SpAM

SpAM (Spatial Analysis and Methods) presents short articles on the use of spatial statistical techniques for housing or urban development research. Through this department of Cityscape*, the Office of Policy Development and Research introduces readers to the use of emerging spatial data analysis methods or techniques for measuring geographic relationships in research data. Researchers increasingly use these new techniques to enhance their understanding of urban patterns but often do not have access to short demonstration articles for applied guidance. If you have an idea for an article of no more than 3,000 words presenting an applied spatial data analysis method or technique, please send a one-paragraph abstract to rwilson@umbc.edu for review.*

Understanding and Enhancing the U.S. Department of Housing and Urban Development's ZIP Code Crosswalk Files

Ron Wilson
University of Maryland, Baltimore County

Alexander Din
Maryland Department of Housing and Community Development

Abstract

ZIP Codes[1] are commonly used for mapping, spatial analyses, creating tables, or other reporting products. Used for these tasks, the results from using these geographies often are distorted because of adverse statistical properties inherent with ZIP Codes. Summarizing ZIP Code data to other large geographies (for example, county, Core Based Statistical Area, state) associates them with these other geographies to create aggregate counts so that metropolitan or county rankings can be reported. This process requires ZIP Codes to be properly allocated to these other geographies to accurately associate a record with that area. Although some companies or government organizations already provide a crosswalk to these geographies, the allocation method used is unclear, leaving it indiscernible as to the accuracy of the assignment of ZIP Codes. In this article, we demonstrate how to use the U.S. Department of Housing and Urban Development

[1] ZIP Code is the acronym for Zoning Improvement Plan Code.

Abstract (continued)

(HUD) United States Postal Service ZIP Code Crosswalk Files to more directly control the ZIP Code allocation process of records to alternative geographies. In meeting this objective, we also provide results of an analysis using the HUD Crosswalk File in associating a ZIP Code with U.S. counties.

ZIP Codes Are Problem Geographies

Organizations use ZIP Codes for many analytical tasks, such as to verify addresses, allocate resources, or create analytical products (for example, maps, tables, or conduct reports). Although ZIP Codes have legitimate—but limited—use in analysis, they have adverse effects on the results. ZIP Codes are problematic, because their boundaries are not created for analytical purposes like other geographies. ZIP Codes were designed to more efficiently deliver mail, not as geographies to be used for analysis. Because of their nature, the boundaries vary in size and shape that amplifies a common, adverse statistical problem when used for analysis. This effect, known as the Modifiable Areal Unit Problem (MAUP), is ever present in analyses that use geography. It is well documented in several studies on how ZIP Codes are notorious for distorting policy-related analyses (Beyer, Schultz, and Rushton, 2007; Cudnick et al., 2012; Dai, 2010; Grubesic and Matisziw, 2006; Hipp, 2007; Krieger et al., 2002; Montalvo and Reynal-Querol, 2017; Wilson, 2015). Further, when mapping ZIP Code data, the choice of the thematic mapping method used to display the data can further exacerbate the misrepresentation of results through erroneous patterns depicted in the map (Wilson, 2011). A final deficiency in the use of ZIP Codes for analysis is that they typically do not contain any social, demographic, or economic data that can be used to create contextualized statistics of ratios, percentages, rates, or densities from the record counts contained within. When data are provided, those values are distorted from the same aggregation problems mentioned in the following paragraphs.

