous boundaries (Duke-Williams and Rees 1998). Avoiding sliver polygons by assuming that datasets will only release contextual data at one spatial scale, I offer a more transparent assessment of risk.

[2] My aggregate measure of disclosure risk follows Lambert's (1993) conceptualization of the "Risk of True Identification" that is defined as the "fraction of released records that an intruder can correctly reidentify."

[3] Clarifying terms used throughout this paper, contextual measures are referred to as "key variables" being how combinations of their values are used to locate uniques within its target population. On the other hand, geographic identifiers are considered "block variables" being how they are used to subset locations within the matching process.

[4] I conduct another set of simulations analyzing contextual measures that are coarsened based on their percentile distribution (i.e., percentile recoding). Twenty percent of all counties have at most 66% of their population being non-Hispanic White (i.e., 20th percentile at 66.14%); while thirty percent of all counties have at most 76% of their population being non-Hispanic White (i.e., 30th percentile at 76.11%). Coarsening the measure into deciles categories, my exemplar county – having 72% of its population being non-Hispanic White – would be characterized as falling between 20th and 30th percentiles (i.e., the third decile).

As illustrated in Appendix Table A-3, disclosure risk is heightened considerably by this global recoding approach. Counties having a rare characteristic – those with an outlying value at the tails of a contextual variable's continuous probability distribution – are less likely to be reidentified with percentile coarsening. However, counties sharing a relatively common characteristic – those within the middle of the distribution – are actually more likely to be reidentified with percentile coarsening.

Building upon my previous example, let us consider my exemplar county which is one among a population of 3,141. With absolute recoding, this county has approximately 346 matches with values between 70% and 80%. With percentile recoding, there are 315 matches with values between 66.14% and 76.11%. In turn, percentile recoding automatically sets an upper bound to the number of matches, resulting in relatively higher risk for more typical counties.

10. List of Tables and Figures in Paper

11. List of Tables and Figures in Appendix

Table 1: Average Proportion of Geographic Units "At-Risk" of Disclosure, By Spatial Scale and Experimental Traits of Contextual Data

	County-Level Contextual Data									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)
National (None)										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.01	(0.01)	0.02	(0.02)
20%-Categories	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.03	(0.04)	0.05	(0.06)
15%-Categories	0.00	(0.00)	0.01	(0.02)	0.03	(0.04)	0.06	(0.08)	0.11	(0.13)
10%-Categories	0.01	(0.01)	0.02	(0.03)	0.07	(0.08)	0.13	(0.16)	0.24	(0.28)
5%-Categories	0.01	(0.01)	0.08	(0.10)	0.22	(0.26)	0.41	(0.47)	0.59	(0.66)
1%-Categories	0.05	(0.05)	0.60	(0.68)	0.88	(0.94)	0.99	(1.00)	1.00	(1.00)
Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.03	(0.03)	0.04	(0.04)
20%-Categories	0.01	(0.01)	0.02	(0.03)	0.04	(0.04)	0.07	(0.08)	0.11	(0.12)
15%-Categories	0.01	(0.01)	0.04	(0.04)	0.07	(0.08)	0.13	(0.15)	0.19	(0.23)
10%-Categories	0.02	(0.02)	0.07	(0.08)	0.13	(0.16)	0.24	(0.28)	0.36	(0.42)
5%-Categories	0.03	(0.03)	0.19	(0.22)	0.36	(0.42)	0.54	(0.60)	0.69	(0.75)
1%-Categories	0.17	(0.17)	0.73	(0.79)	0.92	(0.96)	0.99	(1.00)	1.00	(1.00)
Division										
Top-25%, Bottom-25%, Other	0.01	(0.01)	0.03	(0.03)	0.04	(0.05)	0.06	(0.07)	0.08	(0.09)
20%-Categories	0.02	(0.02)	0.05	(0.06)	0.08	(0.09)	0.13	(0.16)	0.19	(0.23)
15%-Categories	0.03	(0.03)	0.08	(0.09)	0.13	(0.16)	0.21	(0.25)	0.31	(0.36)
10%-Categories	0.04	(0.04)	0.14	(0.17)	0.26	(0.30)	0.38	(0.43)	0.51	(0.58)
5%-Categories	0.08	(0.08)	0.34	(0.39)	0.55	(0.63)	0.69	(0.75)	0.80	(0.85)
1%-Categories	0.37	(0.37)	0.85	(0.90)	0.96	(0.99)	1.00	(1.00)	1.00	(1.00)
Division & Population Density										
Top-25%, Bottom-25%, Other	0.05	(0.05)	0.09	(0.10)	0.11	(0.12)	0.16	(0.17)	0.20	(0.22)
20%-Categories	0.08	(0.08)	0.16	(0.18)	0.20	(0.23)	0.29	(0.33)	0.36	(0.40)
15%-Categories	0.10	(0.10)	0.21	(0.24)	0.29	(0.33)	0.39	(0.44)	0.49	(0.54)
10%-Categories	0.14	(0.14)	0.31	(0.35)	0.43	(0.49)	0.56	(0.61)	0.65	(0.70)
5%-Categories	0.24	(0.24)	0.52	(0.57)	0.68	(0.75)	0.80	(0.84)	0.87	(0.90)
1%-Categories	0.57	(0.57)	0.91	(0.93)	0.98	(0.99)	1.00	(1.00)	1.00	(1.00)
State										
Top-25%, Bottom-25%, Other	0.08	(0.08)	0.13	(0.14)	0.15	(0.17)	0.20	(0.23)	0.25	(0.28)
20%-Categories	0.13	(0.13)	0.23	(0.27)	0.30	(0.34)	0.39	(0.44)	0.47	(0.52)
15%-Categories	0.16	(0.16)	0.31	(0.36)	0.41	(0.47)	0.53	(0.59)	0.63	(0.69)
10%-Categories	0.23	(0.23)	0.43	(0.49)	0.57	(0.64)	0.69	(0.74)	0.76	(0.82)
5%-Categories	0.37	(0.37)	0.65	(0.72)	0.80	(0.86)	0.91	(0.94)	0.95	(0.97)
1%-Categories	0.77	(0.77)	0.97	(0.99)	0.99	(1.00)	1.00	(1.00)	1.00	(1.00)
State & Population Density										
Top-25%, Bottom-25%, Other	0.34	(0.34)	0.39	(0.41)	0.42	(0.44)	0.48	(0.51)	0.52	(0.55)
20%-Categories	0.39	(0.39)	0.49	(0.53)	0.55	(0.60)	0.62	(0.66)	0.67	(0.71)
15%-Categories	0.43	(0.43)	0.56	(0.60)	0.64	(0.69)	0.73	(0.77)	0.78	(0.82)
10%-Categories	0.49	(0.49)	0.64	(0.68)	0.73	(0.78)	0.81	(0.84)	0.85	(0.88)
5%-Categories	0.59	(0.59)	0.78	(0.82)	0.87	(0.92)	0.95	(0.96)	0.97	(0.98)
1%-Categories	0.86	(0.86)	0.98	(0.99)	1.00	(1.00)	1.00	(1.00)	1.00	(1.00)

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).
Note: The upper-bound of the 95% confidence interval for the average proportion at-risk is presented in parentheses. This estimate is adjusted to account for the complex survey design of sampled variables sets.

