# “The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tabulation Geography”1

Philip Steel and Jon Sperling

Key Words: confidentiality, disclosure risk, ZCTA, aggregation

## I. Introduction

Disclosure risk is a complex and evolving concern among statistical agencies. Increasing on-line access to statistical data is creating new demands by the public for more detailed data at multiple user-defined small area geographies. Since statistical agencies must mediate between the right of society to information (public good*)* and the right of individuals to privacy, these requests must continually be weighed against concerns of confidentiality.

The availability of detailed publicly available data at very small, and often non-hierarchical, levels of geography (e.g., census tract, ZIP Code, transportation analysis zone) for many areas of the country is a relatively new phenomenon. The ubiquity of powerful geographic information systems (GIS), advances in computer processing capabilities, more accurate data collection technologies, the prospect of more frequent data releases of small area geographies, and other ongoing changes in the statistical environment may require new perspectives on the relationship between geography and disclosure risk.

This study explores the impact of multiple geographies on disclosure risk for two of the most commonly used small area geographies, the census tract and ZIP Code.

### 1. A measure of disclosure risk

One of the easiest and most intuitive ways to measure disclosure risk both for microdata and tabular data is to count the number of unique records. For microdata this is tied directly to the probability of being able to match the record to an external data set. For tabular data it provides a measure of table sparsity by providing a count of "1"s. In theory, a record is unique if it has any unique combination of values. In practice, statisticians select a limited set of variables and the data is measured with respect to that set (Willenborg & de Waal 2001). The selected variables are often termed "keys" in disclosure avoidance literature.

Disclosure risk defined as the count of unique records with respect to a particular key is unsatisfactory in some respects; for instance it does not distinguish between random uniqueness, where the expected occurrence of a data combination is near 1, and special uniqueness where the expected occurrence is close to 0 (Elliot et al.). The treatment of the types of uniqueness may differ under different disclosure scenarios. In the simplest case, disclosure occurs in a table when a marginal "1" identifies the individual and the internal "1" provides a previously unknown data item. Because we are looking at data without sampling protection, at low levels of geography and use this very simple disclosure scenario, we will not distinguish between random and special uniqueness.

### 2. Table dimension and geography

Most tables are presented as two-dimensional matrices. A third, or paging variable, is added when three-dimensions are required. Geography is often used as a paging variable. Variables most likely to be exclusive are generally "paged". For example, one may be only interested in the income distribution of Hawaiian householders in the Los Angeles MSA. If the data is paged by race and geography, then, by excluding the rest of the US and other races, the table is immediately presentable. An equivalent strategy for reducing the burden on table users is to pack exclusions into the table’s "universe" or title. This style is becoming more popular, since it is natural to the way in which one queries a database. Another motive for placing geography in the page dimension, in addition to exclusion, is eminently practical—it’s too big to do anything else with! The Census age[7] vs. income[16] by race[9] for householders table is iterated 65,082 times (every tract in the US) in the Decennial Summary File 3, so the underlying table has 66 million cells.

The size of a particular dimension is limited by several factors. For the front page (first two dimensions) it is conventional to keep it visually limited to a physical page or to limit it to categories of interest. Different

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

needs may generate different organization of categories; the data provider must find a finer, underlying structure (a **basis**) from which to build up the display categories. For example single year of age is used as a basis for most age categorization. Demographic categories usually have a reasonably small basis from which to build the most commonly used aggregates. Even ranged numeric data can usually be generated from a basis, albeit somewhat larger; e.g. single dollar suffices for income categorization.

Modern technology allows us to create a finite basis for geography, the Global Positioning System (GPS) coordinates, but it is so detailed that at the level of publication you have the possibility of many, more or less independent, geographic hierarchies and a means to tabulate them. Prior to the widespread use of GIS/GPS technologies, tabulation was often limited to the geography under which the data was collected. That situation is changing.

