Challenges in Linking Demographic Data at Different Geographic Levels

Adrijo Chakraborty[1], Rebecca Curtis[2], Ned English[2]
Edward J. Mulrow[1], Ilana Ventura[2]
[1]NORC at the University of Chicago, 4350 East-West Highway. 8th Floor, Bethesda, MD 20814
[2]NORC at the University of Chicago, 55 E. Monroe Street, 31st Floor, Chicago, IL 60603

Abstract
Researchers can be challenged by data sets published at incongruent levels of aggregation. However, there exists the need to combine such data while maintaining its integrity and geographic relationships. We explore two approaches with trade-offs in accuracy and efficiency, with a focus on ZIP codes and US Census tracts. Our first method uses geographic information systems (GIS) to weight tract-level data from the American Community Survey (ACS) based on spatial overlap. The weights are the percent of area overlap for each tract intersecting each ZIP code. This method avoids the duplication of data caused by allocating all of a tract's data to each ZIP code it intersects and allows for a more nuanced distribution of data over matching the tract centroid to the ZIP code. Secondly, we describe a framework that uses calibration techniques to estimate overlapping regions based on published margin totals. The overlap proportion is used to allocate a portion of each tract to the ZIP code. Exploratory analysis provides insight into the strengths and weaknesses of each approach.

Key Words: GIS, geography, ACS, ZIP code, census tract, calibration

1. Challenges Linking Aggregates

Associating data sets at different levels of aggregation is a common problem. For instance, Pun, et al (2017) combined tract-level data with ZIP-level data in order to establish relationships between air pollution measures and mortality. In such cases, data are likely to have many-to-many relationships, and it is often the case that there is no simple way to relate the geographies without losing detailed relationship between areas. How units relate is important to understand due to heterogeneity in phenomena. For example, ZIP Code Tabulation Areas (ZCTA) are generally larger than census tracts, but don't follow Census geographies. Figure 1 shows ZCTA 06105, which is in Hartford, CT, along with the outlines of census tracts that intersect the ZCTA.

Figure 1: Tract/ZCTA Overlap for census tracts in and around ZCTA 06105 (in Hartford, CT). The ZCTA area overlaps with 21 tracts (some overlap obscured by ZCTA boarder thickness).

In this paper, we will concentrate on evaluating methods for linking demographic data from sources that report information at different geographic levels. For example, suppose Source 1 reports statistical estimates at the ZIP code-level, and Source 2 reports statistical estimates at the census tract-level. For linkage, either the values from Source 1 have to be converted from ZIP code estimates to tract estimates, or the estimates from Source 2 need to be converted from tract estimates to ZIP estimates.

We have encountered similar problems that we fit into two categories based on the existence of commonly reported variables.

Conditon 1. No common variables are reported for different geographic levels.
Conditon 2. Some common variables are reported for different geographic levels.

Examples of *Condition 1* come from developing countries for which detailed street-centerline databases are less common, and postal code centroids may be used for geocoding address lists. In such situations, linking postal code level survey variables to geographically reported demographics may be difficult. In El Salvador, geographic reporting units are nested as follows: Enumeration Area, Sub Sector, Sector, and Municipality, with census data reported at the enumeration area. Variables reported in census data are not available for “codigos postales” (postal codes).

An example of *Condition 2* is the reporting of US ZIP code and census tract information. The American Community Survey (ACS) provides estimates for a wide range of variables at both the ZCTA-level and census tract-level. However, a variable such as “the population percentage of those below 200% of the poverty level,” is available at the tract-level but not the ZCTA-level. If a researcher wants to link this variable to another data source with variables reported at the ZIP-level, the researcher has to use the ACS tract-

level information to determine a reasonable ZCTA-level estimate. In this case, ZCTA-level variable estimates related to the variable of interest (e.g. the "Total Population for Whom Poverty Status is Determined") might be used to improve estimates.

We consider two methods for converting estimates of variables from one geographic reporting level to another.

- **GIS-based Methodology**: A Geographic Information System (GIS) is used to determine spatial area of the two geographic units as well as the overlap between units. The ratio of the area overlap to the area of a geographic reporting level is used to apportion reported estimates from one geography to another.
  - This assumes that the variables of interest are evenly distributed across space.
  - This methodology can be used in under either *Condition 1* or *2*.