First, grouping characteristics by an area boundary that is too large, oddly shaped, or a combination of both leads to summary statistics that may not be representative of the population within the boundary. A second adverse effect is one that can hide a pattern of extreme values on either end of the characteristic values (Wilson, 2013). In this instance, extreme differences that exist for a characteristic are canceled out because the summary statistic represents the norm and does not reveal either extreme. A third adverse effect, and related to the second,[2] is the reversal of

[2] For both the second and third adverse effects of ZIP Codes use, the phenomenon is known as Simpson's Paradox that occurs from aggregation bias. Simpson's Paradox occurs when data are aggregated to groups and prevents any analysis from detecting underlying patterns between cases that share a commonality. This problem occurs because statistical techniques treat combined observations as a single group, in which the observations are assumed to be indistinguishable from one another and not have any shared commonalities that might affect the results. However, cases often do share commonalities that may have converse relationships between an outcome and the explanatory factors when analyzed as a subgroup. The repartitioning of the underlying data from smaller to larger geographic units can cancel out or reverse patterns in smaller units. The paradox is a consequence of the MAUP, in which statistical results are affected by modifications to the geographic unit's boundary size or shape or both. Aggregated data are uniquely partitioned by their geography, and when geographic units are changed, the new boundary sizes and shapes are repartitioned.

a relationship if two characteristics are examined together (Hipp, 2007; Montalvo and Reynal-Querol, 2017; Wilson, 2015). Here, a positive or negative relationship between two characteristics is reversed from what it would be if a more appropriate area were used in the analysis because the characteristics would have been assigned to differing areas.

About Allocating ZIP Code Data to Other Geographies

Several private companies obtain address data from the U.S. Postal Service (USPS) or ZIP Code geographies from the U.S. Census Bureau to enhance them and sell the information. Companies in the private industry add value to these products, making them more robust by adding geographic information or the creation of boundaries for mapping.[3] The boundaries created are estimated or modified by delineating areas using topographical point or line landmarks following each organization's proprietary method, including comparisons with the ZIP Code Tabulation Areas (ZCTAs) provided by the Census Bureau. Each company has their own undisclosed method of creating these boundaries, with each claiming theirs is the most accurate.

ZIP Codes typically overlap with other geographies and cannot always be completely associated with areas in another geography. Exhibit 1 demonstrates how a decision must be made in allocating address counts from ZIP Codes to the census tract geography. The map shows a ZIP Code (light gray outline) crosscutting three census tracts (dark gray outline) and the geographic distribution of addresses within each tract. For many analyses, address counts must be associated with only one other geography, lest the addresses be counted multiple times and adversely affect statistical results. The ZIP Code in exhibit 1 has 2,860 addresses, with 1,962 (65.8 percent) in tract C, 872 (30.5 percent) in tract B, and 26 (1.0 percent) in tract C. If the address locations are available, then those proportion counts get assigned to each tract.

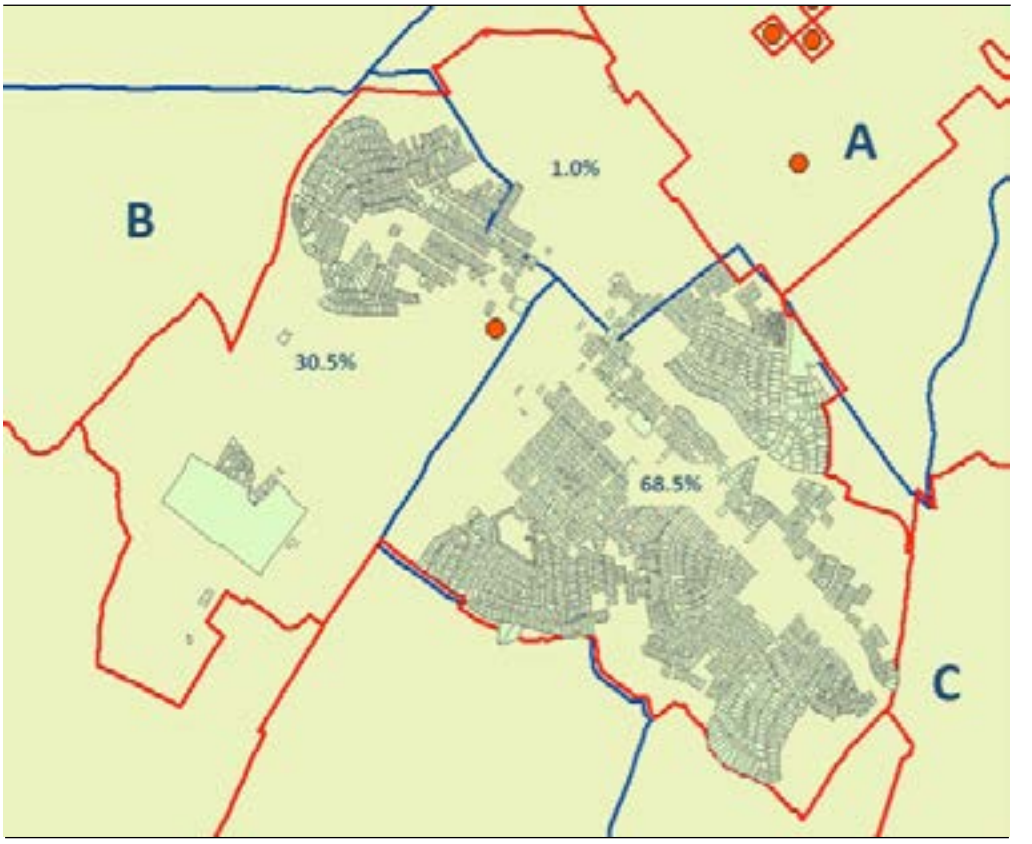
However, without knowing the location of the address, a decision must be made to either assign all the addresses within a ZIP Code to one of the other geographies or proportionally divide the addresses to each of the three tracts.