### 3. Census geography

The creation of the tabulation geography is an often overlooked part of the census effort. The collection block eventually becomes the building block for all census tabulations. It is a major task to correctly assign it to all geographic entities for which data is eventually published. This includes political boundaries like county, congressional district, and municipality as well as other administrative areas such as school district, election district and transportation analysis zones. The assignment process may split the original unit to accommodate these boundaries, and this process is ongoing from initial canvass, through data collection and keying, program and quality checks with the final boundary designation occurring just before final tabulation. While GPS is not currently a part of census collection, we anticipate its use in 2010.

### 4. Postal geography

Even in the traditional geographic framework, when demand is great enough, equivalency is determined for geographies in addition to those used to link collection with tabulation geographies. A mapping is created from the basic elements of one geography to the other. Census 2000 ZIP Code geography, and its associated data, is based on generalized representations derived from the Census Bureau's Master Address File (MAF) and other sources. These new statistical areas, are referred to as ZIP Code Tabulation Areas (ZCTA).

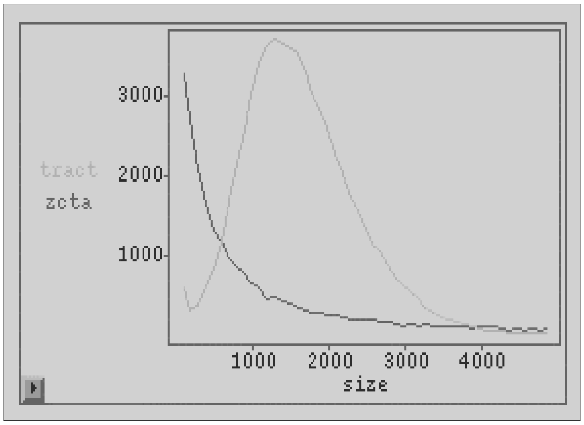
Since the re-tabulation of data uses the map of Census block to ZCTA rather than GPS coordinates of the household, it is necessarily inexact. For example in urban areas, most ZIP Code boundaries follow property lines rather than the centerlines of streets; that is, they do not coincide with census block boundaries. The situation of ZCTA and tract can be viewed as a precursor of the confidentiality problems associated with GIS. The GIS situation is somewhat worse: data published in multiple geographies with exact boundaries lack the noise associated with a manufactured equivalence.

## II. Combining geographies

### 1. Distribution of unit size

To determine the level and type of protection required for tabular disclosure prevention one must know how many units are being placed into how many cells and how evenly these units are distributed. Figure 1 looks at the geographic dimension of the census and compares the distribution of unit size between ZCTAs and tracts. The distribution of ZCTA size has an exponential shape; that is, it has many small units and the frequency of units falls quickly. The distribution of census tract size, on the other hand, is more or less normal. These differences are consistent with the origin of the two geographies.

Figure 1

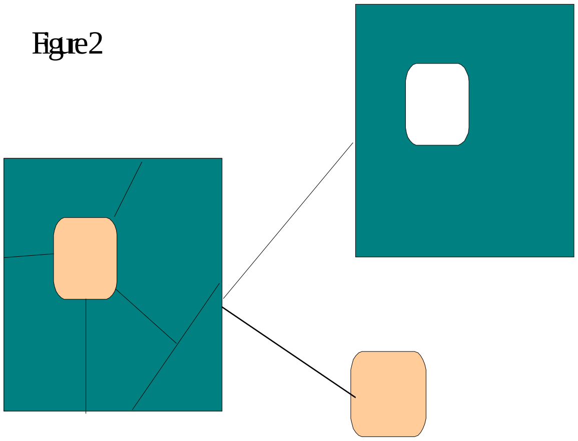


The risk associated with each of the geographies is not captured by the individual unit distributions, though certainly the concentration of small units in the postal geography is troubling. We examine the risk in a more detailed fashion subsequently, but would like to emphasize that the geographies cannot be treated separately - additional consideration must be given when publishing data in more or less independent geographies.

