ICPSR
A PARTNER IN
SOCIAL SCIENCE
RESEARCH

INTER-UNIVERSITY
CONSORTIUM FOR
POLITICAL AND
SOCIAL RESEARCH

UNIVERSITY OF MICHIGAN

Working Paper

Disclosure Risk of Geography Attributes: 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, Version 2

January 2009

Disclosure Risk of Geography Attributes: 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. Contextual data is one way to safely release this information without identifying the location of survey respondents. Analyzing an array of geography attributes, I conduct reidentification experiments for 14,796 simulated datasets to measure the likelihood of pinpointing geographic locations under alternative database designs, relating to: (1) the spatial scale of standard geographies, as determined by the areal size of these administrative units; (2) the scope of study, as determined by the identification of division, state, and MSA-status; (3) the number of geography attributes provided in a dataset; and (4) and coarseness of these contextual measures, as determined by global recoding schema. Using the "data file" as my unit of analysis, the number of geographic units resembling a study location 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 where two distinct scenarios of intruder search behavior are assumed.

Key Words: anonymity, 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 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.

1. Introduction

Many subjects of contemporary social science lend themselves to an analysis in which the individuals under study are placed in their context, particularly a context that can be defined spatially, such as a street, block, town, county, or some other geographic unit. Advances in computer processing power and in the sophistication of analytical tools have facilitated the analysis of large amounts of information, leading to a continuously growing demand for access to new sources of data. The scientific precepts of transparency and replication have further encouraged investigators to distribute their data so that others can verify and build on published findings. Yet a delicate balance must be struck between providing easy access to existing data and guaranteeing privacy to subjects who agree to take part in scientific studies (National Research Council 2007; VanWey et al. 2005). The problem is especially acute when researchers collect information about the geographic location of study participants, as spatial indicators can provide unambiguous clues about the identity of a respondent (Gutmann et al. 2009; National Research Council 2007; Rushton et al. 2006; Armstrong, Rushton, and Zimmerman, 1999; Saalfeld, et. al, 1992). Consequently investigators have become increasingly interested in devising procedures for releasing geographically-rich microdata that protect both the utility of the collected information and the privacy of subjects.

Data producers have found two ways of providing this information, either directly identifying the spatial unit or releasing only the characteristics of these respondent locations. This paper is concerned with this second case, where the *contextualized microdata* file consists of records for study subjects that include both their personal characteristics (e.g., age of respondent) and the attributes of their geographic location (e.g., proportion of population in respondent's neighborhood that is poor). One reason for providing geographic attributes or *contextual* data, rather than the identity of the spatial unit, is that doing so makes it more difficult to pinpoint where the respondent lives (Armstrong, Rushton, and Zimmerman 1999). As a result, intruders must search for a target subject within an enlarged population of individuals virtually amassed across geographies resembling one another. But when the combination of contextual information about a sampled spatial unit is rare among geographies of that type – then identification is more likely, rather than less (Saalfeld, Zayatz, and Hoel 1992). Hence producers must carefully construct measures of geographic attributes (along with the personal characteristics of subjects) so as to maintain the confidentiality of respondent locations and their personal identities, while at the same time ensuring that the data have the maximum analytic value for the broadest user group.

For this study, I compile nearly 15,000 data files composed of geographic-unit records containing a variety of contextual data and measure the likelihood of pinpointing locations under alternative database designs, relating to: (1) the spatial scale of standard geographies, as determined by the areal size of these administrative units; (2) the scope of study, as determined by the identification of division, state, and MSA-status; (3) the number of geography attributes provided in a dataset; and (4) and coarseness of these contextual measures, as determined by global recoding schema. Using the "data file" as my unit of analysis, the number of geographic units resembling a study location 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 where two distinct scenarios of intruder search behavior are assumed.

In the following section (Section 2), I present an overview of how disclosure risk is influenced by releasing geographic attributes onto microdata files and the implications for database design. For the remainder of the paper, I present my study's methodology (Section 3), the results of my analyses (Section 4), and my conclusions (Section 5).

2. Disclosure Risk and the Design of Contextualized Microdata

Three studies have followed the 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 and nearby sampled households (Energy Information Administration 2001). 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). Lastly, Franconi and Stander (2002) obscured the location of a sample of enterprises by assigning them to one of two regions with distinct contextual characteristics, as revealed by a principal components analysis of the geographic attributes for eight Italian territorial units. Because their populations are either particularly conspicuous or geographically constrained, all of these studies utilize perturbative masking techniques in the creation of a single geographic attribute for release.

While these applications are insightful, more research is needed to comprehensively assess disclosure risk as it varies with different design elements of contextualized microdata. As described by Skinner (2007), research assessing identity disclosure may be grouped into two bodies, that of "Population Unique" studies and "Reidentification Probabilities" (Skinner and Holmes 1998). Playing a central role or laying the foundation of different anonymity factors, one important component of disclosure risk is shared by both approaches: the number of units in the population resembling those drawn into a study. Both bodies of research assume that the geographic location of this population is directly known. But when contextual data are released instead, two such frequency components of disclosure risk must be considered that are based on nested populations. Second-level populations of geographic locations bounded by the scope of study, as well as first-level populations of individuals (i.e., persons or establishments) encompassed within these higher-order units, jointly influence the likelihood that respondents are reidentified.

Data producers may reduce the likelihood of respondents being correctly reidentified by fostering the aggregation of first-level populations across locations having similar contextual characteristics. Hence my study's outcome of interest is the degree to which contextual data extends the intruder search across disparate locations and their populations. I then discuss the decisions involved in creating measures of geography attributes that are both anonymizing and of high analytic value, where the same factors determining disclosure risk also influence the utility of information.

2. a. Measurement and Determinants of Disclosure Risk

Several aspects of study design have serious implications for the disclosure of respondent identities. Studies that collect an abundance of detailed information (e.g., longitudinal data, social networks), whose research subjects are drawn from a small or otherwise exceptionally visible population, or that are geographically specific are often highly risky. For a dataset that identifies the geographic location of respondents (e.g., county name), the likelihood of correctly reidentifying the target respondent is then a function of the number of individuals (e.g., persons, establishments) in the known area's population having similar identifying personal attributes or *individual keys* (e.g., gender, number of employees). Given the chosen subjects of a study characterized by a particular set of individual keys, the number of *look-alike individuals* (or *twins*) grows with the size of populations bounded within geographic units (Greenberg and Voshell 1990; Zayatz 1991). Therefore when the land area of known geographies is expanded, or when their *spatial scale* is increased, it becomes more difficult to reidentify respondents drawn from the field's enlarged population.

Instead of releasing identifiers for sufficiently large geographies, producers may choose to release the attributes of second-level units. Contextual measures can be either constructed for standard geographic units that surround sets of individuals, or created for a spatial window uniquely assembled around each subject. Saafeld, Zayatz, and Hoel (1992) argue that windowed contexts make the reidentification of locations much more difficult, while standard contexts pose an insurmountable amount of disclosure risk since these geographies are far more limited in number and well-known. Given the abundance of spatial information that can be easily attached to microdata and the need for more confirmatory research, I have chosen to assess contextual data derived from standard geographies.

When microdata contain the attributes of standard geographies (i.e., *geographic keys*), an intruder must consider all locations in the population sharing the same contextual characteristics as surveyed locations, defined as *look-alike geographies*. As with populations of individuals, the number of look-alike geographies found in these higher-order populations determines the likelihood of pinpointing a surveyed location. But in contrast to first-level populations, the spatial scale of geographies underlying contextual measures is negatively associated with the number of look-alike geographies. Large scale localities carve up finite space into expansive land areas, resulting in a geographic population with few members. However small scale localities are much more numerous, increasing the possibility of finding additional look-alike geographies.

Reflecting the two-fold risk dimension of the areal size of geographies, large-scale contextual measures foster the accumulation of look-alike individuals as indicated by the dramatic growth in the search population with a one-unit increase in look-alike geographies. In contrast, considerably more small-scale geographies are required to gather the same number of look-alike individuals. Therefore the importance of the number of look-alike geographies – for offsetting disclosure risk from each unit's small population – varies with the spatial scale of contextual measures.

$$F = L * W_S \qquad (1)$$

$$L = \frac{1}{n} \sum_{J=1}^{n} (Y_{J|j} * P_j) \qquad (2)$$

$$P_j = f(S, G, K, M) \qquad (3)$$

Where:

F = Extension of full intruder search with the release of "contextual" attributes, through the decline in "pinpointed-respondent" probabilities

W_S = Spatial scale weight, benchmarked to average size of human populations in U. S. counties

L = Extension of limited intruder search with the release of "contextual" attributes, through the decline in "pinpointed-location" probabilities

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

P_j = Population counts of geographies having combination of "contextual" attributes (j)

S = Spatial scale of geographies

G = Scope of study determined by identified geography

K = Number of geography attributes

M = Measurement coarseness of geography attributes

Identity disclosure is driven by the number of look-alike geographies in two ways that depend on intruder search behaviors. When costs are sufficiently high, intruders are likely to conduct a *limited search* by looking for respondents within a subset of all possible locations. As a result, disclosure risk is primarily a function of the proportion of look-alike geographies that are likely explored. Utilizing data from studies that directly identify the spatial units where respondents live, intruders have a 100% chance of perusing surveyed locations. This *pinpointed-location probability* declines at a consistent rate with a one-unit increase in the number of look-alike geographies, regardless of their spatial scale. Representing the denominator for this probability, the degree to which contextual data extends a limited intruder search is indicated by the average population count of look-alike geographies for respondent locations (i.e., L).