Table 1 (cont.): Average Proportion of Geographic Units "At-Risk" of Disclosure, By Spatial Scale and Experimental Traits of Contextual Data

	Tract-Level Contextual Data									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)
National (None)										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
20%-Categories	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)
15%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.05	(0.05)
10%-Categories	0.00	(0.00)	0.00	(0.00)	0.02	(0.02)	0.05	(0.06)	0.14	(0.17)
5%-Categories	0.00	(0.00)	0.01	(0.01)	0.08	(0.10)	0.26	(0.30)	0.47	(0.53)
1%-Categories	0.00	(0.00)	0.22	(0.25)	0.71	(0.78)	0.94	(0.96)	0.98	(1.00)
Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)
20%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.01	(0.02)	0.03	(0.04)
15%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.02)	0.04	(0.04)	0.09	(0.11)
10%-Categories	0.00	(0.00)	0.01	(0.01)	0.04	(0.05)	0.10	(0.12)	0.22	(0.26)
5%-Categories	0.00	(0.00)	0.03	(0.03)	0.16	(0.19)	0.36	(0.42)	0.57	(0.64)
1%-Categories	0.01	(0.01)	0.40	(0.46)	0.81	(0.87)	0.97	(0.98)	0.99	(1.00)
Division										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)
20%-Categories	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.04	(0.05)	0.09	(0.10)
15%-Categories	0.00	(0.00)	0.01	(0.01)	0.04	(0.05)	0.09	(0.11)	0.18	(0.21)
10%-Categories	0.00	(0.00)	0.02	(0.03)	0.09	(0.11)	0.21	(0.24)	0.35	(0.41)
5%-Categories	0.00	(0.00)	0.09	(0.11)	0.31	(0.36)	0.52	(0.58)	0.70	(0.76)
1%-Categories	0.04	(0.04)	0.63	(0.70)	0.90	(0.95)	0.99	(1.00)	1.00	(1.00)
Division & Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.01	(0.01)	0.01	(0.01)	0.02	(0.03)	0.04	(0.05)
20%-Categories	0.00	(0.00)	0.02	(0.02)	0.04	(0.05)	0.08	(0.10)	0.15	(0.18)
15%-Categories	0.00	(0.00)	0.03	(0.04)	0.08	(0.10)	0.16	(0.19)	0.27	(0.32)
10%-Categories	0.01	(0.01)	0.06	(0.07)	0.17	(0.21)	0.31	(0.36)	0.47	(0.53)
5%-Categories	0.02	(0.02)	0.20	(0.23)	0.45	(0.52)	0.64	(0.70)	0.78	(0.84)
1%-Categories	0.12	(0.12)	0.76	(0.81)	0.95	(0.97)	0.99	(1.00)	1.00	(1.00)
State										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.02	(0.02)	0.03	(0.04)	0.05	(0.06)	0.08	(0.09)
20%-Categories	0.01	(0.01)	0.04	(0.05)	0.09	(0.10)	0.14	(0.17)	0.23	(0.27)
15%-Categories	0.01	(0.01)	0.07	(0.08)	0.15	(0.18)	0.25	(0.29)	0.38	(0.43)
10%-Categories	0.02	(0.02)	0.13	(0.16)	0.28	(0.33)	0.42	(0.47)	0.58	(0.64)
5%-Categories	0.06	(0.06)	0.34	(0.39)	0.57	(0.65)	0.74	(0.79)	0.85	(0.90)
1%-Categories	0.27	(0.27)	0.85	(0.89)	0.97	(0.99)	1.00	(1.00)	1.00	(1.00)
State & Population Density										
Top-25%, Bottom-25%, Other	0.01	(0.01)	0.04	(0.04)	0.06	(0.07)	0.09	(0.11)	0.14	(0.16)
20%-Categories	0.03	(0.03)	0.09	(0.10)	0.15	(0.18)	0.22	(0.26)	0.33	(0.38)
15%-Categories	0.04	(0.04)	0.14	(0.16)	0.24	(0.29)	0.35	(0.40)	0.48	(0.55)
10%-Categories	0.06	(0.06)	0.23	(0.27)	0.39	(0.45)	0.53	(0.59)	0.67	(0.74)
5%-Categories	0.13	(0.13)	0.46	(0.52)	0.67	(0.74)	0.83	(0.87)	0.91	(0.94)
1%-Categories	0.44	(0.44)	0.91	(0.94)	0.98	(1.00)	1.00	(1.00)	1.00	(1.00)

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).
Note: The upper-bound of the 95% confidence interval for the average proportion at-risk is presented in parentheses. This estimate is adjusted to account for the complex survey design of sampled variables sets.

Table 1 (cont.): Average Proportion of Geographic Units "At-Risk" of Disclosure, By Spatial Scale and Experimental Traits of Contextual Data