## 2. Subtraction geography

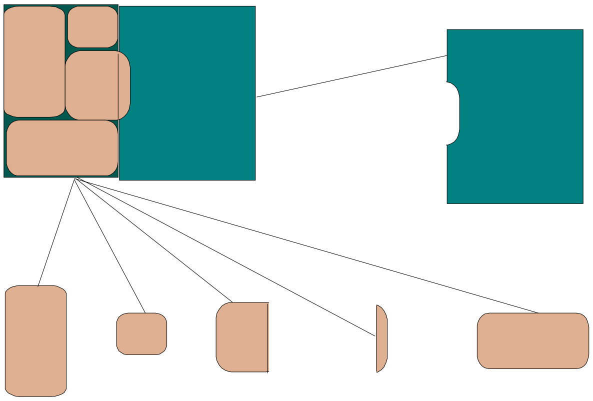
The simplest example of a subtraction geography occurs when one tabulation entity is wholly contained within another, say a tract lying completely inside a ZCTA. In this case, any table created for both tract and ZCTA also implies a table for the balance of ZCTA outside of the tract. Figure 2 shows a simple representation of this case.

Figure2



A slightly more complicated situation occurs when a tract and ZCTA differ only on a small portion of the boundary, i.e. where census geography follows one feature and ZCTA follows another and then meet up again when the features reconnect. As in the preceding case, there is a balance of ZCTA outside of the tract, but that balance of ZCTA lies inside of the neighboring tract, creating in turn a balance of tract (see Figure 3). In fact, the situation can be even more complicated--there can be a domino effect, with balances of geography obtainable through several subtractions.

Figure 3



If one publishes tables in both geographies, then there is a set of tables, on smaller geographic pieces, which is implied. The set of geographies obtainable by subtraction, with the property that they strictly contain no geography that is also obtainable by subtraction, we have dubbed the subtraction geography. Subtraction geographies can be constructed from the product geography, that is, the set of all ZCTA-tracts where ZCTA and tract physically intersect (and contain at least one unit). In this context, the condition of a tract lying inside of a ZCTA means that tract appears only once in the list of ZCTA-tracts. When it is removed from the list, a new piece may become singular in the list of ZCTA-tracts. For example, in figure 3 the balance of ZCTA is contained (and singular) within the boundary-crossing tract, once all of the undivided tracts are removed. Alternately removing singular tracts and then singular ZCTAs eventually exhausts all small subtraction entities.

After all small units are removed, every geographic unit appears two or more times in the remaining list of ZCTA-tracts. What remains of the product geography is then summarized to tract and then ZCTA. We offer without proof that the resulting geography is such that any geography obtainable by subtraction is either on the list or is some union of the geographies listed.

The subtraction geography can be used to build any tract or ZCTA but has the less desirable property that its units may overlap. For the 50 states and the District of Columbia, there are roughly as many subtraction geography entities as the sum of the number of tracts and the number of ZCTAs, but this includes 15,000 new entities, with enough small units to counter balance the addition of many large units in the average size, as indicated in the table below.

| Geography | Number of units | Average unit size | # of units size<500 |
|---|---|---|---|
| | | | |
| ZCTA | 31,969 | 3,299 | 11,319 (35%) |
| Tract | 65,082 | 1,620 | 2,835 (4%) |
| Subtraction | 96,320 | 1,613 | 16,364 (17%) |
| Product | 133,726 | 789 | 66,435 (50%) |

The problem of subtraction is one of the more challenging aspects of building "safe" on-line query tools. The proposed query tool for Census data, the proposed American Fact Finder Tier 3 web interface, constrains the users ability to modify the definition of variables (see Zayatz et al). For example, income is available in three forms, appropriate to small, medium and large populations. No splitting of income categories is then allowed.

Interestingly, geography is the only avenue for which this approach is not taken. Instead, query filters are used to determine if the geographic component is reasonable. For aggregations, the filters test each piece. The size filter can be clearly circumvented if it is applied only to the nominal geography, such as tract or ZCTA, but not to the subtraction of the two. There are several ways to address this limitation: one can add additional checks to the filter, pre-aggregate the geography or add noise to the underlying data.

The disclosure problem is not completely solved by dealing with the subtraction geography. For example, if a balance of ZCTA contains pieces of several tracts and one Hawaiian, the geography associated with that particular Hawaiian can be narrowed considerably (to the product), provided s/he is the only Hawaiian in the union of the intersected tracts. The current Census Bureau procedures guard against narrowing the nominal geography by subjecting households with unique individuals to a data swapping procedure [3]. The protection of households in small geographic areas accounted for a majority of the noise added (via swapping) to census data.

