Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Finding a Needle in a Haystack: The Theoretical and Empirical Foundations of Assessing Disclosure Risk for Contextualized Microdata

Kevin T. Leicht and Kristine Witkowski[1]

Abstract

Our study describes various factors that are of concern when evaluating disclosure risk of contextualized microdata and some of the empirical steps that are involved in their assessment. Utilizing synthetic sets of survey respondents, we illustrate how different postulates shape the assessment of risk when considering: (1) estimated probabilities that unidentified geographic areas are represented within a survey; (2) the number of people in the population who share the same personal and contextual identifiers as a respondent; and (3) the anticipated amount of coverage error in census population counts and extant files that provide identifying information (like names and addresses).

Key Words: Confidentiality; Dissemination; Disclosure Risk

1. Introduction

Many problems in contemporary social science lend themselves to an analysis where the individuals under study are placed in context, defined spatially as a 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 data users can link the appropriate contextual data themselves), or merging the contextual data, effectively adding the characteristics of the spatial unit where the subject lives. In the second case, the record for each individual includes that person's characteristics (e.g., age of respondent) and spatial characteristics (e.g., proportion of population in respondent's neighborhood that is poor).

One reason for providing the contextual data outright rather than identifying the spatial unit is that doing so makes it more difficult to identify the spatial unit where the survey respondent lives (Armstrong, Rushton, and Zimmerman 1999). But the contextual data themselves may constitute enough information to be a geographical unique.
Knowing the county (or census tract, or block group) does not mean that we can identify individuals. That is just a starting point. Because microdata files typically consist of both individual and contextual measures, a full assessment of risk requires a nested approach that incorporates identifying characteristics of survey respondents and their locations. This study helps lay the groundwork for such an evaluation with its needle-in-haystack approach to disclosure and discussing associated methodological concerns.

With an analytical approach bridging two levels, our current study informs the design of public-use data files composed of person-records containing contextual measures from counties, tracts, and block groups. Utilizing a synthetic test data file, we illustrate how the re-identification of individuals is affected by aggregating geographies into look-alike sets. We further assess one dimension of risk associated with contextualized microdata - identifying survey respondents whose personal characteristics (i.e., sex, age, and race) are rarely found among populations sharing the same contexts.

[1] Kevin T. Leicht, Department of Sociology, The University of Illinois Urbana-Champaign, 3120 Lincoln Hall MC-454, 702 South Wright Street, Urbana, IL 61801. Email: kleicht@illinois.edu; Kristine Witkowski, Inter-University Consortium for Political and Social Research (ICPSR), Institute for Social Research, The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106-1248. E-mail: kwitkow@umich.edu

Using geographic-units and their "aggregated look-alike contexts" as our units of analysis, the number of persons in the population with a distinct set of personal characteristics as the outcome of interest, and indicators of underlying risk, we detail the complexity of re-identification patterns by assessing the likelihood that young adult white and black males would be pinpointed within reconstituted geographic haystacks given: (1) the size of the total population of aggregated contexts; (2) the amount of error in population counts; and (3) differential search costs stemming from spatially-dispersed contexts.

2. The Definition of Risk

To help explain the conceptual underpinnings of anonymized data, we rely on the English idiom of the "needle in a haystack", which refers to an item that is difficult to find because it is hidden in a larger set of objects (Cambridge University Press 2003). A survey respondent (or "needle") is a specific person in a public-use data file who has a particular set of individual traits that are easily ascertained by an intruder (sex, age, race, ethnicity, marital status, or education-level). This survey respondent also is a member of a group (or "haystack") of people in the population sharing the same identifying individual and contextual characteristics. As with finding a needle in a haystack, the chance that a respondent is correctly re-identified is considered low when an intruder must search among a sufficiently large number of people sharing the same characteristics in an identifying extant file.

In the initial stage of our re-identification experiment, we assume that an intruder will rely on three sets of compositional variables to assess the difficulty of pinpointing survey respondents. The gender, age, and race/ethnicity of survey respondents define the composition of haystacks where these "needles" are hidden. Enriching microdata files with geographic information, variables describing the contexts of unknown surveyed locations (i.e., % Persons, Foreign-Born; metropolitan status) further constrain where respondents may be found. A third set of composition variables, in the form of hard-to-count scores and racially-specific rentership rates, are used to assess the accuracy of estimated haystack sizes and the likely amounts of coverage error in extant files. The correlation between these three sets of contextual measures is considerable. It is this measurement overlap that determines the shift in chances affecting the correct re-identification of respondents. This overlap reflects the degree to which an aggregated context is residentially segregated, characteristically unique, and difficult to enumerate