If all 2,860 of the addresses in the ZIP Code are assigned to one of the tracts without knowing where the actual addresses are, the counts will be assigned by using one of two general methods that assigns the ZIP Code addresses to other geographies based on the geometric properties of the areas. The first method uses the center location (large gray points) of the ZIP Code—that is, the centroid—and assigns the full count of addresses to the geography in which that center is located. In the previous example, the ZIP Code will be assigned to tract B, because the center is slightly inside the boundary next to tract A. One variant of this approach is the simple assigning of address counts to the geography, with the largest proportion of an area that the ZIP Code crosscuts. In the previous example, the ZIP Code will be assigned to tract C, because it has the largest portion of its area in that county. The second approach assigns a proportion of address counts that is commensurate with proportion of tract areas inside the ZIP Code. In the previous example, tract C will get

[3] The USPS has never made a ZIP Code boundary file available to the public.

Exhibit 1

ZIP Code Boundary Crosscut With Census Tracts



54 percent of the addresses, 33 percent of address will go to tract B, and the remaining 13 percent will go to tract A. Regardless of method, allocating from either of these methods has varying risks to numerical accuracy of the allocation.

Other ZIP Code products provide crosswalks to other geographies that use a similar geocoding method as HUD. An examination of ZIP Code Download, ZIP Boundary, and the HUD data sets shows that the difference between identified counties and Core Based Statistical Areas (CBSAs) is less than 1 percent of each other in regard to which geographies the ZIP Codes become associated. Therefore, the differences between them are negligible. The HUD product, however, is the only one of the three that crosswalks ZIP Codes to census tracts, making it valuable for micro-analyses.

Geocoding Address Data and Identifying Related Geographies

HUD can overcome the aforementioned analytical problems, because the agency has a Geocode Service Center (GSC) that provides a high level of accuracy in assigning corresponding geographies