## III Risk for combined geographies

### 1. Comparing risk

We now undertake a more detailed examination of the geographies in the context of disclosure risk. A census household level summary file was combined with the map of census block to ZCTA. Nine keys were selected and the number of unique households with respect to the nine keys was determined in each tract, ZCTA, subtraction unit and product unit.

| | |
|---|---|
| Key1 | Race of head of household[7] |
| Key2 | Race member[132] |
| Key3 | Tenure[2] |
| Key4 | Number of household under 18[62] |
| Key5 | Hispanic member[2] |
| Key6 | Key1‖Key3[14] |
| Key7 | Key1‖Key5[14] |
| Key8 | Key2‖Key5[264] |
| Key9 | Key1‖Key3‖Key4‖Key5[1736] |

Keys 1-5 represent unique records in two-dimensional tables (with geography). Key2 is the collection of all single race indicators applied to the household. Keys 6-8 represent unique records in 3 dimensional tables, and key 9 represents unique records in a 5 dimensional table. The parenthetic indicates the number of cells represented.

Table 1 shows the substantial variation in total risk depending on the geography and key used. Note that 500,000 is roughly a half percent of all households (105 million). Clearly, the longer the key, the more unique records generated: the percentage of US households unique with respect to key3 is negligible in all the geographies, whereas with key9 we see a range up to roughly 1.5% in the subtraction and product geographies. Any comprehensive treatment of risk for census data would have to examine the full set of pairs in the style of keys 6 and 7 and use the risk function:

$$\sum_j \max_i \left\{ x_{ij} \right\}$$

where $x_{ij}$ indicates whether record j is unique w.r.t. key i. This is the count of records which are unique with respect to **any** key that forms a table margin, where that margin is likely to be known by an intruder. Unfortunately, implementing the maximum poses a technical difficulty for the scale of data examined here.

Clearly choice of keys is critical and can lead to different conclusions. Despite having twice as many tracts as ZCTAs, the risk as indicated by key3 (Table 1) is roughly equivalent, in contrast ZCTA fares very poorly on key5. Note also that the risk is supra-additive on key3, going from the ZCTA and tract to the subtraction geography. Risk is more proportional and more or less additive for the medium sized keys (1, 4, 6, 7) and favors ZCTA for the large keys (2, 8, 9).

Since subtraction geographic units may overlap, some unique records are double counted. This situation implies that the geography associated with the record can be narrowed to a unit below the subtraction geography (it must be in the intersection) and, in some sense increasing the risk of disclosure for that record. For the purposes of this paper, we will ignore both the double counting and the greater risk.

### 2. Risk vs. unit size

In Table 2 we examine the distribution of risk across unit size for the ZCTA geography. The geographic entities are assigned to a size range in increments of 500. As the size distribution suggests, the total risk posed by ZCTA has a large contribution coming from small units. For example, using the measure provided by key1: 43% of the risk comes from units size 1 to 500. Those units contain only 2% of the US total population.

### 3. Effect of aggregation

Aggregation is one of the most commonly used techniques to reduce disclosure risk. For example, most publications involving income categorize the values, with particular care taken for the tail of the distribution. Aggregation for geographic variables can be done many different ways and finding an optimal aggregation is nontrivial. There may be additional problems in selecting which criteria to optimize. Willenborg and deWaal (2001) have a general treatment; Karr et al (2001) have both a theoretical treatment and an implementation. Most optimization affects the utility of the resulting geography, not its disclosure risk.

The skew of the ZCTA size distribution and the concentration of risk in the small units suggests that risk could be reduced by geographic aggregation. We present the result of constraining the ZCTA size to greater than 500 households. We approximate a contiguous aggregation by associating ZCTAs falling below 500 households with the closest ZCTA that shares a common census tract and meets the minimum requirement. Closest is measured by the ZCTA numbering, which may not be indicative of adjacency. If no ZCTA with a common tract is available then numbering alone is used. Finally, aggregations are split from the original ZCTA when the balance exceeds 500. This aggregation reduced the number of ZCTAs from roughly 32,000 to 21,000 (34%) and the associated subtraction geography by 11%.