Covering a broad spectrum of social scientific inquiry, we have selected five contextual variables to be represented in our test dataset: (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) % 14 Persons, In-Poverty; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed. Following our earlier work (Witkowski 2016), we have applied a non-perturbative masking technique in hopes of broadly informing the design of datasets containing contextual information at various spatial scales. After top-coding and bottom-coding these continuous variables to conceal outliers, Iwerecode contextual measures into ten metric spaces of 10% categories (i.e., 0 - 9%, 10 - 19%, 20 - 29%, 30 - 39%, 40 - 49%, 50 - 59%, 60 - 69%, 70 - 79%, 80 - 89%, 90 -100%). Outliers were identified as those within the top and bottom 0.5% of each variables distribution (Zayatz 2005) given geographically-specific distributions defined by the metropolitan status of the geographies. Contextual variables are recoded into aggregated categories based on their absolute values (i.e., absolute recoding).

For each sampled geographic unit (1,785 counties; 8,947 tracts; 10,478 block groups) and look-alike units in an aggregated context (315 counties; 2,280 tracts; and 3,090 block groups) we compile three sets of information regarding the surveyed population as well as the size and composition of the total population. First we count the total number of respondents in each location and context. We then tally the total population size of individual and aggregated areas. Finally, we tally the total number of persons who have a selected set of personal characteristics, providing estimates of haystack size for each location and context. Since analyses for a broad array of haystacks are beyond the scope of this paper, we simplify my study by assessing only one majority and one minority haystack consisting of 20-year-old males who are either (1) non-Hispanic white alone or (2) African-American or black alone. With this approach, we are able to investigate how minority-status influences the re-identification of respondents holding constant their gender and age.

3. What we do here?

In our analytical exercise, we (1) construct a microdata file composed of a single synthetic sample of survey respondents, attaching contextual information to person-level records; (2) identify geographic units that resemble a respondent's location, using available contextual data for counties, tracts, and blockgroups; (3) assemble base population, haystack, coverage error, and spatial dispersion information for geographic units and aggregated contexts, attaching these search estimates to person-level records; and (4) calculate summary statistics for test microdata file, reflecting the distribution of survey respondents.

The analytical unit of ultimate concern is the sampled person or respondent. While synthetic survey respondents are provided a set of contextual characteristics describing their locations, they are not assigned any personal characteristics. Hence we do not know the sex, age, and race of any particular respondent. But we do know the size of haystacks and the total population within each geographic unit and aggregated context as well as the proportion of respondents who have been drawn from these areas. Bringing this information together, we can ascertain the likelihood that a given respondent will have a set of personal characteristics as defined by the haystack of interest (in this case, 20-year old non-Hispanic white or African-American males).

4. Results

We conduct two sets of analyses that evaluate the potential role of haystack size, coverage error, and dispersion in determining disclosure risk of contextualized microdata. For the first set of analyses presented in Table 1, we assess how the re-identification process is changed when we attach contextual information instead of directly identifying geographic units. We illustrate how the above haystack characteristics are functions of sparsely and densely populated geographic units; and how the aggregation of look-alike geographies reduces disclosure risk by increasing uncertainty and search costs. To illustrate we produce separate analyses for survey respondents nested within individual geographic units and aggregated contexts that are highly populated ("dense") or less populated ("sparse"), defined as those whose populations are above and below 100,000 persons.

Looking at Table 1, analyses reveal that the release of contextual information (instead of the direct identification of places) has important ramifications for disclosure risk. The chances of a survey respondent being located in a densely populated area dramatically increases when contextual data are attached to individual records. Most survey respondents (96%) are found in densely populated contexts derived from counties; while a majority (59 to 67%) live in highly-populated contexts aggregated from small-scale geographic units. For all scales of geography, the total population size (on average) rises to over 2.4 million persons for high-density aggregated contexts; and for low-density contexts the population is at least 29,000 persons.

For all geographic units, the accumulation of base populations in aggregated contexts gives rise to significantly larger haystacks. High-density contexts have at least 11,000 non-Hispanic white males age 20, while sparse contexts have at least 125 members of this majority subpopulation (on average). As expected the minority subpopulation of young black males typically has much smaller haystacks than their majority counterpart. But aggregating look-alike geographies increases the size of haystacks emanating from small-scale units to the point that at least 59 twins are found in sparse contexts and as many as 467 are found in dense contexts. Minority haystacks are even larger for dense contexts derived from counties, with an average size of 2,641 target members.

Looking at the shifting patterns of coverage error resulting from aggregation, we find evidence for the selection of unique geographies in contexts with less than 100,000 persons. With twice the difficulty of being enumerated, 47 to 71% of look-alike geographies in sparse contexts are likely undercounted. The concentration of hard-to-count populations within relatively unpopulated and unique contexts is further indicated by the exceptionally low likelihood of double-counting (i.e., 1 to 5%). With national rates of undercounting ranging between 1 to 2%, the disproportionate number of difficult-to-enumerate tracts may prove to be a significant barrier to re-identification by intruders.