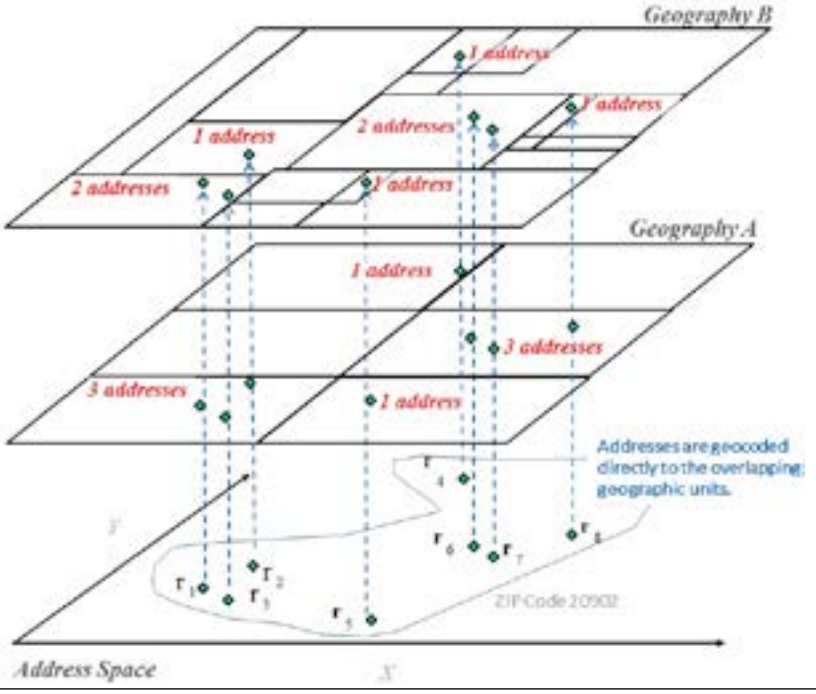
to the addresses received from USPS. On a quarterly basis, HUD receives domestic address data from USPS that are sent to the GSC to match, standardize, and return geographic information that staff can use in HUD products. This additional information offers a wide selection of geographic information to meet the many varied analytical questions the agency asks from its data systems. Diagnostics are also returned from the GSC helping to ensure the data are correctly geocoded, enabling staff to evaluate the quality of address matching and geographic positioning of the location.

Having the actual locations of the addresses facilitates HUD's capabilities in accurately analyzing applicant data with mapping and spatial analysis. With the longitude (X) and latitude (Y) that the GSC produces, mapping and spatial analysis can be conducted directly on the locations, eliminating error when examining groups.

The GSC also returns the names and codes of geographies in which the addresses fall so that other types of analyses can be done that will not require the locations. With location geocoding of addresses, an address can be directly associated to other geographies to which they are contextually associated. For example, an address can be associated to the census block, block group, tract, county, and CBSA they are within or other non-Census geographies that are available. These geographies can facilitate the (1) simple tabulating of counts by geography, (2) mapping or spatial analysis of socioeconomic data, or (3) creation of geographic data for advanced statistical modeling. Exhibit 2 demonstrates this spatial overlay process.

Exhibit 2

Geocoding Process of Addresses by ZIP Code to Overlapping Geographies



HUD cannot release the geocoded USPS address data at the location level, but the agency can release a file that takes advantage of this geocoding to produce a file that can crosswalk ZIP Codes to several other geographies so that other agencies may use those geographies instead of using ZIP Codes as analytical units. The value of HUD's USPS ZIP Code Crosswalk Files is that they enable users to escape the adverse effects of using ZIP Codes by allocating the data to more appropriate geographies for analysis.

HUD USPS ZIP Code Crosswalk Files

HUD produces USPS ZIP Code Crosswalk Files that correspond to four Census Bureau geographies, which are (1) census tracts, (2) counties and county equivalents, (3) CBSAs, and (4) congressional districts. Each file contains two general pieces of information.[4] The first is the ZIP Code and geographic identification of the corresponding geography. The second is the ratios of address type in the geography the ZIP Code overlaps. Each crosswalk file contains the following address type ratios: (1) residential, (2) business, or (3) other. Exhibit 3 contains the full descriptions of the each element in each of the crosswalk files.[5]

Because many ZIP Codes overlap the boundaries of the other geographies, duplicate ZIP Code records will exist, requiring the user to make a decision about which geography to associate the ZIP Code. Exhibit 4 shows the structure of the HUD ZIP Code to County Crosswalk File. As an example, in Mt. Airy, Maryland, four records for ZIP Code 21771 are highlighted, with each recording the proportion of residential addresses in that ZIP Code. That ZIP Code overlaps four counties, with residential ratios distributed at 0.4652 (47 percent), 0.4143 (41 percent), 0.1088 (11 percent), and 0.0117 (1 percent), all adding up to 1.0 (100 percent).