Note that the roll up of ZCTA with a minimum of 500 households does not necessarily reduce the (key1) risk by 43%, since some households would remain unique within their new, larger geography and the addition of new households may affect the uniqueness of already resident households. A new subtraction geography was calculated and unique households with respect to the 9 keys were counted. Table 3 shows the unique counts with after the rollup of ZCTA geography to a minimum of 500 and its companion subtraction geography with respect to census tract.

The risk reduction gained by the rollup is represented as the percent decrease in the number of unique records, see Table 4. While risk reduction seems worthwhile for ZCTA considered by itself, a 10% reduction (keys 6 and 7) may not be sufficient, considering that the addition of ZCTA publication increased the overall risk by nearly 50%. This does not suggest aggregation is not a good means for reducing risk in other situations, only that it must be applied to the subtraction geography. However, any aggregation of the subtraction geography forces aggregation in **both** its parent geographies.

## IV Conclusion

Geography poses special difficulty for disclosure limitation, both in its ability to define small categories and its ability to support multiple hierarchies. GIS technology sharpens these long standing problems--once data subjects have been assigned a position code, it becomes cost effective to tabulate data under any GIS supported geography. For example, re-tabulation of data after boundary changes, now undertaken only at cost and in circumstances of great need, will become remarkably easy.

GIS has allowed local governments and other entities to publish a great deal of data, both as a public service and for profit. They primarily rely on the interpretation of what is a “public” record to determine what data can be made available. In many instances, electronic publication strips away a considerable privacy protection inherent in paper records. This may cause reconsideration of what is “public”. In order to continue to expand data publication and react to changes in law, having the ability to publish aggregate data with disclosure prevention would both meet any obligation to ensure privacy and also foster public support for such practices.

We have tried to give the reader a sense of the scale and difficulty of evaluating risk for multiple geographic hierarchies. While current methods can be applied successfully, these methods are resource intensive and require a trade off with data quality: the more data requiring protection, the less reliable the statistical results derived from the protected data. Though it would by no means solve the problem completely, a standard which encouraged the creation and use of a set of tabulation geographies would reduce the problem to a more manageable scale. Such a standard would have to provide a universally agreed upon set of basic units, which are designed to be of a size appropriate for “safely” displaying data. In some cases, this might eliminate the need for other measures to ensure that aggregate data does not inadvertently breach privacy.

## References:

[1] Elliot, M., Skinner, C. and Dale, A. (1998). **Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk**, Research in Official Statistics (Vol. 1 No. 2) EUROSTAT.

[2] Willenborg, L. and de Waal, T. (2001). **Elements of Statistical Disclosure Control**. Springer.

[3] Zayatz, L, Steel, P and Rowland, S. (2000). **Disclosure Limitation for Census 2000**, Proceedings of the Section on Government Statistics, American Statistical Association meetings.

[4] Wallace, M. and Sperling, J. (2000), **User Integrated Statistical Solutions,** URISA Journal (Vol. 12 No. 4).

[5] Karr, A., Lee, J. and Sanil, (forthcoming), **Web-based Systems that Disseminate Information from Data but Protect Confidentiality**, Advances in Digital Government, A. K. Elmagarmid and W. M. McIver, editors. Kluwer, Amsterdam, 2001.