	Blockgroup-Level Contextual Data									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)	Prop.	UB (95%CI)
National (None)										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
20%-Categories	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)
15%-Categories	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.03	(0.03)
10%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.03	(0.03)	0.09	(0.11)
5%-Categories	0.00	(0.00)	0.00	(0.00)	0.04	(0.05)	0.17	(0.20)	0.37	(0.43)
1%-Categories	0.00	(0.00)	0.09	(0.11)	0.55	(0.61)	0.85	(0.88)	0.96	(0.98)
Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
20%-Categories	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)
15%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.05	(0.07)
10%-Categories	0.00	(0.00)	0.00	(0.00)	0.02	(0.02)	0.06	(0.07)	0.16	(0.19)
5%-Categories	0.00	(0.00)	0.01	(0.01)	0.09	(0.11)	0.27	(0.31)	0.48	(0.54)
1%-Categories	0.00	(0.00)	0.22	(0.25)	0.69	(0.75)	0.91	(0.93)	0.98	(0.99)
Division										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)
20%-Categories	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.02	(0.02)	0.05	(0.06)
15%-Categories	0.00	(0.00)	0.00	(0.00)	0.02	(0.02)	0.05	(0.06)	0.12	(0.14)
10%-Categories	0.00	(0.00)	0.01	(0.01)	0.05	(0.06)	0.13	(0.15)	0.26	(0.31)
5%-Categories	0.00	(0.00)	0.04	(0.04)	0.19	(0.23)	0.41	(0.46)	0.61	(0.67)
1%-Categories	0.01	(0.01)	0.43	(0.49)	0.83	(0.88)	0.95	(0.97)	0.99	(1.00)
Division & Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.00	(0.00)	0.01	(0.01)	0.01	(0.01)	0.02	(0.02)
20%-Categories	0.00	(0.00)	0.01	(0.01)	0.02	(0.03)	0.04	(0.05)	0.09	(0.11)
15%-Categories	0.00	(0.00)	0.01	(0.01)	0.04	(0.05)	0.09	(0.11)	0.18	(0.22)
10%-Categories	0.00	(0.00)	0.03	(0.03)	0.10	(0.12)	0.21	(0.24)	0.36	(0.42)
5%-Categories	0.01	(0.01)	0.10	(0.11)	0.31	(0.36)	0.52	(0.57)	0.70	(0.76)
1%-Categories	0.04	(0.04)	0.58	(0.64)	0.90	(0.93)	0.97	(0.99)	0.99	(1.00)
State										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.01	(0.01)	0.01	(0.02)	0.02	(0.03)	0.04	(0.05)
20%-Categories	0.00	(0.00)	0.02	(0.02)	0.05	(0.06)	0.08	(0.10)	0.15	(0.18)
15%-Categories	0.01	(0.01)	0.03	(0.04)	0.09	(0.10)	0.16	(0.19)	0.27	(0.32)
10%-Categories	0.01	(0.01)	0.07	(0.08)	0.18	(0.21)	0.30	(0.35)	0.46	(0.53)
5%-Categories	0.02	(0.02)	0.20	(0.23)	0.43	(0.49)	0.63	(0.68)	0.78	(0.83)
1%-Categories	0.12	(0.12)	0.71	(0.76)	0.94	(0.96)	0.98	(0.99)	1.00	(1.00)
State & Population Density										
Top-25%, Bottom-25%, Other	0.00	(0.00)	0.01	(0.02)	0.03	(0.03)	0.05	(0.05)	0.08	(0.09)
20%-Categories	0.01	(0.01)	0.04	(0.05)	0.08	(0.10)	0.13	(0.16)	0.22	(0.26)
15%-Categories	0.01	(0.01)	0.07	(0.08)	0.14	(0.17)	0.23	(0.27)	0.36	(0.41)
10%-Categories	0.02	(0.02)	0.12	(0.14)	0.26	(0.31)	0.40	(0.45)	0.55	(0.62)
5%-Categories	0.05	(0.05)	0.30	(0.34)	0.54	(0.60)	0.72	(0.77)	0.84	(0.88)
1%-Categories	0.23	(0.23)	0.80	(0.85)	0.96	(0.98)	0.99	(1.00)	1.00	(1.00)

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).
Note: The upper-bound of the 95% confidence interval for the average proportion at-risk is presented in parentheses. This estimate is adjusted to account for the complex survey design of sampled variables sets.

Figure A: Aggregate Disclosure Risk of Geographic Units by Number of Contextual Variables, County-Level

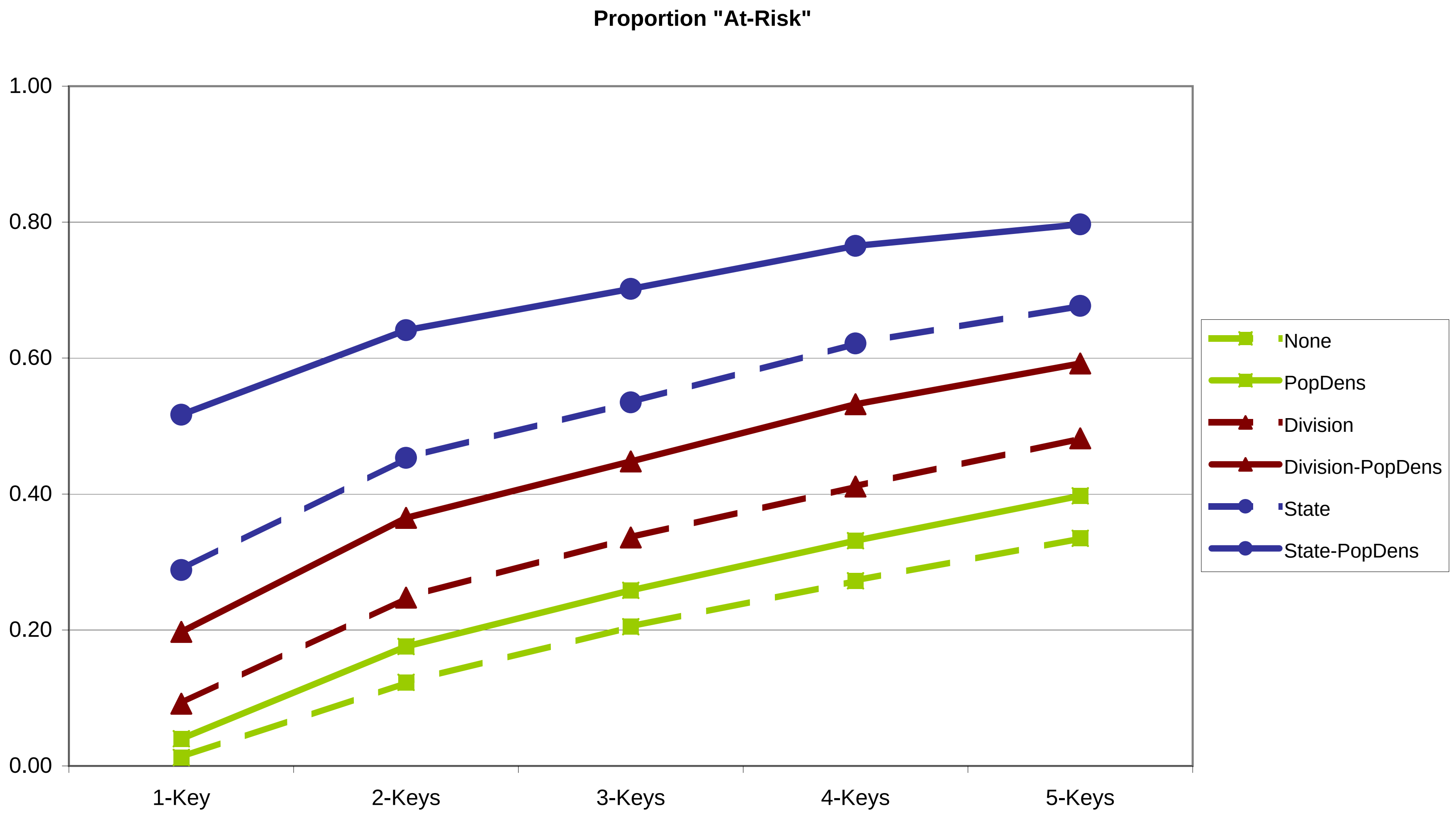

Figure B: Aggregate Disclosure Risk of Geographic Units by Number of Contextual Variables, Tract-Level