I estimate this risk component by tallying the number of counties, tracts, and blockgroups within the geographic population that resemble my sampled locations (i.e., P_j). The distribution of contextual variables characterizing surveyed geographies (i.e., $Y_{J|j}$) is determined by my experiment's sampling methodology. Using simple combinations of contextual variables derived from perfectly accurate information, the number of look-alike geographies within the population depends on four characteristics of a dataset: the spatial scale of context (S), geographic identifiers (G), the number of contextual variables (K), and the coarseness of their measures (M).

A conservative assessment of risk assumes that intruders will conduct a *full search* by extensively looking for a respondent within all pertinent geographies and their populations, setting the pinpointed-location probability back to 100%. As a result, disclosure risk then becomes a function of the number of look-alike individuals compiled across all look-alike geographies. The rate at which look-alike geographies contribute to this *pinpointed-respondent probability* depends on the concentration of lower-level populations within the spatial unit and systematically varies with their spatial scale. Facilitating comparisons between database designs, I weight the average number of look-alike geographies (i.e., L) by their ability to accumulate first-level populations (i.e., W_S) as benchmarked to a common population-size standard. By standardizing the number of look-alike geographies, my second outcome measure then indicates the degree to which contextual data extends the full intruder search as measured by the multiplicative factor (i.e., F) associated with the size of first-level populations within the benchmarking area.

In creating the above multiplier F, I allow for the absolute size of study populations to vary since researchers ponder a variety of subjects within their spatial context. Furthermore, in constructing the weight W_S , I assume that the spatial distribution of specialized human and establishment populations mirror that of the general human population. There are dramatic differences in the size of human populations within U.S. counties, tracts, and blockgroups, with an average of 89,596; 4,318; and 1,352 persons (respectively). Weighing the original second-level population counts by the relative size of first-level populations, I multiply tract and blockgroup values by 0.0482 and 0.0151 respectively, such that these estimates are benchmarked to county-level counts (i.e., 4,318 and 1,352 divided by 89,596).

2. b. Decisions Involved in Creating Safe and Useful Contextual Data

When designing a dataset, the producer seeks to release as much information as possible while ensuring that subjects face minimal disclosure risk. After choosing the mode of distributing data files to the user community and considering associated intruder behavior, he must first define what constitutes anonymity for survey respondents by selecting the minimal probability of correct reidentification. Given this definition of risk, he estimates the likelihood of subjects being reidentified for a variety of database configurations. The producer then makes a wish list of data utility characteristics by prioritizing (1) the geographic scope of study or the release of geographic identifiers, (2) the scale of contextual variables, (3) the number of these geography attributes, and (4) their measurement detail. Subsequent decisions seeking to minimize risk can then be guided by these priorities.

Setting the stage for the intruder search, the designated geographic area of study is usually bounded within nation-states and is further restricted by the direct identification of regions or by the initial confinement to a particular area (e.g., small-area studies). If geographic attributes can adequately capture important spatial determinants, contextual information is a likely alternative to releasing of direct geographic identifiers. In turn, choosing the scale of geographic attributes is of paramount concern when compiling contextual information for release. This decision depends on the most likely research questions to be answered with the data file and the nature of spatial processes to be studied. Producers must carefully consider whether a particular scale of context may be too large of an aggregation, smoothing over environmental patterns better captured by smaller geographies.

While small-scale contexts may provide the most analytically useful information, a higher number of look-alike geographies is required to offset risk from their lesser populations. Small geographic units, which are large in number, offer more opportunities to locate look-alike units or matches. Hence the ability to pinpoint geographic units declines with their areal size, given the high probability of finding multiple matches. However the anonymity benefits of a relatively large number of potential matches is offset by another statistical artifact. Because space is delineated into units that are relatively small in area, large in number, and heterogeneous, small-scale locations exhibit considerably more variation in contextual characteristics, increasing the chance of identifying unique contexts. Bringing together these factors, the selection of contextual scale requires complex decisions that carefully consider intruder behavior.

Data producers can further influence the accumulation of look-alike geographies (and individuals therein) by carefully selecting the number and measurement detail of contextual variables. The analytical utility of data generally increases with these design elements. However locations are generally more easily reidentified in datasets with relatively large numbers of geographic attributes. The amount of risk resulting from these keys depends on the coarseness of their measurement.

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 two important techniques of statistical disclosure control should be considered, namely *global recoding* and *local suppression* (DeWaal and Willenborg 1995, 1996). Aggregating continuous measures into various levels of coarseness, utilizing global recoding schema, decreases the likelihood that locations are reidentified. But for geographic units that remain easily pinpointed, their contextual characteristics are not to be released on a microdata file and are locally suppressed.

Synthetic data techniques can then be used to construct contextual information that replaces a particular missing value or to formulate a completely new set of measures. Fewer measurement categories and high suppression rates increase the amount of information lost, while large amounts of ascribed data may distort analyses. Hence a producer needs to consider how data utility varies with these methods and whether a group of geographic units is particularly affected by these aberrations.

3. Assessing Disclosure Risk of Masked Contextual Data

Informing the above decision-making process, I conduct experiments to assess the amount of disclosure risk associated with a dataset's contextual data (Domingo-Ferrer and Torra, 2001a). In doing so, I followed seven methodological steps:

- draw a sample of contextual 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 study subjects;
- 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 two measures of disclosure risk components for each test microdata file; and

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

3. a. Sources, Measures, and Sampled Sets of Contextual Variables

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 demographic concepts listed in my Appendix Table A-2.

The conceptual content of datasets varies with (1) the number of contextual variables (i.e., k = 1, 2, 3, 4, 5) and (2) which contextual variables are included in the sets. All possible sets of contextual variables are constructed for those 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 contextual 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 contextual variables composed of one to five concepts. All seventeen variable sets with a single concept are included in my study. Thirty datasets are randomly sampled from each stratum of the multiple-concept variable sets. I expect that my results will be robust whether my experiment is composed of 30 or 300 variables sets since the statistical properties of contextual information packages remain fairly constant (see Appendix Table 1-A for further details).

3. b. Experimental Traits

Datasets ($C_{b,s,g,m}$) are varied along the [B x S x G x M] matrix of: (1) 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 sets consisting of five contextual 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).[1]

Given these contextual variable sets, I then constrain my matching process by geographic identifiers released in the dataset. A dataset can directly identify the state and region of respondent locations, where 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.

The population density of unidentified geographies is a contextual variable that also broadly confines the scope of study. 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. Given its general analytical appeal, I consider MSA-status as a "pseudo" direct geographic identifier and analyze how this element determines risk separately from and jointly with other contextual variables.

Finally I systematically vary the amount of measurement detail across my experimental datasets; thereby assessing how these risk components fluctuate with global recoding schema. Geographic units are considered *outliers* when their attributes place them within the top and

[1] Presented in Appendix Table A-2, equal proportions of datasets across categories of spatial scale, geographic identifiers, and masking techniques reflect my experimental design. However, the proportions of datasets across categories of the number and conceptual composition of contextual variables reflect the random sampling of variables set, stratified by the number of keys (i.e., including all 17 sets with 1 key; including 30 sets each with 2, 3, 4, and 5 keys). Every conceptual domain 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.

bottom 0.5% of their population's distribution (Zayatz, 2005), defined by each dataset's identified geography. These extreme attribute values typically lead to unique geographies (i.e., population count equal to "1"), requiring that a value representing the upper or lower bound of the distribution be assigned. 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). 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%.[2]

[Exhibit 1 Here]

[2] 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-4, 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 – have a larger number of look-alike geographies with percentile coarsening. However, counties sharing a relatively common characteristic – those within the middle of the distribution – have a smaller number of look-alikes 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.

3. 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. I have chosen the block as my sampling unit because it most closely approximates the residential location of our theoretically ideal sample of individual study subjects (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-3, 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.

3. d. Generalized Estimates of Risk Components

As detailed in Section 2.a., the foundation of my measures of disclosure risk components (i.e., F and L) is the number of geographies within a second-level population sharing a particular set of contextual characteristics (i.e., P_j). These estimates are derived from a single sample of locations associated with a survey of respondents (i.e., $Y_{J|j}$), instead of drawing separate samples of counties, tracts, and blockgroups.