- **Calibration-based Methodology**: A calibration method (Folsom et al, 2000; Kott, 2006) is used to determine cross-tabulation cell estimates for common reported variables. The ratio of the cell value to a marginal total is used to apportion a non-common variable reported for one geographic level to the other.
  - This requires both geographies to have a common superset boundary, e.g. often a state boundary is the union of all ZCTA and also the union of all tracts.
  - This assumes the phenomena in question are distributed similarly to the variables used for the calibration.
  - This methodology can only be used under *Condition 2*.

We describe how to implement each approach, and use exploratory examples to review the strengths and weaknesses of each. Our evaluation relies on ACS 2015 census tract and ZCTA estimates reported for the state of Connecticut (CT). In practice, interest is in constructing estimates of a variable not reported for one of the two geographic areas. For evaluation purposes consider a variable, *Total Population for Whom Poverty Status is Determined*, which is report for census tracts and ZCTA. For brevity, we refer to this variable as the *Poverty Status Population*.

We first describe and evaluate a GIS-based method. This is followed by a similar review of a Calibration-based method. We conclude with some observations on the strengths and weaknesses of each method.

2. GIS-based Methodology

A GIS-based methodology uses the area of spatial overlap between two geographies to apportion a variable estimate from one geographic level to the other. In the case of ZCTA and census tracts more variable estimates are report at the tract-level than the ZCTA-level. This suggests that it is more likely that a ZCTA-level estimate will not be reported, and it will need to be constructed from tract-level estimates.

Using GIS tools, the spatial area of each tract and the overlap of each tract with a ZCTA can be determined. From this information, we derive the Tract Overlap Ratio (TOR) = $\frac{\text{Tract/ZCTA Area Overlap}}{\text{Tract Area}}$, and use it to construct a ZCTA-level estimate for each variable of interest.

We illustrate how this is done for the *Poverty Status Population* using ZCTA 06105. Figure 2 illustrates the calculation. To compute the estimate for ZCTA 06105, we apportion part of each tracts estimate to the ZTCA, and total these amounts from all overlapping tracts.

Census Tracts Overlapping ZCTA 06105	ACS 2015 Poverty Status Population (PovStatus)	Tract Area (sq mi)	Tract\ZCTA Overlap Area (sq mi)	Tract Overlap Ratio (TOR)	TOR x PovStatus
9003496700	3,985	0.50	0.00	0.00	0.76
9003496900	6,103	0.83	0.01	0.01	73.53
9003497000	4,393	0.93	0.06	0.06	283.42
9003497100	4,107	0.52	0.00	0.00	0.00
9003497200	1,700	0.69	0.00	0.00	0.00
9003497300	4,198	2.05	0.07	0.03	143.35
9003501700	1,479	0.11	0.00	0.00	0.00
9003502100	1,998	1.01	0.08	0.08	158.26
9003502900	3,311	0.30	0.00	0.00	0.00
9003503000	3,244	0.16	0.00	0.00	0.00
9003503100	3,859	0.27	0.26	0.96	3,716.07
9003503300	2,672	0.11	0.11	1.00	2,672.00
9003503500	1,608	0.09	0.00	0.00	0.00
9003503700	2,462	0.23	0.00	0.00	0.00
9003503800	587	0.72	0.01	0.01	8.15
9003504100	1,684	0.13	0.00	0.00	0.00
9003504200	5,015	0.45	0.43	0.96	4,792.11
9003504300	2,676	0.31	0.00	0.00	0.00
9003524501	2,681	0.16	0.12	0.75	2,010.75
9003524502	2,108	0.66	0.66	1.00	2,108.00
9003524600	3,228	0.50	0.46	0.92	2,969.76
					18,936.16

Figure 2: Example calculation of the *Poverty Status Population* based on reported ACS tract estimates of 21 tracts that overlap the ZCTA 06105. This method uses the Tract Overlap Ratio (TOR = $\frac{\text{Tract/ZCTA Area Overlap}}{\text{Tract Area}}$) to estimate the portion of each tracts estimate attributable to the ZCTA. The sum of the last column is the estimate.