However the protection offered by coverage error, and the translation of hard-to-count scores to actual rates of undercounting, is intricate. First, there is evidence that the level of enumeration difficulty captured by hard-to-count scores does not sufficiently reflect the complexity of residential segregation patterns surrounding underrepresented race and ethnic groups. Regardless of the spatial scale, population density, and geographic detail of contexts, non-Hispanic white households are actually easier to enumerate than other populations within their tract, while African-American or black households are actually more difficult to count. Compared to others in their tract, approximately 2 to 11% more black households are renters while 13 to 17% more non-Hispanic white households are homeowners. Given this offsetting pattern of homeownership, coverage error tends to close the minority-majority gap in disclosure risk. However the chance that a small minority haystack has been largely hidden may be reduced in contexts with small base populations. For instance, a sparse context (derived from tracts) typically has 34,668 persons in its total population. Even if 71% of these contexts' tracts are probably undercounted, it is unlikely that this small population would contain an exorbitant number of hidden members.

While the potential size of hidden haystacks in sparse contexts may not be sufficiently large, the same high error rates could indicate the lack of coverage in identifying extant files. The intruder will likely have to perform supplemental search activities for respondents from these areas. While cost estimates for these activities are not available, we do know that the search – for respondents in aggregated contexts with less than 100,000 persons – would cover an average of: 2 counties, 7,503 square miles, and 2 states; 9 tracts, 581 square miles, and 5 states; or 23 blockgroups, under 1 square mile, and 10 states. If an intruder is willing to take a calculated risk, they could lower these costs by ignoring geographic units that are unlikely to have a survey respondent drawn from the area. The savings could be pronounced, especially for minority populations residing in sparse aggregated contexts, where 58 to 78% of look-alike geographies do not have a single black male aged 20.

Table 1. Aggregation of Sampled Geographic Units into Look-Alike Contexts, Weighted to Reflect Spatial Distribution of Survey Respondents (N=1,785; 8,947; and 10,478 of Sampled Counties, Tracts, and Blockgroups; N=315; 2,280; and 3,090 of Sampled Aggregated Contexts Based on Counties, Tracts, and Blockgroups; N=11,562 of Synthetic Survey Respondents)

	Survey Respondents Residing in											
	County as Contextual-Base				Tract as Contextual-Base				Blockgroup as Contextual-Base			
	Geographic Unit		Aggregated Context		Geographic Unit		Aggregated Context		Geographic Unit		Aggregated Context	
	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+
Proportion of Respondents in Context	0.34	0.66	0.04	0.96	1.00	0.00	0.33	0.67	1.00	0.00	0.41	0.59
Size of Total Population within Sampled Context	41,756	955,163	51,233	3,469,881	4,898	---	34,668	2,484,499	1,613	---	29,243	2,489,206
Average Size of Haystack Subpopulation of												
White alone, non-Hispanic, Males Age 20	275	3,128	264	17,715	25	---	155	11,808	8	---	126	11,155
African-American or Black alone, Males Age 20	33	1,222	130	2,641	5	---	76	467	2	---	59	318
Proportion of Tracts in Sampled Geographic Units and Aggregated Contexts that are Likely Under-Counted	0.17	0.28	0.47	0.23	0.29	---	0.71	0.08	0.29	---	0.59	0.07
Proportion of Tracts in Sampled Geographic Units and Aggregated Contexts that are Likely Over-Counted	0.18	0.28	0.05	0.25	0.19	---	0.01	0.28	0.19	---	0.03	0.32
Difference in Proportion Renters												
White alone, non-Hispanic (minus Others)	-0.16	-0.15	-0.13	-0.16	-0.16	---	-0.15	-0.17	-0.16	---	-0.16	-0.16
African-American or Black alone (minus Others)	0.02	0.11	0.04	0.09	0.10	---	0.09	0.10	0.10	---	0.11	0.11
Average Number of Geographic Units Resembling Sampled County	117	14	---	---	381	---	---	---	1,098	---	---	---
Average Number of Geographic Units in Aggregated Context [1]	---	---	2	51	---	---	9	566	---	---	23	1,838
Average Sq.Miles of Land Area of Sampled Geographic Units and Aggregated Contexts	1,848	1,272	7,503	44,691	113	---	581	39,759	0.03	---	0.44	32.52
Average Number of States in Aggregated Context	---	---	2	12	---	---	5	31	---	---	10	37
Proportion of Geographic Units with Subpopulation (in Context)												
White alone, non-Hispanic, Males Age 20	1.00	1.00	0.99	1.00	0.95	---	0.98	1.00	0.88	---	0.96	1.00
African-American or Black alone, Males Age 20	0.79	1.00	0.79	1.00	0.59	---	0.94	1.00	0.38	---	0.92	1.00

Note: Excluded from analyses are those geographic units with no population, resulting in 3,141 counties; 65,174 tracts; and 208,125 blockgroups considered from the geographic-unit population.