Since geographic attributes of surveyed locations are not perturbed, identifying matches is exact. Consequently I simply count the number of geographic units in the population file sharing the same attributes as units found in a test database, as determined by selected design elements (Winkler 2004). Analyzing metadata characterizing my sample of 14,796 datasets (see Appendix Table A-4 for further details), I produce approximations of my two outcome measures that are generalized to all possible collections. In doing so, I provide point and interval estimates of the degree to which limited and full intruder search efforts are extended with contextual data 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).

4. Presentation of Results

In turn I present estimates of these risk components in matrices that represent a tool for creating contextualized microdata. Encapsulating the complexity of results, I also produce a series of summary figures and provide a specific example of how this information may be integrated into the design of datasets. These experimental findings directly apply only to studies within the U.S. context, as the construction of such geographies in other countries is very different.

4. a. Tool for Designing Contextual Data

In Tables 1 and 2, I present the average unstandardized and standardized number of look-alike geographic units for each dataset typology. [3] I also present the lower bound of these estimates in that the largest, probable amount of risk is also of concern when designing public-use datasets. These tables are 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, risk tends to increase as one reads from left-to-right and from top-to-bottom.[4]

[Tables 1 and 2]

[3] Confidence intervals for extreme measurement values 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 matching algorithm. It is easy to predict that the number of look-alike geographies will be close to “1” 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 these outcomes (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 my risk components. This variation should be considered when using this study’s results for designing datasets.

[4] For those wishing to fine-tune results presented in Supplemental Table 2, Supplemental Table 1’s estimates can be associated with alternative spatial scale weights that are constructed for particular study populations, thereby better reflecting their absolute size and spatial distribution.

4. b. Likelihood of Pinpointing Sampled Locations in a Limited Intruder Search

Summarizing Table 1 results, the spatial scale of contexts and the scope of study dramatically influence the chance of pinpointing respondent locations when conducting a limited search. For contexts characterized by 1 to 5 attributes at 6 coarseness-levels, there is a 1-in-1,082 average chance of correctly selecting a surveyed county from those across the nation. Dividing space among larger sets of tracts and blockgroups, uncertainty drops to less than a tenth of this amount (i.e., 46,709 and 15,001 look-alike blockgroups and tracts across the nation).

Placing these geographic units into searchable subsets, risk dramatically rises again when the scope of study is constrained. For instance, collections identifying the state and MSA-status have pinpointed-location probabilities whose magnitudes are over 50 times that of national databases (i.e., 20, 200, and 565 look-alike counties, tracts, and blockgroups). Besides determining the absolute number of look-alike units (i.e., the denominator of pinpointed-location probability), the scope of study also influences whether all units are searched (i.e., the numerator of this probability). If we assume that the ease of compiling identifying information is spatially clustered, search costs should be relatively lower for finite areas, increasing the likelihood of an extensive exploration that heightens disclosure risk.

However little is known about factors that shape intruder behaviors; and, therefore, a more conservative assessment of risk is warranted. Assuming that all look-alike geographies will be searched, it then becomes necessary to assess whether the aggregation of first-level populations can offset risk from an exceptionally small number of look-alike geographies.

4. c. Likelihood of Pinpointing Respondents in a Full Intruder Search

For all spatial scales of context, the uncertainty associated with a full-search generally declines with the release of any direct geographic identifier. Summarizing Table 2 results in Figures 1 to 4, an average of 1,082; 723; and 705 standardized look-alike counties are identified when national databases release between 1 to 5 geographic attributes at 6 coarseness-levels of county-, tract-, and blockgroup-level contexts (respectively). Illustrated in Figure 1, knowledge of the population density (i.e., MSA-status) of contexts typically reduces the size of aggregated populations by 48, 62, and 63% (across spatial scales). Disclosure risk continues to rise dramatically with the release of division only, resulting in an 85 to 88% reduction in standardized look-alike counties. As compared to national datasets, the ability to aggregate anonymizing individuals falls by 99% when the state and population density of surveyed locations is known. While the protection offered by contextual data has dwindled significantly, these spatially constrained collections can still compile first-level populations across an average of 9, 10, or 20 standardized geographies (for blockgroup-, tract-, and county-level contexts, respectively). However these aggregation rates depend on the number and coarseness of contextual measures as well as the absolute size of individual populations.

[Figures 1, 2, 3, and 4]

Looking at Figures 2 and 3, the amount of risk associated with contextual information generally increases with the number of geography keys and their measurement detail. The largest increase in risk occurs with the addition of a second attribute, for contexts of any spatial scale across all scopes of study (Figure 2), with a 32% reduction in the aggregation of individual populations from county-level contextual data and an even more pronounced decline of 44% for tract- and blockgroup-level contexts. Releasing a third and a fourth attribute also decreases uncertainty but at a moderate rate (13 to 17% decrements); while a fifth key has relatively little effect on risk (7% decrement).

Seeking to offset risk associated with the number of contextual keys, we see from Figure 3 that the amount of uncertainty rises by 362 to 467% when the most detailed measures represented in 1%-categories are collapsed into 5%-categories. With an 80% reduction in the number of measurement categories (i.e., from 100 to 20 metric spaces), collections are able to achieve an average of 82 to 156 standardized counties. For the remaining recoding schemes, the aggregation of first-level populations follows a consistent increasing pattern, mirroring the fall in measurement detail. By lowering the number of measurement categories from 100 to 10, 7, 5 and 3, uncertainty rises by approximately 800; 1,100; 1,400; and 1,900% (respectively).

4. c. Example of Decision-Making Process

The U.S. Census Bureau has initiated dissemination practices that allow for highly populated locations to be identified on microdata files (Cox et al. 1986; U.S. Census Bureau 2001, 2003b). The foundation of these guidelines is the *population-size threshold* that indicates: the minimal size of a population required to produce a sufficiently small number of individuals with unique characteristics (within said population) who are subsequently drawn into a survey (Greenberg and Voshell 1990; Zayatz 1991). For instance, the Census Bureau has found that risk falls to a negligible level when there are 100,000 or more persons in the population. As a result, the Current Population Survey has been allowed to directly identify such densely populated counties in their public-use files.

But let us consider a hypothetical cross-sectional survey with a similar set of personal identifiers and sampling rate, whose study population is composed of persons aged 65 years and over (rather than the civilian non-institutional population 15 years of age and older).[5] Based on the 2000 decennial census, an average of 11,140 persons from this specialized population resides within a county. Reflecting upon the above rule-of-thumb as applied to a full intruder search, it would require approximately 9 standard look-alike counties (representing 186 tracts or 595 blockgroups) to reach 100,000 members of this aged population, thereby ensuring the confidentiality for approximately 17,500 respondents.

Developing a microdata file for this study population that contains blockgroup-level contextual data, I review Table 2 to identify design elements that are likely to yield a sufficient amount of uncertainty. Ideally I would like to release four or five contextual variables as well as a geographic identifier; therefore I consider the trade-offs in risk with identifying division alone, division and population density, and state alone. Furthermore I investigate how risk will be offset by aggregating my measures into 10%-, 15%- and 20%- categories.

Presented in Figure 4, three optimal designs are likely to produce the 9 standard counties required to meet our population-size threshold: (1) division, 5-keys, 10%-categories; (2) division and population density, 5-keys, 15%-categories; and (3) state, 4-keys, 20%-categories. The final choice among these schemes ultimately hinges on the analytical value of the measures and the scientific relevance of the contextual content.

[5] Typically the Current Population Survey samples roughly 50,000 households. With 2000 census counts of 105,480,101 households, this translates into an approximate coverage (or sampling) rate of 0.000474.

5. Conclusions

This research assesses on how different design elements of contextual information influence a study's geographic specificity and, consequently, its level of disclosure risk. Because this is an analysis of spatial units defined and characterized by United States' administrative systems and its population, results can be directly applied to U.S. studies. The lessons learned may also be of use to a much wider audience since the fundamental components of disclosure risk of contextualized microdata (based on standard units) are abstractions that can be empirically characterized and modeled. The size of first- and second- level populations, coupled with measures summarizing distributions of their identifying characteristics, are important variables that can predict risk for a variety of individuals nested within a variety of environments. But as I describe below, there are contingencies not represented in my experimental design that indicate the need for further research.

First of all, my study's synthetic set of respondents is (in essence) a random sample of persons scattered across each of the U.S. states, reflecting the areal distribution of the population. This approach provides a more comprehensive sample of contexts and enhances the robustness of my estimates. While appropriate for this study, my sampling methodology does not reflect how survey data is typically collected. In reality, surveys are likely to have a clustered design that result in public-use data files containing indicators of the strata and clusters from which respondents are drawn. Although my current study does not directly consider these methodological indicators, a nationally representative two-stage cluster design is closely akin to a study that has released: (1) state identifiers, representing strata; (2) MSA-status, representing the first cluster; and (3) another geographic attribute characterizing the area from which a second cluster is drawn from (e.g., high- and low-income schools).