The estimated population for whom poverty status is determined within ZCTA 06105 is a fraction above 18,936. The ACS 2105 reported value for this ZCTA is 18,972 with a margin of error[1] of 1,157. Thus, for this example, the GIS-based method provided a reasonable estimate.

Will this be true in general? We can get a sense for this by computing similar estimates for all ZCTA in CT, and comparing the estimates to ACS 2015 reported values for each of the ZCTA.

[1] Half-width of a 95% confidence interval.

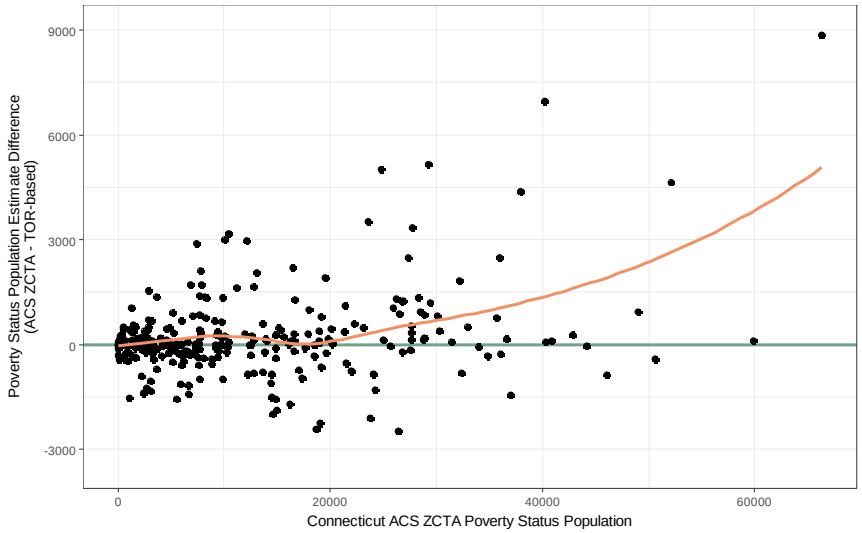

Figure 3: *Poverty Status Population Estimate Difference* (ACS ZCTA – TOR-based) by ACS ZCTA *Poverty Status Population* for CT ZCTA. A loess smoother is added to help discern a pattern in the residuals.

Figure 3 is a scatterplot of the ACS 2015 *Poverty Status Population* estimate differences (ACS estimate less the TOR-based estimate) versus the ACS 2015 *Poverty Status Population* estimate for each of the 278 ZCTA in CT. A loess smoother is added to the plot to help discern a pattern. The loess smooth suggests that for smaller ZCTA (less than 20,000) the GIS TOR-based method works well, on average. For larger ZCTA, TOR-based estimates tend to underestimate the ACS ZCTA values. A check of the differences versus that ACS ZCTA margins of error shows that only 121 out of 278 (44%) TOR-based estimates are within the margin of error.

We have not carried out an exhaustive review of the methodology and its potential, and therefore one should be careful generalizing the results. As previously noted, a GIS-based methodology might be the only alternative when *Condition 1* is true. So, a researcher should look for additional ways to evaluate GIS-based estimates before using them.

3. Calibration-based Methodology

A calibration-based methodology uses commonly reported variable estimates across two reporting levels (e.g. ZCTA and tract) as marginal control totals of a cross-tabulation to estimate a variable's value within the intersection of the two reporting levels. Cells in the cross-tabulation matrix for which the tract/ZCTA do not geographically overlap have a cell value of 0. The remaining cells are assumed to be greater than or equal to zero. A calibration algorithm is used to estimate the non-zero cell values so that row sums and column sums all equal the corresponding marginal control values. Calibration techniques are described in Folsom et al (2000) and Kott (2006).

The ACS Summary Files provide estimates of many variables at various levels of geography, including census tract and ZCTA. However, estimates for tract/ZCTA overlap are rare. If tract/ZCTA overlap estimates for a variable are derived via calibration, the proportion of this estimate relative to a known tract estimate can be determine for the variable. Similar to the TOR, we can derive the Tract Cell Proportion (TCP) =

$\frac{\text{Tract/ZCTA Calibration Estimate}}{\text{ACS Tract Total Population Estimate}}$, and use it to construct a ZCTA-level estimate for each variable of interest.