Note: Dataset contains five county, tract, and blockgroup-level contextual measures of (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) % Persons, In-Poverty; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed, recoded into 10% categories (i.e., 0 - 9%, 10 - 19%, 20 - 29%, 30 - 39%, 40 - 49%, 50 - 59%, 60 - 69%, 70 - 79%, 80 - 89%, 90 -100%).This dataset also directly identifies MSA-status of geographic units: (1) MSA 1-million or more, (2) MSA less than 1-million, and (3) Non-MSA.

Note: Weighted to reflect the spatial distribution of survey respondents, averaged values are derived from sets of geographic units and aggregated contexts having a total population of size that is eitherless than 100,000 persons or 100,000 persons or more, indicating: (1) number of geographic units resembling sampled geography; (2) number of geographic units in aggregated context; (3) size of totalpopulation in a geographic unit or an aggregated context (i.e., distributed across all look-alike geographic units); (4) proportion of tracts in an individual geographic units or geographic units in an aggregated context having a "high" or "low" Hard-To-Count score; (5) number of states in an aggregated context; and (6) number of square miles of land area in a geographic unit or an aggregated context (i.e., distributed across all look-alike geographic units).

Note: Details of the construction of Hard-to-Count scores are provided by Bruce and Robinson (2003). Tract-level scores are assigned to nested blockgroups and are aggregated into county- and context- level estimates. A "high" and "low" levels of error in extant data are reflected by the top and bottom quartiles of Hard-to-Count scores, as derived from tract distributions.

Note: "0.00" indicates a proportion of respondents in contexts greater than zero but less than 0.05, while "---" indicates an absolute value of zero. "---" also indicates that a particular population-size category did not apply to a set of geographies and, therefore, associated statistics were not calculated.

[1] The low number of geographic units in aggregated contexts with less than 100,000 persons reflects the selection of a limited set of relatively unpopulated "look-alike" units into these contexts.

5. Conclusion

Our study indicates that contextualized microdata may prove to be a viable method of safely distributing geographically rich information. This finding is particularly pertinent for county-level contextual information, where only 4% of survey respondents are typically located in aggregated contexts withfewer than 100,000 persons. However, more work needs to done to fully understand the implications of premises underlying population-size thresholds. While our results show the potentially important role of coverage error in ensuring the anonymity of respondents, further research is needed to create and analyze data that better captures spatial variation in undercounting for different subpopulations. This study is also limited to a single set of contextual information and two emergent haystacks. A more complete assessment is needed for a comprehensive set of haystacks as well as an expanded set of contextual information that varies in measurement composition and detail.

References

Armstrong, Marc P., Gerard Rushton, and Dale L. Zimmerman. 1999. “Geographically Masking Health Data to Preserve Confidentiality.” *Statistics in Medicine* 18: 497-525.

Cambridge University Press. (2003). *Cambridge Dictionary of American Idioms*. Cambridge University Press: Cambridge, UK.

Witkowski, Kristine M. (2016). “Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files.” Submitted to *Journal of Official Statistics*.

Zayatz, Laura. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Revised August 31, 2005: Research Report Series (Statistics #2005-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

U.S. Census Bureau, Population Division. 2002. Census 2000 PHC-T-3. Ranking Tables forMetropolitan Areas: 1990 and 2000 (Table 3: Metropolitan Areas Ranked by Population). Last Revised: July 31, 2002 < Web Page: http://www.census.gov/population/www/cen2000/phc-t3.html>

< Direct Link: http://www.census.gov/population/cen2000/phc-t3/tab03.xls>

U.S. Census Bureau, Geography Division, Cartographic Products Management Branch. 2005. *Cartographic Boundary Files*. Last Revised: August 24, 2005. <http://www.census.gov/geo/www/cob/index.html>

U.S. Census Bureau, Population Division. 2006a. *Geographic Relationship Files: 1999 MA to 2003 CBSA* (Excel file). Last Modified: August 18, 2006.
<http://www.census.gov/population/www/estimates/metroarea.html>
<Direct Link: http://www.census.gov/population/www/estimates/CBSA03_MSA99.xls>

U.S. Census Bureau. 2006b. 2000 *Census of Population and Housing, Summary File 1 (Matrices P1)* generated by Kristine Witkowski; using American FactFinder; <http://factfinder.census.gov>; (6 November 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]: 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],