While these strata and clusters may remain unnamed, disclosure risk may be heightened by estimating the likelihood that a sampled cluster matches a particular set of geographies in the population as a function of contextual keys. The search for look-alike geographies can then be limited to those within newly pinpointed clusters. The accuracy of these posterior probabilities depends on the degree to which primary sampling units (PSU) and contexts overlap or are bounded within one another.

The spatial intersection between PSUs and selected contexts also increases risk by limiting the aggregation of first-level populations. Sampling units may take the form of blended geographies, where respondents are drawn across administrative units. Subtracting overlapping areas, slivers of geography can emerge that encompass a small subset of the population. Assessing risk in terms of person-counts, Duke-Williams and Rees' (1998) found that these *geographic slivers* significantly increased the chances of pinpointing respondents residing within extemporaneous boundaries. Therefore, if PSUs do not have exactly the same contexts, a clustered sampling design may fail to aggregate a sufficient number of individuals with shared personal and contextual characteristics, regardless of the number of look-alike geographies.

Besides assessing how risk is shaped by a study's sampling methodology, it is also important to consider other types of contextual variables. The current study analyzes a sample of contextual variable sets, where all keys share the same level of coarseness. While these simplifications were necessary for the initial stages of this work, further research needs to be conducted to (1) closely study how specific contextual measures (e.g. % persons, in-poverty) shape disclosure risk; (2) identify more optimal recoding schemes; and (3) quantify the utility of masked contextual data by measuring amounts of information lost and suppression bias (e.g., Domingo-Ferrer and Torra, 2001a and 2001b; Raghunathan, et. al. 2003; Winkler 2004).

Furthermore this study only considers census-based contextual variables that distinguish standard geographic-units independently of one other. However geographic attributes may also be spatially or temporally linked, consisting of: (1) a single location characterized by contextual variables of differing spatial scales; (2) a single respondent associated with two locations characterized by contextual variables; and (3) a single location characterized by contextual variables gathered at multiple points in time. These dependencies appreciably cap the number of look-alike geographies as compared to datasets having the same number of independent geographic attributes.

Producers may also want to release non-census based contextual variables, such as distance of respondent's home from a river or the average rainfall of the immediate area, along with contextual measures for standard spatial units. Since both kinds of information can be mapped onto the same spatial surface, an intruder's search can be further narrowed to geographies that share all of these demographic, topological, and climatic characteristics.

Although I have only discussed the role of geographic attributes, respondent characteristics may also be used to pinpoint surveyed locations. A new set of approximated contextual measures can be derived from the personal characteristics of sampled individuals within shared contexts. As an example of such an indicator, let us consider respondents living in unidentified counties with 50 to 59% its population in-poverty, where we calculate the proportion of these sampled persons who are 18 years of age. The accuracy of these supplemental keys and their subsequent contributions to the reidentification process is enhanced by clustered sampling designs that increase the numbers of respondents sharing the same geographic attributes.

6. References

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

Cox, Lawrence H., Sarah-Kathryn McDonald, and Dawn Nelson. 1986. Confidentiality Issues at the United States Bureau of the Census. *Journal of Official Statistics* 2(2): 135-160.

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.

------. 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.

------. 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.

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

Franconi, Luisa and Julian Stander. 2002. A model-based method for disclosure limitation of business microdata. *The Statistician* 51(1): 51-61.

Greenberg, Brian and Laura Voshell. 1990. Two Notes on Relating the Risk of Disclosure for Microdata and Geographic Area Size. *Survey of Income and Program Participation Working Paper Series* (#9029). Washington, DC: U.S. Census Bureau.

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.

National Research Council. 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. M.P. Gutmann and P.C. Stern, Eds. Committee on the Human Dimensions of Global Change. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

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.

Skinner, C.J. 2007. The probability of identification: Applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society (Series A)* 170(1): 195-212.

Skinner, C.J. and D.J. Holmes. 1998. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* 14(4): 361-372.

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.

U.S. Census Bureau. 2003. *Census 2000, Public Use Microdata Sample (PUMS), United States, Technical Documentation*. http://www.census.gov/prod/cen2000/doc/pums.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 Voshell. 1991. Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample. *Research Report Series* (CENSUS/SRD/RR-91/08). 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.

7. Data Sources

U. S. Census Bureau, Population Division. 2001. Census 2000 PHC-T-13. Population and Ranking Tables of the Older Population for the United States, State, Puerto Rico, Places of 100,000 or More Population, and Counties (Table 6. Counties Ranked by Percent 65 Years and Over: 2000). http://www.census.gov/population/www/cen2000/briefs/phc-t13/index.html (accessed December 27, 2008)

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, 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.

8. List of Exhibits in Paper

9. List of Tables and Figures in Paper

10. List of Tables and Figures in Appendix

[Exhibit 1]

Global Recoding of Geography Attributes					
Coarseness of Measures	Metric Spaces	Absolute Values	Coarseness of Measures	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