We illustrate this using the ACS 2015 commonly reported variable *Total Population*, which has estimates for all tracts and all ZCTA within the state of Connecticut. For calibration to work, there should be a common superset boundary, i.e. the union of all tracts considered should be the same as the union of all ZCTA considered. In Connecticut's case, the state is the only superset boundary.

Figure 4a is a "zoomed-out" view of a CT tract × ZCTA matrix containing 828 tracts (rows) and 278 ZCTA (columns). Shaded cells are those for which the tract and ZCTA overlap. Only a small percentage of the cells represent geographic overlap. Thus, only these cells can have non-zero *Total Population*. Figure 4b is a "zoomed-in" view that focuses on ZCTA 06105 and a few neighboring ZCTA. Eleven tracts and five ZCTA are shown.

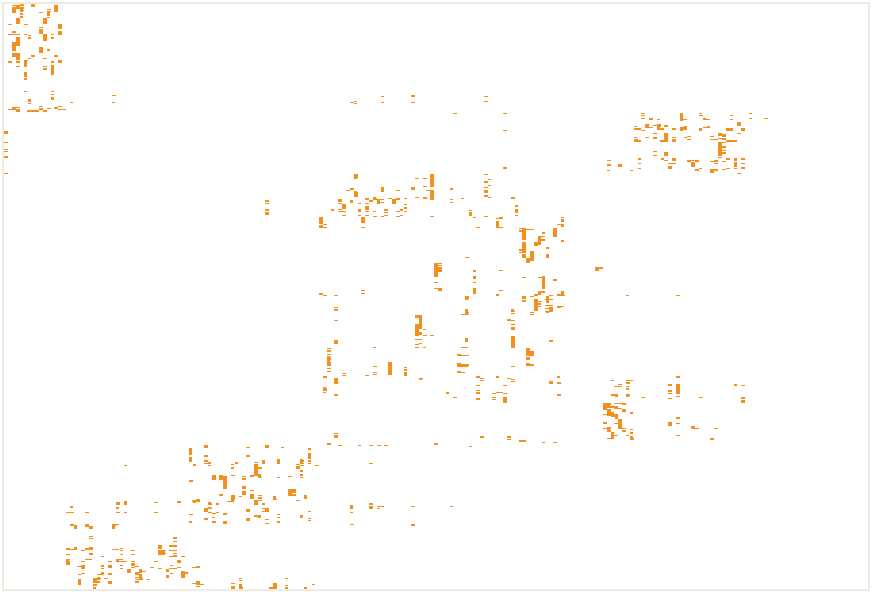

Figure 4a: Matrix view of all 828 tracts (rows) and 278 ZCTA (columns) in CT. Overlap is represented by the shaded cells. Tract/ZCTA pairs that do not overlap are unshaded cells; values for these cells estimates must be 0.