Exhibit 3

Data Elements in the HUD USPS ZIP Code Crosswalk Files

Element	Description
ZIP	5 digit USPS ZIP code
TRACT	11 digit unique 2000 or 2010 Census tract GEOID consisting of state FIPS + county FIPS + tract code. The decimal is implied and leading and trailing zeros have been preserved.
COUNTY	5 digit unique 2000 or 2010 Census county GEOID consisting of state FIPS + county FIPS.
CBSA	5 digit CBSA code for Micropolitan and Metropolitan Areas as defined by OMB in February of 2013. ZIP codes with a CBSA code of '99999' are not located within a CBSA. In Metropolitan Areas that are broken out into Divisions, the code reported is the CBSA Metropolitan Division code.
RES_RATIO	The ratio of residential addresses in the ZIP – Tract, County, or CBSA part to the total number of residential addresses in the entire ZIP.
BUS_RATIO	The ratio of business addresses in the ZIP – Tract, County, or CBSA part to the total number of business addresses in the entire ZIP.
OTH_RATIO	The ratio of other addresses in the ZIP – Tract, County, or CBSA part to the total number of other addresses in the entire ZIP.
TOT_RATIO	The ratio of all addresses in the ZIP – Tract, County, or CBSA part to the total number of all types of addresses in the entire ZIP.

CBSA = Core Based Statistical Area. FIPS = Federal Information Processing Standards. GEOID = geographic identification. HUD = U.S. Department of Housing and Urban Development. OMB = Office of Management and Budget. USPS = United States Postal Service.

[4] huduser.gov/portal/datasets/usps_crosswalk.html.

[5] More details can be found on the HUD USPS ZIP Code Crosswalk Files website at huduser.gov/portal/datasets/usps_crosswalk.html - codebook.

Exhibit 4

Example of the ZIP Code to County Crosswalk File

	ZIP_CODE	COUNTY_CODE	res_ratio	bus_ratio	oth_ratio	tot_ratio
9597	21766	24001	1.000000000	1.000000000	0.000000000	1.000000000
9598	21767	24043	1.000000000	1.000000000	1.000000000	1.000000000
9599	21769	24021	0.995981246	0.988304094	1.000000000	0.995749203
9600	21769	24043	0.004018754	0.011695906	0.000000000	0.004250797
9601	21770	24021	1.000000000	1.000000000	1.000000000	1.000000000
9602	21771	24013	0.465155904	0.807650273	0.752212389	0.492245026
9603	21771	24021	0.414295451	0.155191257	0.230088496	0.394059686
9604	21771	24027	0.108877151	0.033879781	0.017699115	0.102694658
9605	21771	24031	0.011671494	0.003278689	0.000000000	0.010966630
9606	21773	24021	1.000000000	1.000000000	1.000000000	1.000000000
9607	21774	24021	1.000000000	1.000000000	1.000000000	1.000000000
9608	21775	24021	0.000000000	1.000000000	0.000000000	1.000000000
9609	21776	24013	0.879154079	1.000000000	1.000000000	0.884915742
9610	21776	24021	0.120845921	0.000000000	0.000000000	0.115084258
9611	21777	24021	1.000000000	1.000000000	1.000000000	1.000000000
9612	21778	24021	1.000000000	1.000000000	0.000000000	1.000000000
9613	21779	24043	1.000000000	1.000000000	0.000000000	1.000000000
9614	21780	24021	0.947813822	1.000000000	0.000000000	0.948965517
9615	21780	24043	0.052186178	0.000000000	0.000000000	0.051034483
9616	21781	24043	0.000000000	1.000000000	0.000000000	1.000000000
9617	21782	24043	1.000000000	1.000000000	1.000000000	1.000000000
9618	21783	24021	0.137710633	0.061728395	0.000000000	0.135540464
9619	21783	24043	0.862289367	0.938271605	1.000000000	0.864459536
9620	21784	24013	0.939181882	0.990284525	0.990811639	0.945208556
9621	21784	24027	0.060818118	0.009715475	0.009188361	0.054791444