Table 1. Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	County Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	2,782	(2,782)	2,383	(2,231)	2,224	(2,058)	1,761	(1,597)	1,607	(1,418)
20%-Categories	2,485	(2,485)	1,867	(1,675)	1,551	(1,331)	1,143	(988)	933	(787)
15%-Categories	2,218	(2,218)	1,494	(1,279)	1,163	(933)	714	(590)	516	(413)
10%-Categories	1,921	(1,921)	1,177	(966)	810	(612)	435	(354)	296	(221)
5%-Categories	1,397	(1,397)	592	(448)	325	(206)	119	(86)	68	(45)
1%-Categories	408	(408)	54	(36)	15	(5)	2	(2)	1	(1)
Population Density										
Top, Bot-25%, Oth	1,406	(1,406)	1,226	(1,158)	1,135	(1,056)	923	(848)	858	(771)
20%-Categories	1,262	(1,262)	971	(875)	805	(690)	619	(533)	512	(430)
15%-Categories	1,133	(1,133)	784	(679)	613	(496)	396	(330)	289	(234)
10%-Categories	987	(987)	620	(517)	432	(331)	251	(205)	174	(133)
5%-Categories	731	(731)	323	(249)	183	(117)	74	(53)	43	(29)
1%-Categories	237	(237)	33	(22)	10	(3)	2	(1)	1	(1)
Division										
Top, Bot-25%, Oth	390	(390)	341	(323)	320	(300)	265	(244)	243	(218)
20%-Categories	353	(353)	278	(253)	235	(206)	183	(161)	156	(135)
15%-Categories	320	(320)	229	(201)	182	(151)	124	(105)	96	(80)
10%-Categories	282	(282)	185	(156)	132	(103)	82	(68)	60	(47)
5%-Categories	211	(211)	102	(79)	59	(40)	26	(19)	17	(12)
1%-Categories	67	(67)	12	(8)	4	(2)	1	(1)	1	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	County Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	217	(217)	194	(185)	181	(170)	155	(144)	145	(132)
20%-Categories	198	(198)	161	(148)	137	(120)	112	(99)	98	(85)
15%-Categories	181	(181)	135	(120)	108	(91)	78	(67)	62	(52)
10%-Categories	161	(161)	110	(94)	80	(64)	54	(45)	41	(32)
5%-Categories	122	(122)	62	(50)	38	(26)	18	(14)	12	(9)
1%-Categories	43	(43)	8	(6)	3	(2)	1	(1)	1	(1)
State										
Top, Bot-25%, Oth	78	(78)	69	(65)	65	(62)	55	(51)	51	(46)
20%-Categories	71	(71)	57	(52)	49	(43)	39	(35)	33	(29)
15%-Categories	65	(65)	47	(42)	38	(32)	27	(23)	22	(18)
10%-Categories	57	(57)	38	(33)	28	(23)	18	(15)	14	(11)
5%-Categories	43	(43)	22	(18)	13	(10)	7	(5)	5	(4)
1%-Categories	15	(15)	3	(3)	2	(1)	1	(1)	1	(1)
State & Pop.Dens.										
Top, Bot-25%, Oth	43	(43)	39	(37)	37	(35)	32	(30)	30	(28)
20%-Categories	40	(40)	33	(30)	28	(25)	24	(21)	21	(19)
15%-Categories	37	(37)	28	(25)	23	(20)	17	(15)	14	(12)
10%-Categories	33	(33)	23	(20)	17	(14)	12	(10)	10	(8)
5%-Categories	25	(25)	14	(11)	9	(7)	5	(4)	4	(3)
1%-Categories	10	(10)	3	(2)	2	(1)	1	(1)	1	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Tract Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	49,334	(49,334)	35,764	(32,211)	30,047	(26,346)	21,425	(18,806)	16,495	(13,988)
20%-Categories	42,686	(42,686)	26,741	(22,477)	20,277	(16,235)	12,435	(10,304)	8,716	(6,835)
15%-Categories	37,845	(37,845)	21,452	(17,298)	14,832	(11,178)	7,605	(6,049)	4,926	(3,640)
10%-Categories	31,128	(31,128)	15,142	(11,557)	9,117	(6,332)	3,795	(2,944)	2,242	(1,536)
5%-Categories	20,939	(20,939)	6,989	(4,800)	3,231	(1,883)	819	(557)	437	(217)
1%-Categories	5,080	(5,080)	454	(273)	77	(29)	9	(4)	4	(1)
Population Density										
Top, Bot-25%, Oth	18,994	(18,994)	13,703	(12,402)	11,546	(10,186)	8,278	(7,338)	6,377	(5,496)
20%-Categories	16,349	(16,349)	10,144	(8,558)	7,717	(6,220)	4,729	(3,940)	3,348	(2,652)
15%-Categories	14,472	(14,472)	8,104	(6,573)	5,622	(4,288)	2,923	(2,342)	1,903	(1,425)
10%-Categories	11,851	(11,851)	5,673	(4,361)	3,437	(2,437)	1,458	(1,141)	880	(610)
5%-Categories	7,921	(7,921)	2,622	(1,826)	1,226	(741)	324	(226)	178	(92)
1%-Categories	1,965	(1,965)	178	(112)	32	(13)	4	(2)	2	(1)
Division										
Top, Bot-25%, Oth	5,808	(5,808)	4,248	(3,837)	3,576	(3,149)	2,583	(2,279)	2,003	(1,709)
20%-Categories	5,029	(5,029)	3,187	(2,688)	2,420	(1,952)	1,514	(1,265)	1,074	(850)
15%-Categories	4,473	(4,473)	2,578	(2,092)	1,789	(1,364)	956	(772)	632	(477)
10%-Categories	3,700	(3,700)	1,839	(1,414)	1,115	(785)	490	(385)	299	(209)
5%-Categories	2,534	(2,534)	885	(612)	415	(245)	116	(80)	66	(34)
1%-Categories	653	(653)	66	(39)	13	(6)	2	(2)	2	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Tract Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	2,508	(2,508)	1,823	(1,656)	1,540	(1,365)	1,119	(995)	870	(754)
20%-Categories	2,158	(2,158)	1,352	(1,147)	1,030	(837)	647	(544)	466	(373)
15%-Categories	1,916	(1,916)	1,089	(891)	759	(588)	415	(338)	278	(213)
10%-Categories	1,576	(1,576)	768	(597)	469	(339)	213	(169)	133	(95)
5%-Categories	1,069	(1,069)	368	(260)	174	(108)	52	(37)	31	(17)
1%-Categories	282	(282)	29	(18)	6	(3)	2	(1)	1	(1)
State										
Top, Bot-25%, Oth	1,229	(1,229)	897	(816)	760	(676)	557	(496)	431	(373)
20%-Categories	1,057	(1,057)	662	(563)	506	(412)	321	(271)	227	(182)
15%-Categories	938	(938)	533	(436)	373	(288)	204	(167)	135	(104)
10%-Categories	774	(774)	378	(294)	232	(167)	106	(84)	66	(47)
5%-Categories	530	(530)	183	(129)	87	(53)	26	(19)	16	(9)
1%-Categories	140	(140)	15	(9)	4	(2)	1	(1)	1	(1)
State & Pop.Dens.										
Top, Bot-25%, Oth	657	(657)	475	(434)	404	(361)	296	(266)	229	(201)
20%-Categories	561	(561)	345	(294)	264	(217)	167	(141)	119	(97)
15%-Categories	497	(497)	276	(228)	193	(152)	107	(88)	72	(56)
10%-Categories	406	(406)	193	(152)	118	(87)	55	(45)	35	(25)
5%-Categories	275	(275)	93	(67)	44	(29)	14	(10)	9	(5)
1%-Categories	73	(73)	8	(6)	2	(2)	1	(1)	1	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Blockgroup Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	154,293	(154,293)	110,025	(98,322)	88,607	(76,730)	61,535	(53,964)	48,423	(40,399)
20%-Categories	135,594	(135,594)	85,528	(72,070)	62,437	(49,998)	38,267	(31,961)	27,979	(21,707)
15%-Categories	119,547	(119,547)	67,310	(54,208)	44,637	(33,254)	22,512	(17,968)	14,943	(10,789)
10%-Categories	99,080	(99,080)	47,868	(36,850)	27,831	(19,201)	11,859	(9,214)	7,302	(4,856)
5%-Categories	66,936	(66,936)	22,128	(15,654)	9,980	(5,735)	2,853	(1,964)	1,468	(706)
1%-Categories	20,252	(20,252)	1,637	(1,044)	351	(89)	60	(15)	13	()
Population Density										
Top, Bot-25%, Oth	57,682	(57,682)	40,867	(36,669)	33,051	(28,822)	23,019	(20,379)	18,012	(15,241)
20%-Categories	50,570	(50,570)	31,679	(26,836)	23,199	(18,752)	14,241	(11,983)	10,502	(8,262)
15%-Categories	44,483	(44,483)	24,791	(20,078)	16,466	(12,424)	8,395	(6,755)	5,605	(4,095)
10%-Categories	36,749	(36,749)	17,538	(13,596)	10,219	(7,186)	4,412	(3,462)	2,751	(1,858)
5%-Categories	24,663	(24,663)	8,087	(5,794)	3,657	(2,172)	1,070	(752)	560	(280)
1%-Categories	7,440	(7,440)	606	(399)	132	(38)	23	(7)	6	(1)
Division										
Top, Bot-25%, Oth	17,630	(17,630)	12,657	(11,342)	10,223	(8,895)	7,182	(6,328)	5,662	(4,751)
20%-Categories	15,502	(15,502)	9,854	(8,330)	7,217	(5,818)	4,493	(3,776)	3,298	(2,578)
15%-Categories	13,710	(13,710)	7,806	(6,317)	5,201	(3,914)	2,708	(2,186)	1,813	(1,326)
10%-Categories	11,419	(11,419)	5,609	(4,340)	3,279	(2,288)	1,455	(1,140)	908	(611)
5%-Categories	7,810	(7,810)	2,669	(1,895)	1,220	(709)	377	(260)	198	(97)
1%-Categories	2,421	(2,421)	212	(134)	49	(12)	10	(3)	3	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Blockgroup Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	7,396	(7,396)	5,269	(4,744)	4,273	(3,747)	3,013	(2,677)	2,367	(2,014)
20%-Categories	6,478	(6,478)	4,078	(3,469)	2,990	(2,436)	1,872	(1,585)	1,388	(1,102)
15%-Categories	5,714	(5,714)	3,211	(2,617)	2,139	(1,634)	1,133	(922)	765	(568)
10%-Categories	4,735	(4,735)	2,289	(1,788)	1,337	(955)	607	(482)	384	(264)
5%-Categories	3,207	(3,207)	1,080	(780)	492	(298)	158	(111)	85	(44)
1%-Categories	994	(994)	87	(58)	21	(6)	5	(2)	2	(1)
State										
Top, Bot-25%, Oth	3,595	(3,595)	2,572	(2,320)	2,093	(1,839)	1,488	(1,324)	1,170	(994)
20%-Categories	3,140	(3,140)	1,973	(1,679)	1,449	(1,180)	910	(772)	668	(528)
15%-Categories	2,772	(2,772)	1,555	(1,269)	1,038	(793)	550	(449)	369	(274)
10%-Categories	2,303	(2,303)	1,112	(868)	651	(462)	296	(234)	186	(127)
5%-Categories	1,572	(1,572)	528	(380)	242	(145)	79	(55)	42	(21)
1%-Categories	492	(492)	44	(28)	11	(4)	3	(1)	1	(1)
State & Pop.Dens.										
Top, Bot-25%, Oth	1,889	(1,889)	1,335	(1,209)	1,089	(963)	772	(692)	600	(517)
20%-Categories	1,640	(1,640)	1,013	(866)	742	(610)	464	(396)	342	(274)
15%-Categories	1,442	(1,442)	792	(651)	526	(407)	280	(230)	189	(142)
10%-Categories	1,189	(1,189)	559	(440)	324	(236)	149	(119)	95	(66)
5%-Categories	799	(799)	261	(191)	118	(74)	39	(28)	22	(12)
1%-Categories	247	(247)	22	(15)	6	(2)	2	(1)	1	(1)

Table 1 (cont.). Average Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

SOURCE. -- 2000 U.S. Census of Population and Housing

NOTE. -- The upper-bound of the 95% confidence interval for the average number of look-alike geographic units is presented in parentheses. This estimate is adjusted to account for the complex survey design of sampled variables sets.