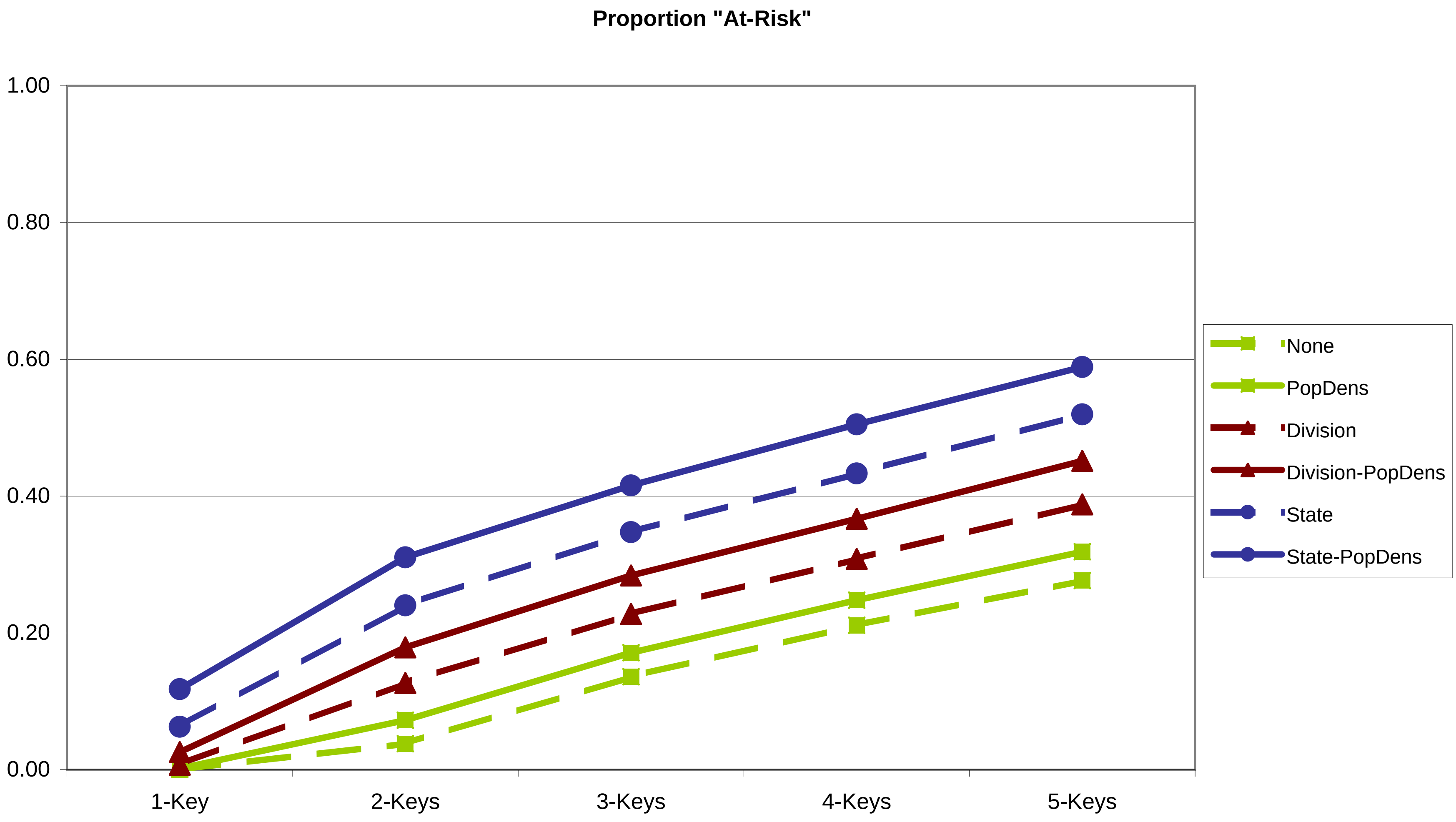

Figure C: Aggregate Disclosure Risk of Geographic Units by Number of Contextual Variables, Blockgroup-Level