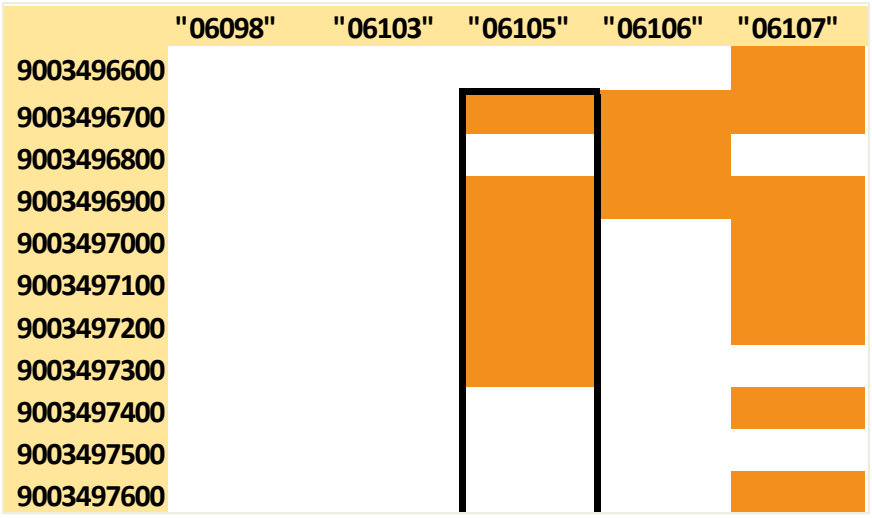

Figure 4b: Matrix view of selected CT tracts and ZCTA. Overlap is represented by the shaded cells. Tract/ZCTA pairs that do not overlap are unshaded cells; values for these cells estimates must be 0.

Many calibration algorithms exist. We used a simple raking algorithm—iterative proportional fitting (Fienberg and Meyer 2014)—to derive cell tract/ZCTA estimates for all of Connecticut. Figure 5 provides this for *Total Population* within tracts that intersect ZCTA 06105. Analogous to the TOR (GIS-based) example, we illustrate how to estimate the *Poverty Status Population* for ZCTA 06105 using the TCP and tract-level *Poverty Status Population* estimates.

The estimated population for whom poverty status is determined within ZCTA 06105 is about 18,400. The ACS 2105 reported value for this ZCTA is 18,972 with a margin of error[2] of 1,157.

Thus, for this example, the calibration-based method provided a reasonable estimate.

As was done for the GIS-based method, we can get a sense of how well this method estimates ZCTA *Poverty Status Population* by computing similar estimates for all ZCTA in CT, and comparing the estimates to ACS 2015 reported values for each of the ZCTA.

[2] Half-width of a 95% confidence interval.

Census Tract	ACS 2015 Poverty Status Population (PovStatus)	ACS Tract Total Population Estimate	Total Population Tract\ZCTA Calibration Estimate	Tract Cell Proportion (TCP)	TCP x PovStatus
9003496700	3,985	3,990	467	0.12	466.08
9003496900	6,103	6,166	908	0.15	898.69
9003497000	4,393	4,393	570	0.13	569.98
9003497100	4,107	4,399	779	0.18	727.65
9003497200	1,700	2,321	301	0.13	220.57
9003497300	4,198	4,871	822	0.17	708.57
9003501700	1,479	1,495	381	0.26	377.30
9003502100	1,998	2,001	174	0.09	173.74
9003502900	3,311	3,326	1,549	0.47	1,542.33
9003503000	3,244	3,299	1,537	0.47	1,511.12
9003503100	3,859	3,886	1,810	0.47	1,797.60
9003503300	2,672	2,675	2,675	1.00	2,672.01
9003503500	1,608	1,608	705	0.44	705.48
9003503700	2,462	2,478	1,087	0.44	1,080.16
9003503800	587	3,401	574	0.17	99.08
9003504100	1,684	1,694	789	0.47	784.44
9003504200	5,015	5,238	2,440	0.47	2,336.09
9003504300	2,676	2,761	408	0.15	395.80
9003524501	2,681	2,686	537	0.20	535.99
9003524502	2,108	2,122	295	0.14	292.88
9003524600	3,228	3,298	515	0.16	504.17
					18,399.73

Figure 5: Example calculation of the Percentage Below 200% Poverty for ZCTA 06105 based on reported ACS tract estimates of 21 tracts that overlap the ZCTA. This method uses the Tract Cell Proportion (TCP = $\frac{\text{Tract/ZCTA Calibration Estimate}}{\text{ACS Tract Total Population Estimate}}$) to estimate the portion of each tracts estimate attributable to the ZCTA. The sum of the last column is the estimate.

Figure 6 is a scatterplot of the ACS 2015 *Poverty Status Population* estimate differences (ACS estimate less the TCP-based estimate) versus the ACS 2015 *Poverty Status Population* estimate for each of the 278 ZCTA in CT. A loess smoother is added to the plot to help discern a pattern.