The res_ratio, bus_ratio, and oth_ratio columns show the proportions of the corresponding address type within each county. The tot_ratio column is the total number of all address types within each county. Again, for ZIP Code 21771 and rounding up at the fourth decimal place, from top to bottom, the tot_ratio column is 0.4922 + 0.3931 + 0.1027 + 0.011 = 1.0.

The ratios can be used to help decide which county to assign to the ZIP Code. For ZIP Codes with ratios of 1.0, no decision needs to be made, because all the addresses are contained within a single county. For ZIP Codes with ratio proportions, a decision can be made to assign the addresses to a county based on one of two approaches. The first approach is to assign all addresses to the county with the largest ratio. With probability theory, any address with that ZIP Code has a greater chance of being in the county with the largest proportion of addresses.

A second approach is to proportionally assign the addresses to each county through geoprocessing in a geographic information system or cross-tabulating in statistical or general database software. This approach can allocate addresses in one of two ways. The first approach is to randomly assign each address to a county by proportion if only counts, rates, and densities are all that are needed for analysis. In this instance, it would not matter if an address was assigned to the wrong county, because only the aggregate value from its assignment is used and not a characteristic such as sex, ethnicity, or age. The second approach is to use an additional geography file, such as a street

network, to determine if the street of the address was primarily in one county or a neighboring county. This file is used if the desire is to have more accurate counts or rates for characteristics such as sex, ethnicity, or age.

Examining Geographic Overlap in the HUD USPS ZIP Code Crosswalk Files

The number of overlaps between a ZIP Code and another geography is a function of scale. The smaller the overlapping geography is, greater is the overlap of the ZIP Codes. Exhibit 5 shows the total number of ZIP Codes overlaps with the CBSA, county, and census tract boundaries. For the ZIP Codes that overlap multiple geographic boundaries, tracts have the smallest number of no overlaps—about 36 percent, 74 percent for counties, and 83 percent for CBSAs. These data translate into the chance an address is not in the boundary (error) being 61 percent for tracts, 26 percent for counties, and 17 percent for CBSAs. For CBSAs and counties, the opportunity for error is even lower if the proportion of addresses in one of the overlaps is examined.

For counties, the distribution of residential ratios also shows a low potential for error in misassociation of county to ZIP Code. Of the 26 percent of residential address that are in multiple counties, 16 percent of residences are 90 to 99 percent in one area, four percent are 80 to 89 percent in one area, leaving 6 percent of residences 30 to 79 percent in one or more counties. The crosswalking of ZIP Codes to census tracts are of particular concern when assigning a one-to-one relationship.

Exhibit 5

ZIP Code Overlap Results Across Three Geographies

ZIP Code Overlap Counts, by Geography						
	Tract		County		CBSA	
Cross-cuts	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*
None	14,106	35.89	29,107	73.8	32,908	83.4
2	5,730	14.58	7,772	19.7	5,658	14.3
3	4,313	10.97	2,141	5.4	829	2.1
4	3,090	7.86	399	1.0	66	0.2
5	2,218	5.64	36	0.1	1	0.0
6	1,550	3.94	6	0.0	0	0.0
7	1,226	3.12	0	0.0	0	0.0
8	1,059	2.69	0	0.0	0	0.0
9	930	2.37	0	0.0	0	0.0
10	816	2.08	1	0.0	0	0.0
11 to 68	4,263	10.86	0	0.0	0	0.0
Total	39,301	100	39,462	100.0	39,462	100.0

CBSA = Core Based Statistical Area.