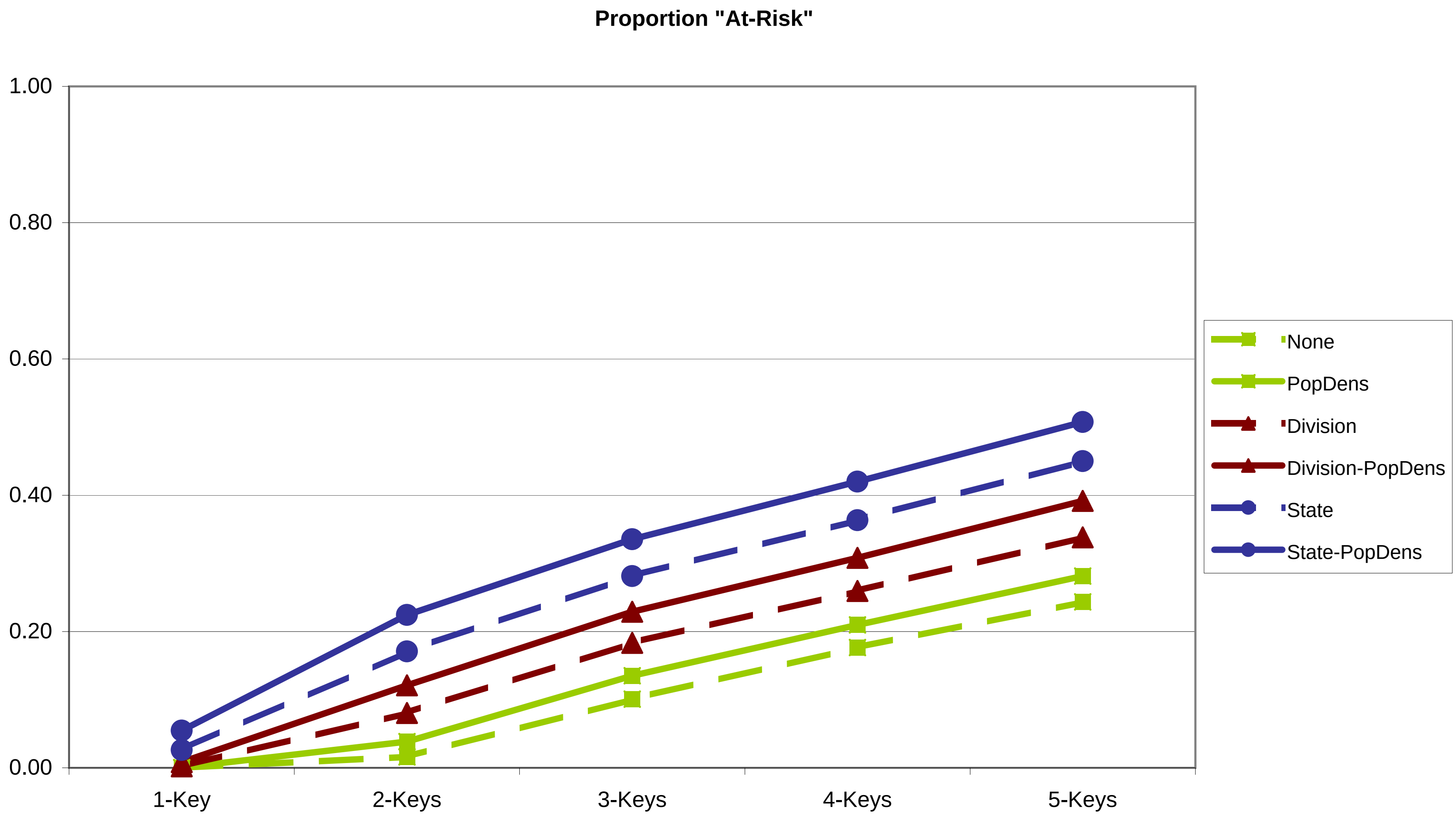

Figure D: Aggregate Disclosure Risk of Geographic Units by Coarseness of Contextual Variables, County-Level

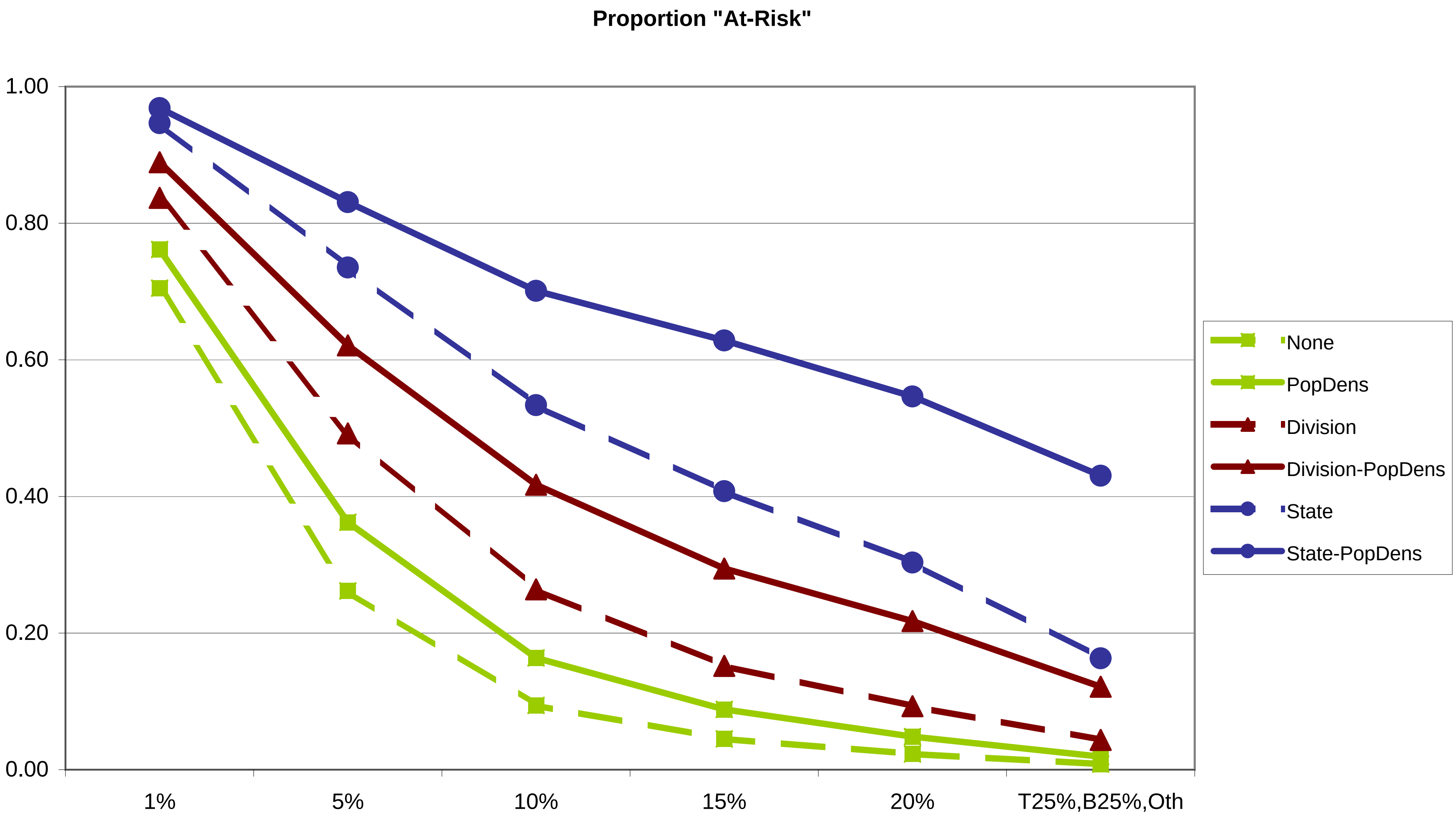

Figure E: Aggregate Disclosure Risk of Geographic Units by Coarseness of Contextual Variables, Tract-Level