Table 2. Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	County Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	2,782	(2,782)	2,383	(2,231)	2,224	(2,058)	1,761	(1,597)	1,607	(1,418)
20%-Categories	2,485	(2,485)	1,867	(1,675)	1,551	(1,331)	1,143	(988)	933	(787)
15%-Categories	2,218	(2,218)	1,494	(1,279)	1,163	(933)	714	(590)	516	(413)
10%-Categories	1,921	(1,921)	1,177	(966)	810	(612)	435	(354)	296	(221)
5%-Categories	1,397	(1,397)	592	(448)	325	(206)	119	(86)	68	(45)
1%-Categories	408	(408)	54	(36)	15	(5)	2	(2)	1	(1)
Population Density										
Top, Bot-25%, Oth	1,406	(1,406)	1,226	(1,158)	1,135	(1,056)	923	(848)	858	(771)
20%-Categories	1,262	(1,262)	971	(875)	805	(690)	619	(533)	512	(430)
15%-Categories	1,133	(1,133)	784	(679)	613	(496)	396	(330)	289	(234)
10%-Categories	987	(987)	620	(517)	432	(331)	251	(205)	174	(133)
5%-Categories	731	(731)	323	(249)	183	(117)	74	(53)	43	(29)
1%-Categories	237	(237)	33	(22)	10	(3)	2	(1)	1	(1)
Division										
Top, Bot-25%, Oth	390	(390)	341	(323)	320	(300)	265	(244)	243	(218)
20%-Categories	353	(353)	278	(253)	235	(206)	183	(161)	156	(135)
15%-Categories	320	(320)	229	(201)	182	(151)	124	(105)	96	(80)
10%-Categories	282	(282)	185	(156)	132	(103)	82	(68)	60	(47)
5%-Categories	211	(211)	102	(79)	59	(40)	26	(19)	17	(12)
1%-Categories	67	(67)	12	(8)	4	(2)	1	(1)	1	(1)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	County Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	217	(217)	194	(185)	181	(170)	155	(144)	145	(132)
20%-Categories	198	(198)	161	(148)	137	(120)	112	(99)	98	(85)
15%-Categories	181	(181)	135	(120)	108	(91)	78	(67)	62	(52)
10%-Categories	161	(161)	110	(94)	80	(64)	54	(45)	41	(32)
5%-Categories	122	(122)	62	(50)	38	(26)	18	(14)	12	(9)
1%-Categories	43	(43)	8	(6)	3	(2)	1	(1)	1	(1)
State										
Top, Bot-25%, Oth	78	(78)	69	(65)	65	(62)	55	(51)	51	(46)
20%-Categories	71	(71)	57	(52)	49	(43)	39	(35)	33	(29)
15%-Categories	65	(65)	47	(42)	38	(32)	27	(23)	22	(18)
10%-Categories	57	(57)	38	(33)	28	(23)	18	(15)	14	(11)
5%-Categories	43	(43)	22	(18)	13	(10)	7	(5)	5	(4)
1%-Categories	15	(15)	3	(3)	2	(1)	1	(1)	1	(1)
State & Pop.Dens.										
Top, Bot-25%, Oth	43	(43)	39	(37)	37	(35)	32	(30)	30	(28)
20%-Categories	40	(40)	33	(30)	28	(25)	24	(21)	21	(19)
15%-Categories	37	(37)	28	(25)	23	(20)	17	(15)	14	(12)
10%-Categories	33	(33)	23	(20)	17	(14)	12	(10)	10	(8)
5%-Categories	25	(25)	14	(11)	9	(7)	5	(4)	4	(3)
1%-Categories	10	(10)	3	(2)	2	(1)	1	(1)	1	(1)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Tract Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	2,378	(2,378)	1,724	(1,553)	1,448	(1,270)	1,033	(906)	795	(674)
20%-Categories	2,057	(2,057)	1,289	(1,083)	977	(783)	599	(497)	420	(329)
15%-Categories	1,824	(1,824)	1,034	(834)	715	(539)	367	(292)	237	(175)
10%-Categories	1,500	(1,500)	730	(557)	439	(305)	183	(142)	108	(74)
5%-Categories	1,009	(1,009)	337	(231)	156	(91)	39	(27)	21	(10)
1%-Categories	245	(245)	22	(13)	4	(1)	0	(0)	0	(0)
Population Density										
Top, Bot-25%, Oth	916	(916)	660	(598)	557	(491)	399	(354)	307	(265)
20%-Categories	788	(788)	489	(412)	372	(300)	228	(190)	161	(128)
15%-Categories	698	(698)	391	(317)	271	(207)	141	(113)	92	(69)
10%-Categories	571	(571)	273	(210)	166	(117)	70	(55)	42	(29)
5%-Categories	382	(382)	126	(88)	59	(36)	16	(11)	9	(4)
1%-Categories	95	(95)	9	(5)	2	(1)	0	(0)	0	(0)
Division										
Top, Bot-25%, Oth	280	(280)	205	(185)	172	(152)	124	(110)	97	(82)
20%-Categories	242	(242)	154	(130)	117	(94)	73	(61)	52	(41)
15%-Categories	216	(216)	124	(101)	86	(66)	46	(37)	30	(23)
10%-Categories	178	(178)	89	(68)	54	(38)	24	(19)	14	(10)
5%-Categories	122	(122)	43	(30)	20	(12)	6	(4)	3	(2)
1%-Categories	31	(31)	3	(2)	1	(0)	0	(0)	0	(0)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Tract Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	121	(121)	88	(80)	74	(66)	54	(48)	42	(36)
20%-Categories	104	(104)	65	(55)	50	(40)	31	(26)	22	(18)
15%-Categories	92	(92)	52	(43)	37	(28)	20	(16)	13	(10)
10%-Categories	76	(76)	37	(29)	23	(16)	10	(8)	6	(5)
5%-Categories	52	(52)	18	(13)	8	(5)	3	(2)	1	(1)
1%-Categories	14	(14)	1	(1)	0	(0)	0	(0)	0	(0)
State										
Top, Bot-25%, Oth	59	(59)	43	(39)	37	(33)	27	(24)	21	(18)
20%-Categories	51	(51)	32	(27)	24	(20)	15	(13)	11	(9)
15%-Categories	45	(45)	26	(21)	18	(14)	10	(8)	7	(5)
10%-Categories	37	(37)	18	(14)	11	(8)	5	(4)	3	(2)
5%-Categories	26	(26)	9	(6)	4	(3)	1	(1)	1	(0)
1%-Categories	7	(7)	1	(0)	0	(0)	0	(0)	0	(0)
State & Pop.Dens.										
Top, Bot-25%, Oth	32	(32)	23	(21)	19	(17)	14	(13)	11	(10)
20%-Categories	27	(27)	17	(14)	13	(10)	8	(7)	6	(5)
15%-Categories	24	(24)	13	(11)	9	(7)	5	(4)	3	(3)
10%-Categories	20	(20)	9	(7)	6	(4)	3	(2)	2	(1)
5%-Categories	13	(13)	4	(3)	2	(1)	1	(1)	0	(0)
1%-Categories	4	(4)	0	(0)	0	(0)	0	(0)	0	(0)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Blockgroup Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
National (None)										
Top, Bot-25%, Oth	2,330	(2,330)	1,661	(1,485)	1,338	(1,159)	929	(815)	731	(610)
20%-Categories	2,047	(2,047)	1,291	(1,088)	943	(755)	578	(483)	422	(328)
15%-Categories	1,805	(1,805)	1,016	(819)	674	(502)	340	(271)	226	(163)
10%-Categories	1,496	(1,496)	723	(556)	420	(290)	179	(139)	110	(73)
5%-Categories	1,011	(1,011)	334	(236)	151	(87)	43	(30)	22	(11)
1%-Categories	306	(306)	25	(16)	5	(1)	1	(0)	0	(0)
Population Density										
Top, Bot-25%, Oth	871	(871)	617	(554)	499	(435)	348	(308)	272	(230)
20%-Categories	764	(764)	478	(405)	350	(283)	215	(181)	159	(125)
15%-Categories	672	(672)	374	(303)	249	(188)	127	(102)	85	(62)
10%-Categories	555	(555)	265	(205)	154	(109)	67	(52)	42	(28)
5%-Categories	372	(372)	122	(87)	55	(33)	16	(11)	8	(4)
1%-Categories	112	(112)	9	(6)	2	(1)	0	(0)	0	(0)
Division										
Top, Bot-25%, Oth	266	(266)	191	(171)	154	(134)	108	(96)	85	(72)
20%-Categories	234	(234)	149	(126)	109	(88)	68	(57)	50	(39)
15%-Categories	207	(207)	118	(95)	79	(59)	41	(33)	27	(20)
10%-Categories	172	(172)	85	(66)	50	(35)	22	(17)	14	(9)
5%-Categories	118	(118)	40	(29)	18	(11)	6	(4)	3	(1)
1%-Categories	37	(37)	3	(2)	1	(0)	0	(0)	0	(0)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