Exhibit 6 shows the residential ratio distribution across overlapping ZIP Codes.[6,7] The exhibit gives an indication of the chance an address will be misassociated with a particular geography when a ZIP Code overlap occurs. For CBSAs and counties, the potential for error is minimal, with most of the residential ratios being largely in one area more than the others.

Of the 17 percent of residential addresses that are in multiple CBSAs, about 10 percent of residences are 90 to 99 percent in one CBSA, 3 percent are 80 to 89 percent in one, leaving only 4 percent of residences 50 to 79 percent in one or more CBSAs—with the possibility of being outside of a CBSA, because they do not completely cover the United States.

Exhibit 6

Nonduplicate ZIP Code Results Across Three Geographies

Residential Ratio Distribution of Zip Code Overlap, by Geography						
	Tract		County		CBSA	
Res Ratio	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*
0 (Outside)	4,897	12.5	4,938	12.5	4,938	12.5
0.01 to 0.04	0	0.0	0	0.0	0	0.0
0.05 to 0.09	124	0.3	0	0.0	0	0.0
0.10 to 0.14	1,003	2.6	0	0.0	0	0.0
0.15 to 0.19	1,778	4.5	0	0.0	0	0.0
0.20 to 0.29	3,183	8.1	1	0.0	0	0.0
0.30 to 0.39	2,327	5.9	20	0.1	6	0.0
0.40 to 0.49	2,059	5.2	91	0.2	32	0.1
0.50 to 0.59	2,688	6.8	633	1.6	378	1.0
0.60 to 0.69	2,217	5.6	741	1.9	433	1.1
0.70 to 0.79	2,216	5.6	1,002	2.5	607	1.5
0.80 to 0.84	1,272	3.2	695	1.8	435	1.1
0.85 to 0.89	1,378	3.5	900	2.3	574	1.5
0.90 to 0.94	1,547	3.9	1,409	3.6	926	2.4
0.95 to 0.99	2,971	7.6	4,724	12.0	3,110	7.9
1 (Inside)	9,641	24.5	24,308	61.6	28,023	71.0
Total	39,301	100	39,462	100.0	39,462	100.0
Total Whole:	14,538	39.3	29,246	74.1	32,961	83.5
Total Part:	24,763	60.7	10,216	25.9	6,501	16.5

CBSA = Core Based Statistical Area.

[6] Duplicate ZIP Code records were removed by keeping only the ZIP Code with the highest residential ratio in a geography, the geography to which all the ZIP Codes would get assigned.

[7] The ZIP Codes are those that are outside of a county or census tract, or are in some other geographic boundary other than a county or its equivalent in U.S. territories or military bases.

The distribution in exhibit 6 shows that of the 61 percent of residential address that are in multiple tracts, only about 12 percent of residences are 90 to 99 percent in one area, 7 percent are 80 to 89 percent in one area, but 45 percent of residences are in one tract or another. These percentages translate into approximately of one-third of ZIP Codes having the greatest chance of an address being assigned to the wrong census tract if a one-to-one approach is used in associating ZIP Codes to a tract using the highest residential proportion of addresses.

Even more precarious is that some ZIP Codes are split among numerous tracts, leaving the highest residential ratio less than 50 percent, with the other ratios similar to the other tracts. Exhibit 7 shows the records for ZIP Code 36067 in Autauga County, Alabama, which overlaps 13 census tracts. Exhibit 8 shows the size and shape disparity among the geographies.

Exhibit 7 shows the highest tract residential ratio to be in tract 0100102082 at only approximately 23 percent, followed by four tracts with 10 to 14 percent. The remaining eight tracts have between 1 and 7 percent of residential addresses. This even distribution makes the assignment of a ZIP Code to any of these tracts dubious, because 77 percent of the remaining residential population is spread over a large enough area to have equal representation of that ZIP Code.