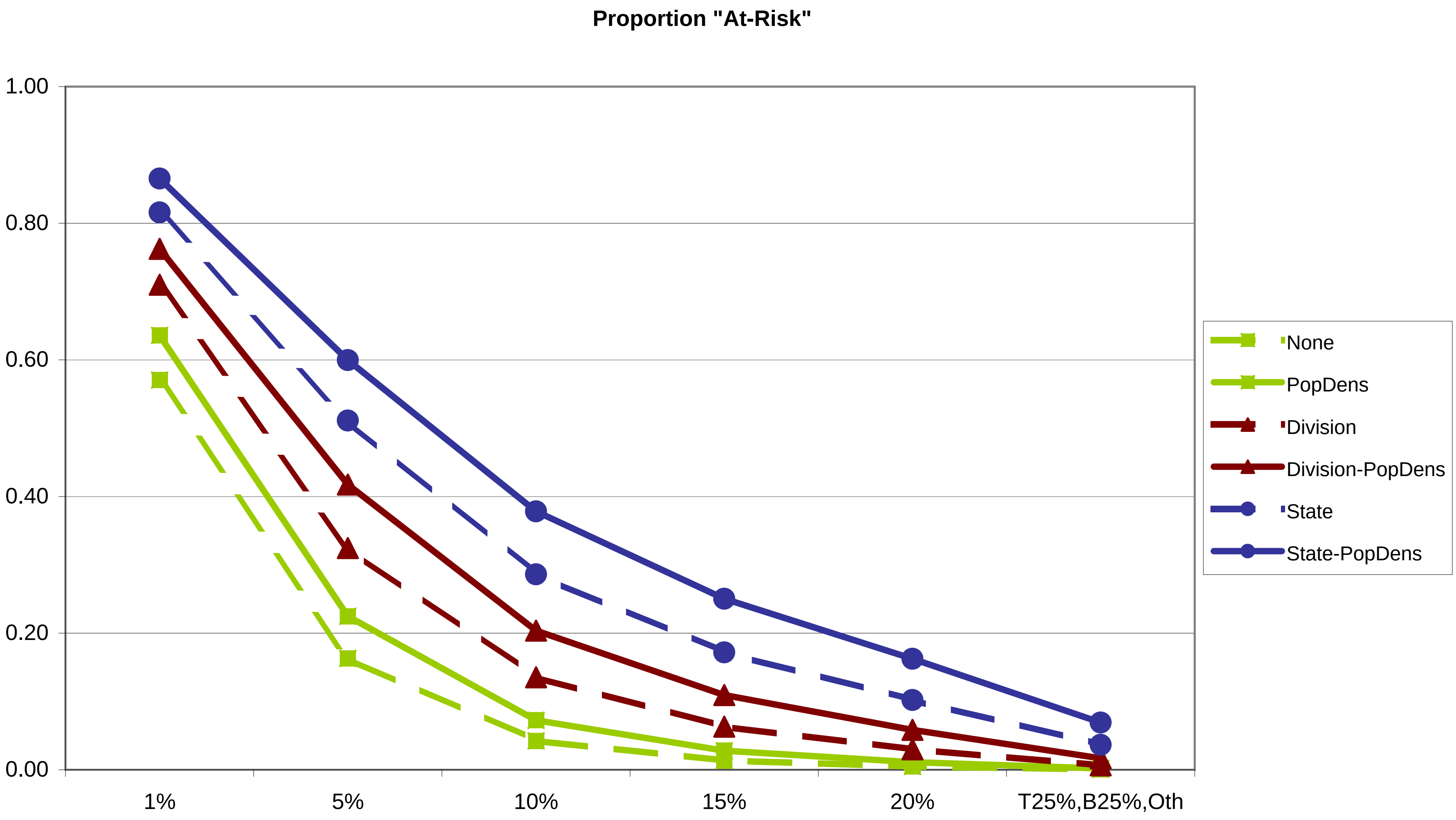

Figure F: Aggregate Disclosure Risk of Geographic Units by Coarseness of Contextual Variables, Blockgroup-Level

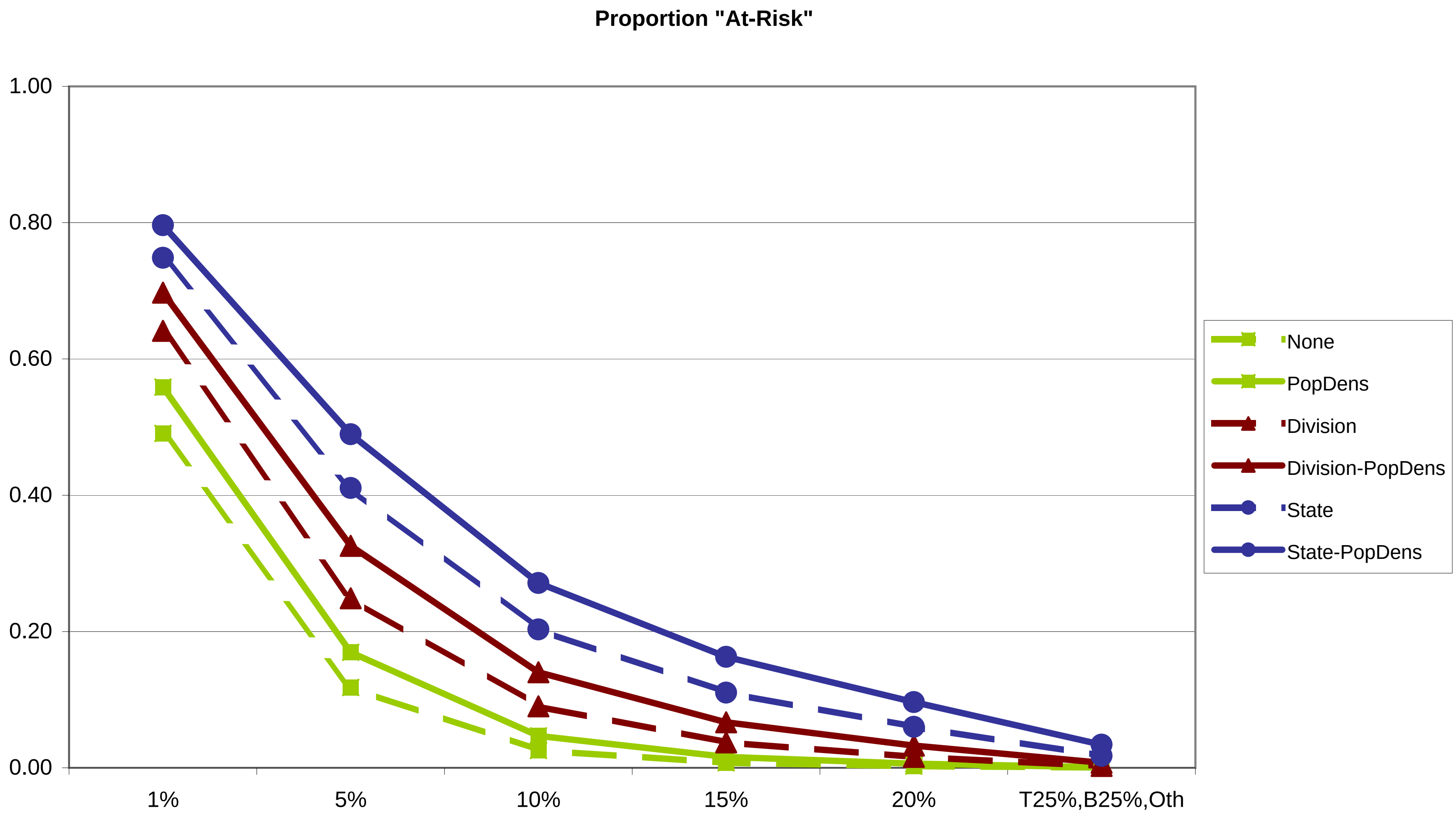

Figure G. Aggregate Disclosure Risk of Geographic Units, Tract-Level

Proportion "At-Risk"