	Blockgroup Attributes									
	1-Key		2-Keys		3-Keys		4-Keys		5-Keys	
	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI	Number	95% CI
Division & Pop.Dens.										
Top, Bot-25%, Oth	112	(112)	80	(72)	65	(57)	45	(40)	36	(30)
20%-Categories	98	(98)	62	(52)	45	(37)	28	(24)	21	(17)
15%-Categories	86	(86)	48	(40)	32	(25)	17	(14)	12	(9)
10%-Categories	72	(72)	35	(27)	20	(14)	9	(7)	6	(4)
5%-Categories	48	(48)	16	(12)	7	(5)	2	(2)	1	(1)
1%-Categories	15	(15)	1	(1)	0	(0)	0	(0)	0	(0)
State										
Top, Bot-25%, Oth	54	(54)	39	(35)	32	(28)	22	(20)	18	(15)
20%-Categories	47	(47)	30	(25)	22	(18)	14	(12)	10	(8)
15%-Categories	42	(42)	23	(19)	16	(12)	8	(7)	6	(4)
10%-Categories	35	(35)	17	(13)	10	(7)	4	(4)	3	(2)
5%-Categories	24	(24)	8	(6)	4	(2)	1	(1)	1	(0)
1%-Categories	7	(7)	1	(0)	0	(0)	0	(0)	0	(0)
State & Pop.Dens.										
Top, Bot-25%, Oth	29	(29)	20	(18)	16	(15)	12	(10)	9	(8)
20%-Categories	25	(25)	15	(13)	11	(9)	7	(6)	5	(4)
15%-Categories	22	(22)	12	(10)	8	(6)	4	(3)	3	(2)
10%-Categories	18	(18)	8	(7)	5	(4)	2	(2)	1	(1)
5%-Categories	12	(12)	4	(3)	2	(1)	1	(0)	0	(0)
1%-Categories	4	(4)	0	(0)	0	(0)	0	(0)	0	(0)

Table 2 (cont.). Average Standardized Number of Look-Alike Geographic Units for Respondent Locations, By Spatial Scale and Experimental Traits of Geography Attributes

SOURCE. -- 2000 U.S. Census of Population and Housing

NOTE. -- The upper-bound of the 95% confidence interval for the Average Standardized Number of look-alike geographic units is presented in parentheses. This estimate is adjusted to account for the complex survey design of sampled variables sets.

NOTE. -- Geography population counts for tracts and blockgroups (presented in Table 1) are multiplied by 0.0482 and 0.0151 respectively, such that these estimates are benchmarked to county-level counts.

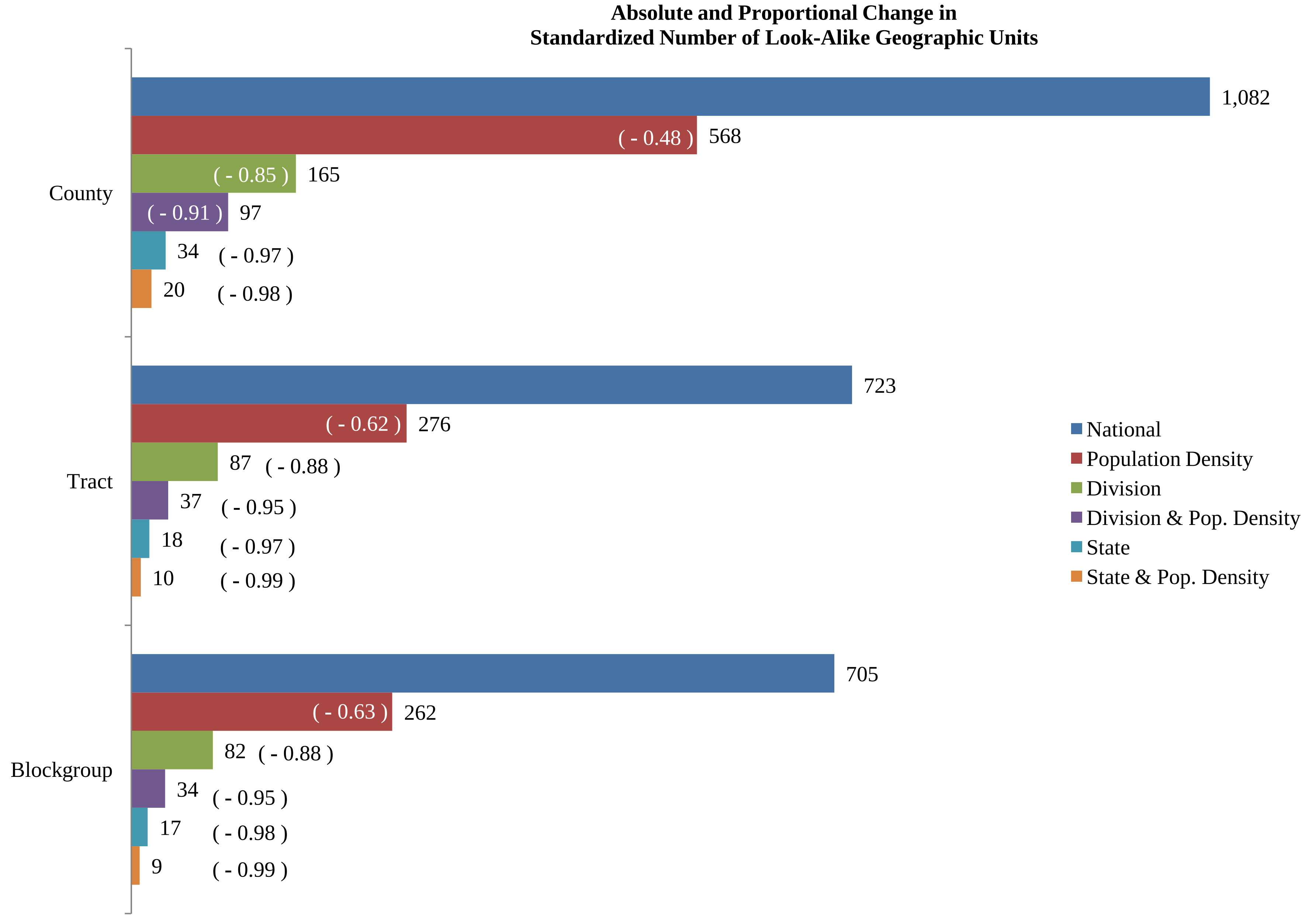


Figure 1. Scope of Study and Reductions in Full Intruder Search Effort

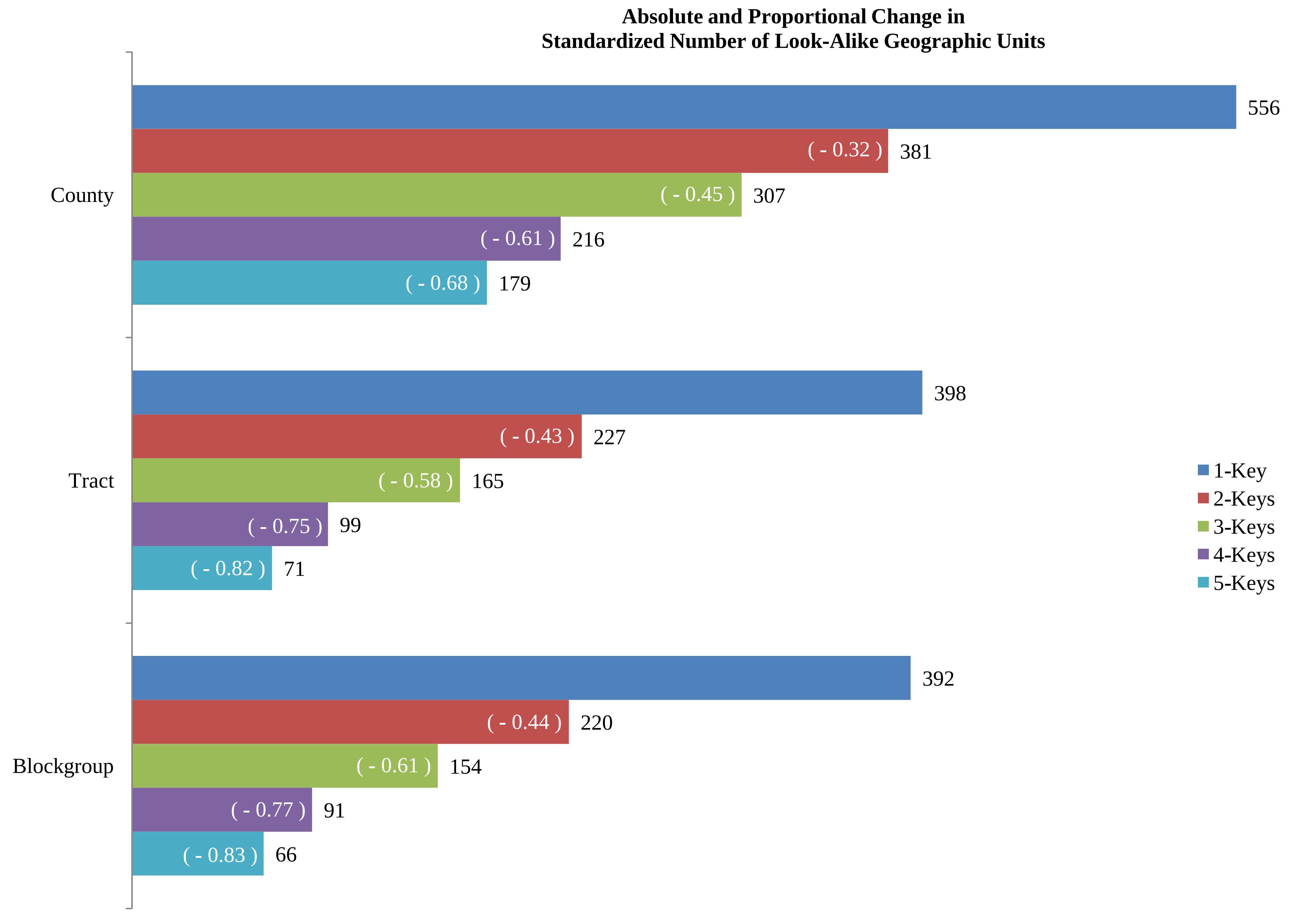


Figure 2. Number of Geographic Attributes and Reductions in Full Intruder Search Effort