Table 1 Counts of unique records with respect to 9 keys.

| | Key1 | Key2 | Key3 | Key4 | Key5 | Key6 | Key7 | Key8 | Key9 |
|---|---|---|---|---|---|---|---|---|---|
| Tract | 37566 | 353042 | 200 | 61314 | 2238 | 98498 | 96736 | 504840 | 1150297 |
| ZCTA | 23031 | 142902 | 188 | 29836 | 3432 | 46389 | 44653 | 199233 | 438537 |
| Subtraction | 61606 | 478530 | 574 | 90003 | 6252 | 146012 | 141463 | 673375 | 1521521 |
| Product | 90191 | 504045 | 7350 | 122545 | 15065 | 188238 | 176336 | 684583 | 1511354 |

Table 2. Distribution of risk across size for the ZCTA geography.

| Unit size range | % of units in range | % of HH in range | % of key1 unique in range | % of key2 unique in range | % of key3 unique in range | % of key4 unique in range | % of key5 unique in range | % of key6 unique in range | % of key7 unique in range | % of key8 unique in range | % of key9 unique in range |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | |
| 500 | 35.4 | 2.2 | 42.9 | 18.8 | 100.0 | 32.9 | 66.4 | 33.9 | 32.3 | 15.8 | 14.2 |
| 1000 | 14.4 | 3.1 | 21.4 | 12.8 | 0.0 | 14.2 | 23.8 | 20.9 | 18.8 | 11.4 | 10.7 |
| 1500 | 7.9 | 2.9 | 9.8 | 7.8 | 0.0 | 7.7 | 6.7 | 11.7 | 9.8 | 7.1 | 7.7 |
| 2000 | 5.1 | 2.7 | 5.4 | 5.4 | 0.0 | 5.4 | 2.2 | 7.0 | 6.1 | 5.1 | 5.8 |
| 2500 | 3.8 | 2.6 | 3.2 | 4.4 | 0.0 | 3.7 | 0.7 | 4.6 | 4.1 | 4.3 | 4.8 |
| 3000 | 2.9 | 2.4 | 2.0 | 3.4 | 0.0 | 2.9 | 0.1 | 3.2 | 3.0 | 3.4 | 4.0 |
| 3500 | 2.3 | 2.2 | 1.4 | 2.9 | 0.0 | 2.3 | 0.1 | 2.1 | 2.3 | 2.9 | 3.3 |
| 4000 | 2.0 | 2.3 | 1.3 | 2.6 | 0.0 | 2.1 | 0.1 | 1.7 | 2.0 | 2.7 | 3.1 |
| 4500 | 1.9 | 2.4 | 1.2 | 2.5 | 0.0 | 2.0 | 0.0 | 1.5 | 1.9 | 2.6 | 3.0 |
| 5000 | 1.4 | 2.1 | 0.8 | 1.9 | 0.0 | 1.5 | 0.0 | 1.0 | 1.3 | 2.0 | 2.3 |
| 5500 | 1.5 | 2.3 | 0.9 | 2.0 | 0.0 | 1.7 | 0.0 | 1.0 | 1.4 | 2.2 | 2.3 |
| 6000+ | 21.5 | 72.7 | 9.7 | 35.4 | 0.0 | 23.5 | 0.0 | 11.2 | 17.1 | 40.4 | 38.8 |

Table 3 Counts of unique households for ZCTA, the subtraction geography and the corresponding geographies after rollup.

| | Key1 | Key2 | Key3 | Key4 | Key5 | Key6 | Key7 | Key8 | Key9 |
|---|---|---|---|---|---|---|---|---|---|
| ZCTA | 23031 | 142902 | 188 | 29836 | 3432 | 46389 | 44653 | 199233 | 438537 |
| Z<500 | 13489 | 118707 | 0 | 20535 | 1046 | 31653 | 30986 | 171296 | 387045 |
| Subtraction | 61606 | 478530 | 574 | 90003 | 6252 | 146012 | 141463 | 673375 | 1521521 |
| Z<500sub | 52149 | 456068 | 345 | 80804 | 3761 | 131847 | 128157 | 647850 | 1477591 |

Table 4 Risk reduction for ZCTA aggregation.

| | Key1 | Key2 | Key3 | Key4 | Key5 | Key6 | Key7 | Key8 | Key9 |
|---|---|---|---|---|---|---|---|---|---|
| ZCTA alone | 41.3% | 16.9% | 100.0% | 31.2% | 70.0% | 31.8% | 30.6% | 14.0% | 11.7% |
| Subtraction | 15.4% | 4.7% | 40.0% | 10.2% | 39.8% | 9.7% | 9.4% | 3.8% | 2.9% |