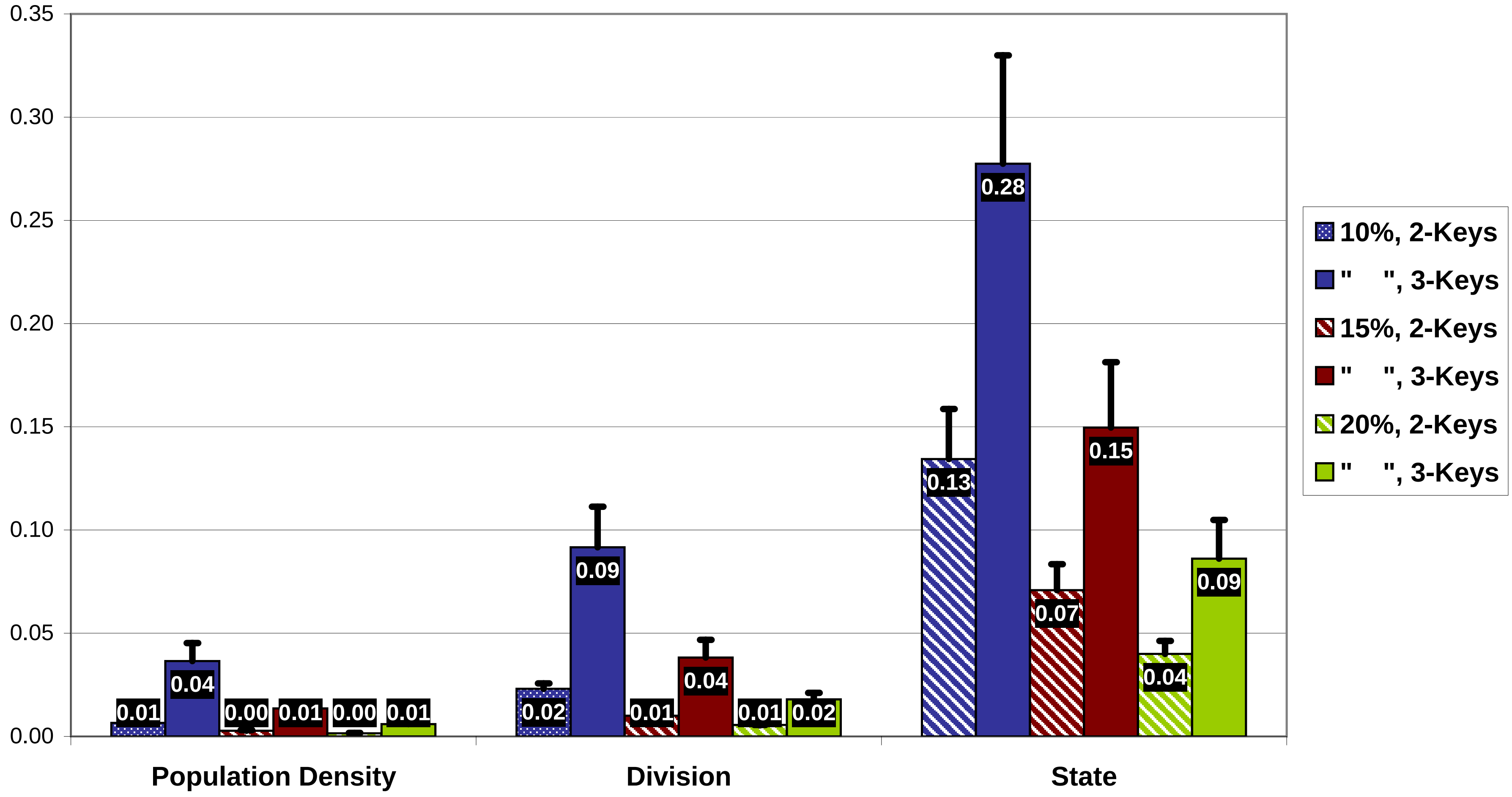

Table A-1: Characteristics of Test Datasets (N=14,796)

Unweighted Averages Across Datasets	Proportion of Datasets
Spatial Scale	
Counties	0.33
Tracts	0.33
Blockgroups	0.33
Geographic Identifiers	
None	0.17
Population Density	0.17
Division	0.17
Division & Population Density	0.17
State	0.17
State & Population Density	0.17
Coarseness	
1%-Categories	0.17
5%-Categories	0.17
10%-Categories	0.17
15%-Categories	0.17
20%-Categories	0.17
Top-25%, Bottom-25%, Other	0.17
Number of Contextual Variables	
1-Key	0.12
2-Keys	0.22
3-Keys	0.22
4-Keys	0.22
5-Keys	0.22
Conceptual Composition	
% Persons, Non-Hispanic White	0.31
% Persons, Non-Hispanic African-American	0.17
% Persons, Non-Hispanic Asian or Pacific Islander	0.07
% Persons, Non-Hispanic Other Race	0.24
% Persons, Hispanic	0.18
% Persons, Foreign-Born	0.17
% Foreign-Born, Naturalized Citizens	0.18
% Households, Linguistically Isolated	0.23
% Persons, In-Poverty	0.19
% Households, With Wage Income	0.18
% Households, Receiving Public Assistance	0.15
% Persons Age 25+, College Degree	0.18
% Families, Female-Headed	0.20
% Persons Age 16-19, Neither Enrolled nor Graduated from High School	0.15
% Housing Units, Owner-Occupied	0.15
% Persons Age 16+, Civilian Labor Force	0.23
% Civilian Labor Force, Unemployed	0.22

Table A-2: Sampling of Synthetic Persons, Resulting Geographic Contexts, and Size of Geographic Unit Populations

Sampling	Total	Sample	Percent
Synthetic Persons	281,421,906	11,562	0.0041
Geographic Units			
Blockgroups	208,125	10,478	5.03
Tracts	65,174	8,947	13.73
Counties	3,141	1,785	56.83
States & Washington DC	51	51	100.00

Population Size	Average	Mininum	Maximum
Counties			
Geographic Identifiers			
None	3,141	3,141	3,141
Population Density	1,047	390	2,294
Division	349	67	618
Division & Population Density	116	15	546
State	62	1	254
State & Population Density	21	0	196
Tracts			
Geographic Identifiers			
None	65,174	65,174	65,174
Population Density	21,725	13,832	36,156
Division	7,242	3,203	11,328
Division & Population Density	2,414	576	7,398
State	1,278	127	7,038
State & Population Density	426	0	5,830
Blockgroups			
Geographic Identifiers			
None	208,125	208,125	208,125
Population Density	69,375	47,646	112,294
Division	23,125	11,006	36,686
Division & Population Density	7,708	1,856	23,347
State	4,081	398	22,066
State & Population Density	1,360	0	17,932

Note: Excludes tracts and blockgroups with no population.