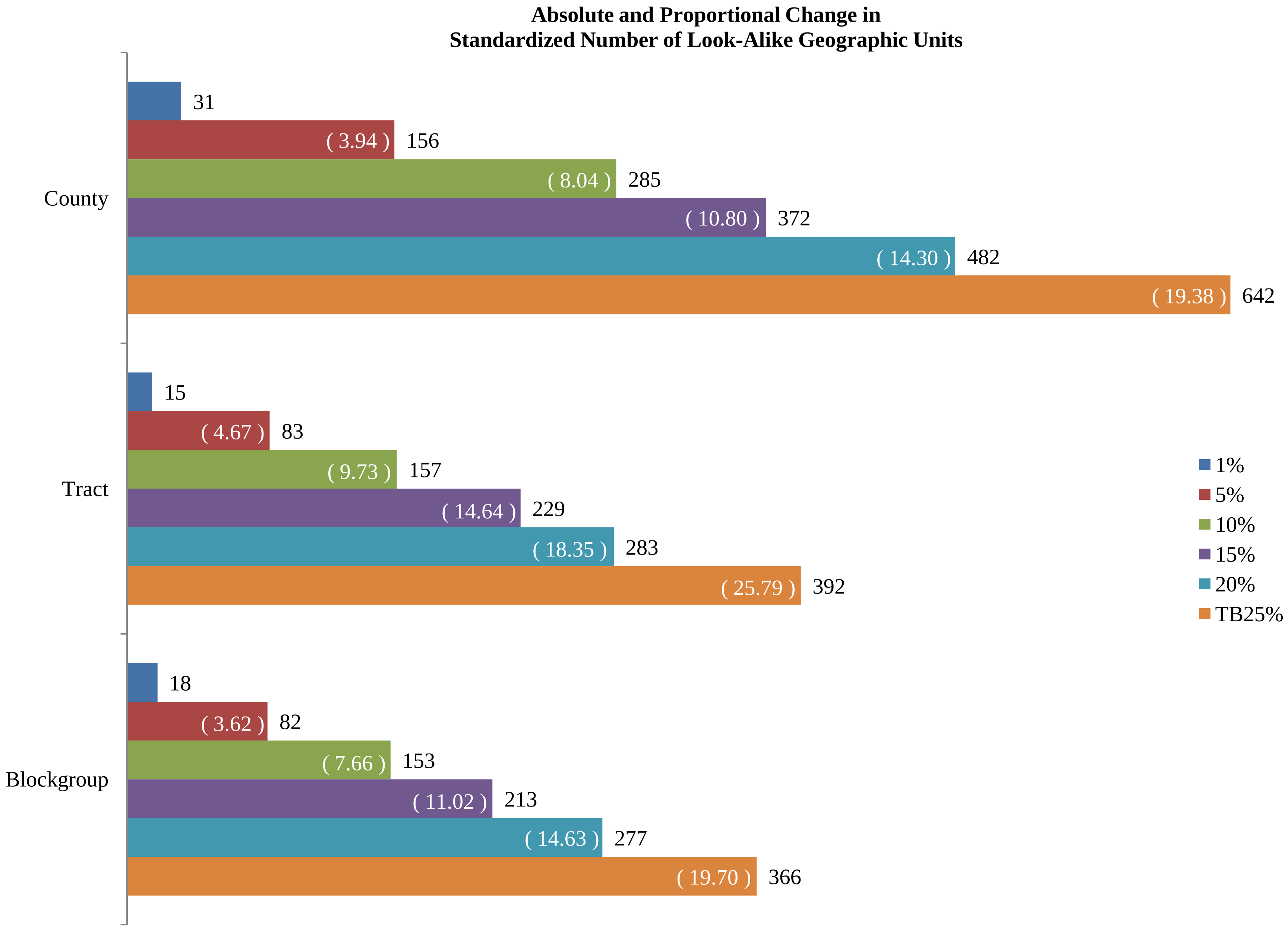
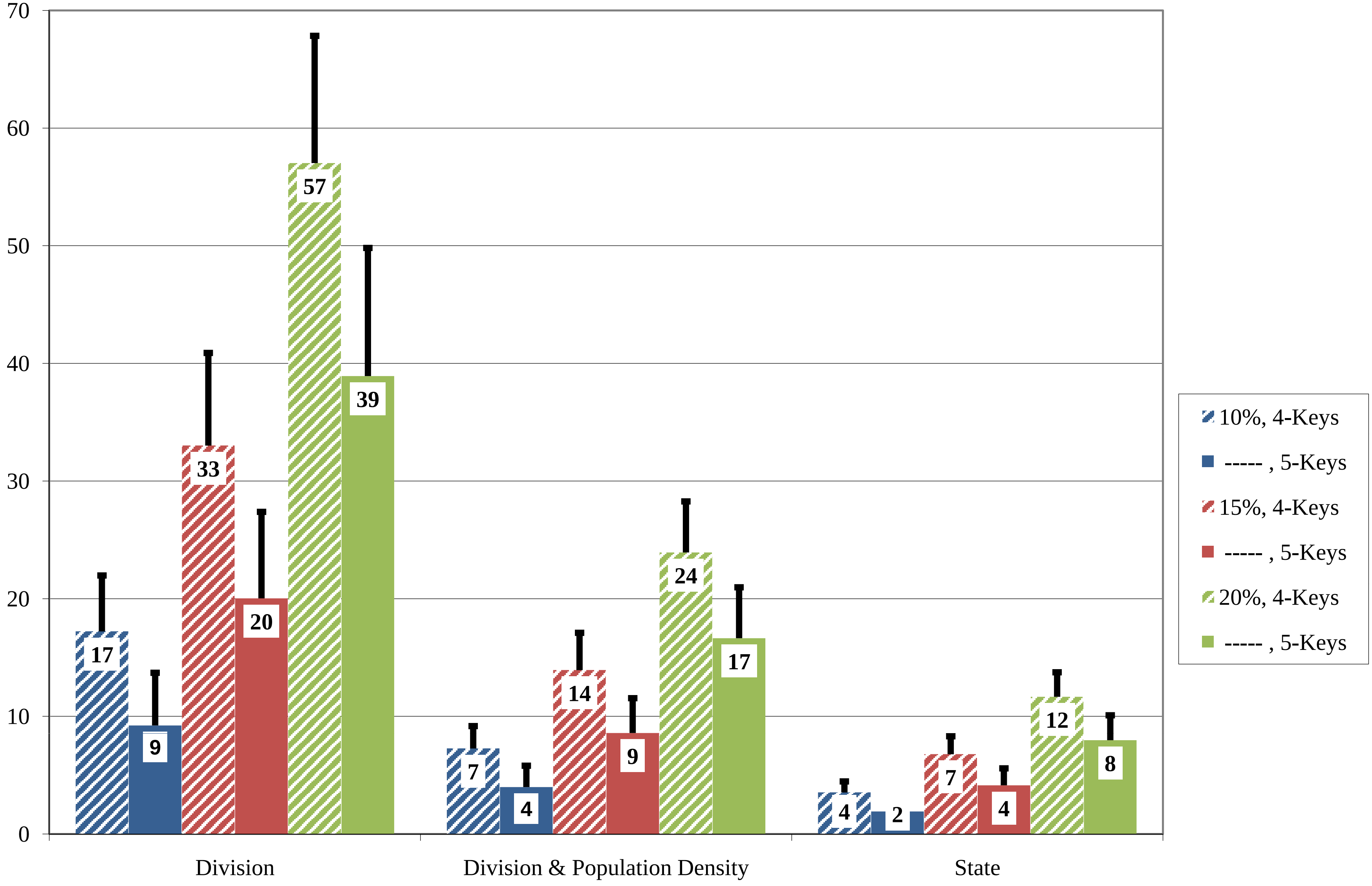


Figure 3. Measurement Detail of Geographic Attributes and Increases in Full Intruder Search Effort


70
60
50
40
30
20
10
0
17
9
33
20
57
39
7
4
14
9
24
17
4
2
7
4
12
8
Division
Division & Population Density
State
10%, 4-Keys
----- , 5-Keys
15%, 4-Keys
----- , 5-Keys
20%, 4-Keys
----- , 5-Keys