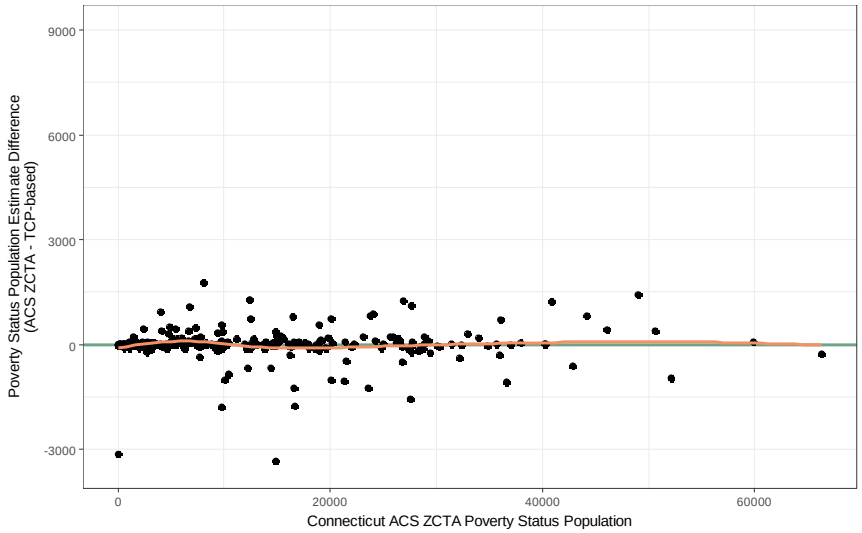

Figure 6: *Poverty Status Population Estimate Difference* (ACS ZCTA – TCP-based) by ACS ZCTA *Poverty Status Population* for CT ZCTA. A loess smoother is added to help discern a pattern in the residuals.

The loess smoothed data suggests that the calibration TCP-based method works well, on average. A check of the differences versus that ACS ZCTA margins of error shows that only 223 out of 278 (80%) TOR-based estimates are within the margin of error.

The results for the calibration-based methodology appear better than the GIS-based results. However, a general conclusion cannot be made because we have not done an exhaustive review of the methodology. As previously noted, calibration can only be used when *Condition 2* is true, i.e., some common variables are reported for different geographic levels.

4. Observations

For the examples considered, the calibration-based method is the better choice for estimating ZCTA quantities. This may be true because *Poverty Status Population* and the calibration variable, *Total Population*, are highly correlated. More investigation is needed to determine if this, and other conditions, make calibration a better method.

Based on our investigations so far, we make the following observations:

1. Both methods require special software in order to implement the methodology.
    a. GIS software such as ArcGIS, MapInfo Professional, or similar is needed, along with geographic boundary files, to calculate the area of overlap between geographies.
        i. Given you have such software, GIS-based methods are simpler to implement compared to calibration.
    b. Calibration algorithms can be complex, but are available in many common statistical software packages such as R.
2. Preliminary investigations indicate that a GIS-based methodology can be improved when *Condition 2* is true; that is, some common variables are reported for each of the geographies. For example, if the population percentage below the poverty level is of interest, the total population below the poverty level (the

numerator) may need to be estimated using the GIS-based methodology, but the denominator, which is the *Poverty Status Population*, is available for all ZCTA. This should be more accurate than estimating the percentage from tract report percentages.

3. As previously noted, a GIS-based methodology may be the only choice under *Condition 1*. If there are no commonly reported variables across geographies, a calibration-based method is not feasible.
4. Calibration-based methods can be extended to use more than one set of control totals. For example, demographic and economic variables can be included as marginal control totals. We conjecture that calibration-based estimates will improve with the use of more control totals. If the variable of interest is related to one or more control total variables, the variance will likely be reduced.

References

Fienberg, Stephen E., and Michael M. Meyer (2014). Iterative proportional fitting including generalizations. Accessed 10 1, 2017. http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat00364/pdf.

Folsom, R. E., and Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In Proceedings of the Section on Survey Research Methods, 598–603 American Statistical Association, Alexandria, VA.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, 32(2), 133-142.

Pun VC, Kazemiparkouhi F, Manjourides J, Suh H.H. (2017). Long-Term PM2.5 Exposures and Respiratory, Cancer and Cardiovascular Mortality in American Older Adults. Am J Epidemiology. 2017 May 24. doi: 10.1093/aje/kwx166.