Table A-3: Aggregate Disclosure Risk of Geographic Units in Test Datasets, By Experimental Traits (N=4,932 Datasets At Each Spatial Scale)

Coarseness Based on Absolute Values

	Proportion At-Risk					
	Counties		Tracts		Blockgroups	
Weighted Averages Across Datasets	Mean	(SE)	Mean	(SE)	Mean	(SE)
Total	0.52	(0.01)	0.40	(0.01)	0.34	(0.01)
Geographic Identifiers						
None	0.31	(0.02)	0.25	(0.02)	0.22	(0.02)
Population Density	0.37	(0.02)	0.29	(0.02)	0.26	(0.02)
Division	0.46	(0.02)	0.36	(0.02)	0.31	(0.02)
Division & Population Density	0.57	(0.02)	0.42	(0.02)	0.36	(0.02)
State	0.65	(0.02)	0.49	(0.02)	0.42	(0.02)
State & Population Density	0.78	(0.01)	0.56	(0.02)	0.48	(0.02)
Number of Contextual Variables						
1-Key	0.19	(0.00)	0.04	(0.00)	0.02	(0.00)
2-Keys	0.33	(0.01)	0.16	(0.01)	0.11	(0.01)
3-Keys	0.41	(0.01)	0.26	(0.01)	0.21	(0.01)
4-Keys	0.49	(0.01)	0.35	(0.01)	0.29	(0.01)
5-Keys	0.55	(0.01)	0.42	(0.01)	0.37	(0.01)
Coarseness						
1%-Categories	0.99	(0.00)	0.98	(0.00)	0.96	(0.00)
5%-Categories	0.77	(0.01)	0.66	(0.01)	0.57	(0.01)
10%-Categories	0.53	(0.02)	0.36	(0.01)	0.27	(0.01)
15%-Categories	0.39	(0.02)	0.21	(0.01)	0.14	(0.01)
20%-Categories	0.29	(0.01)	0.12	(0.01)	0.08	(0.01)
Top-25%, Bottom-25%, Other	0.17	(0.01)	0.04	(0.00)	0.02	(0.00)

Coarseness Based on Percentiles

	Proportion At-Risk					
	Counties		Tracts		Blockgroups	
Weighted Averages Across Datasets	Mean	(SE)	Mean	(SE)	Mean	(SE)
Total	0.94	(0.00)	0.75	(0.01)	0.54	(0.01)
Geographic Identifiers						
None	0.84	(0.02)	0.56	(0.02)	0.38	(0.02)
Population Density	0.88	(0.01)	0.65	(0.02)	0.44	(0.02)
Division	0.94	(0.01)	0.74	(0.02)	0.51	(0.02)
Division & Population Density	0.97	(0.00)	0.80	(0.02)	0.57	(0.02)
State	0.99	(0.00)	0.86	(0.01)	0.63	(0.02)
State & Population Density	1.00	(0.00)	0.89	(0.01)	0.68	(0.02)
Number of Contextual Variables						
1-Key	0.39	(0.00)	0.07	(0.00)	0.02	(0.00)
2-Keys	0.69	(0.01)	0.32	(0.01)	0.18	(0.01)
3-Keys	0.85	(0.01)	0.54	(0.01)	0.35	(0.01)
4-Keys	0.92	(0.01)	0.69	(0.01)	0.48	(0.01)
5-Keys	0.95	(0.00)	0.79	(0.01)	0.57	(0.01)
Coarseness						
1%-Categories	1.00	(0.00)	1.00	(0.00)	0.99	(0.00)
5%-Categories	1.00	(0.00)	0.99	(0.00)	0.94	(0.00)
10%-Categories	1.00	(0.00)	0.93	(0.00)	0.78	(0.01)
15%-Categories	0.99	(0.00)	0.80	(0.01)	0.06	(0.00)
20%-Categories	0.96	(0.00)	0.59	(0.02)	0.37	(0.02)
Top-25%, Bottom-25%, Other	0.68	(0.02)	0.19	(0.01)	0.07	(0.01)

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).

Note: Standard errors (in parentheses) are adjusted to account for the complex survey design of sampled variables sets.

Table A-4: Aggregate Disclosure Risk of Geographic Units by Number of Contextual Variables, Comparisons across Spatial Scale and Geographic Identifiers

Average of Predicted Values Across Coarseness	County-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1-Key	0.01	0.04	0.09	0.20	0.29	0.52
2-Keys	0.12	0.18	0.25	0.36	0.45	0.64
3-Keys	0.20	0.26	0.34	0.45	0.53	0.70
4-Keys	0.27	0.33	0.41	0.53	0.62	0.77
5-Keys	0.33	0.40	0.48	0.59	0.68	0.80

Average of Predicted Values Across Coarseness	Tract-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1-Key	0.00	0.00	0.01	0.03	0.06	0.12
2-Keys	0.04	0.07	0.13	0.18	0.24	0.31
3-Keys	0.14	0.17	0.23	0.28	0.35	0.42
4-Keys	0.21	0.25	0.31	0.37	0.43	0.51
5-Keys	0.28	0.32	0.39	0.45	0.52	0.59

Average of Predicted Values Across Coarseness	Blockgroup-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1-Key	0.00	0.00	0.00	0.01	0.03	0.05
2-Keys	0.02	0.04	0.08	0.12	0.17	0.22
3-Keys	0.10	0.14	0.18	0.23	0.28	0.34
4-Keys	0.18	0.21	0.26	0.31	0.36	0.42
5-Keys	0.24	0.28	0.34	0.39	0.45	0.51

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).

Table A-5: Aggregate Disclosure Risk of Geographic Units by Coarseness of Contextual Variables, Comparisons across Spatial Scale and Geographic Identifiers

Average of Predicted Values Across Number of Contextual Keys	County-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1%	0.70	0.76	0.84	0.89	0.95	0.97
5%	0.26	0.36	0.49	0.62	0.74	0.83
10%	0.09	0.16	0.26	0.42	0.53	0.70
15%	0.05	0.09	0.15	0.29	0.41	0.63
20%	0.02	0.05	0.09	0.22	0.30	0.55
T25%,B25%,Oth	0.01	0.02	0.04	0.12	0.16	0.43

Average of Predicted Values Across Number of Contextual Keys	Tract-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1%	0.57	0.64	0.71	0.76	0.82	0.87
5%	0.16	0.22	0.32	0.42	0.51	0.60
10%	0.04	0.07	0.14	0.20	0.29	0.38
15%	0.01	0.03	0.06	0.11	0.17	0.25
20%	0.00	0.01	0.03	0.06	0.10	0.16
T25%,B25%,Oth	0.00	0.00	0.01	0.02	0.04	0.07

Average of Predicted Values Across Number of Contextual Keys	Blockgroup-Level Contextual Data					
	None	Population Density	Division	Division & Population Density	State	State & Population Density
1%	0.49	0.56	0.64	0.70	0.75	0.80
5%	0.12	0.17	0.25	0.33	0.41	0.49
10%	0.03	0.05	0.09	0.14	0.20	0.27
15%	0.01	0.02	0.04	0.07	0.11	0.16
20%	0.00	0.01	0.02	0.03	0.06	0.10
T25%,B25%,Oth	0.00	0.00	0.00	0.01	0.02	0.03

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).