However, unlike assigning a ZIP Code to county, CBSA, or congressional district geographies, where each will get a ZIP Code assignment, this approach has an analytical consequence when using census tracts. Because tracts are often smaller than ZIP Code coverage, a problem of incomplete coverage occurs—that is, geographic holes form. Exhibit 9 shows the census tracts in the Mid-Atlantic states after assigning the ZIP Codes to the tract with the highest residential ratio. The dark gray polygons are the tracts assigned ZIP Codes because they had the highest residential ratio, with the light gray areas being the holes in which a tract was not associated with a ZIP Code.

The geographic holes occur because the ZIP Codes in those areas are larger than the tracts, as exhibit 6 indicates of the more even distribution of residential ratios. When a ZIP Code is assigned to

Exhibit 7

Census Tract Address Distribution for ZIP Code 36067, Autauga County, Alabama

	ZIP_CODE	TRACT_CODE	res_ratio	bus_ratio	oth_ratio	tot_ratio
1	36067	01001020100	0.061892061	0.030518820	0.057471264	0.059618842
2	36067	01001020200	0.072874176	0.309257375	0.206896552	0.091262136
3	36067	01001020300	0.111939128	0.109867752	0.264367816	0.113700108
4	36067	01001020400	0.003686853	0.093591048	0.022988506	0.010284070
5	36067	01001020500	0.000392218	0.016276704	0.000000000	0.001510248
6	36067	01001020600	0.138766865	0.083418108	0.172413793	0.135275081
7	36067	01001020700	0.126843426	0.207527976	0.074712644	0.131895002
8	36067	01001020801	0.107075620	0.046796524	0.063218391	0.102265372
9	36067	01001020802	0.232271729	0.072227874	0.103448276	0.219345559
10	36067	01001020900	0.062558055	0.008138352	0.028735632	0.058324344
11	36067	01001021000	0.042751804	0.006103764	0.000000000	0.039626034
12	36067	01001021100	0.038751177	0.007121058	0.005747126	0.036102122
13	36067	01051031300	0.000156887	0.009155646	0.000000000	0.000791082

Exhibit 8

Census Tract Overlap With ZIP Code 36067, Autauga County, Alabama

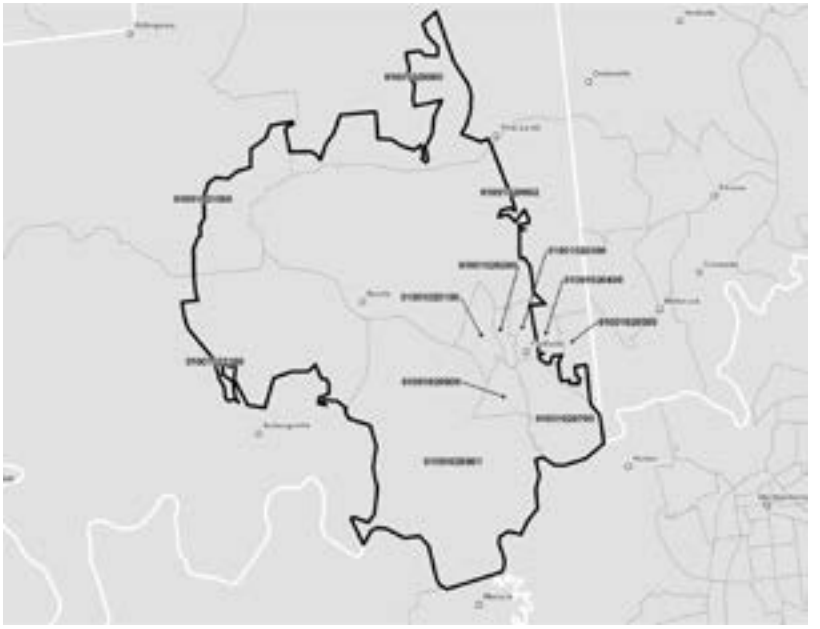
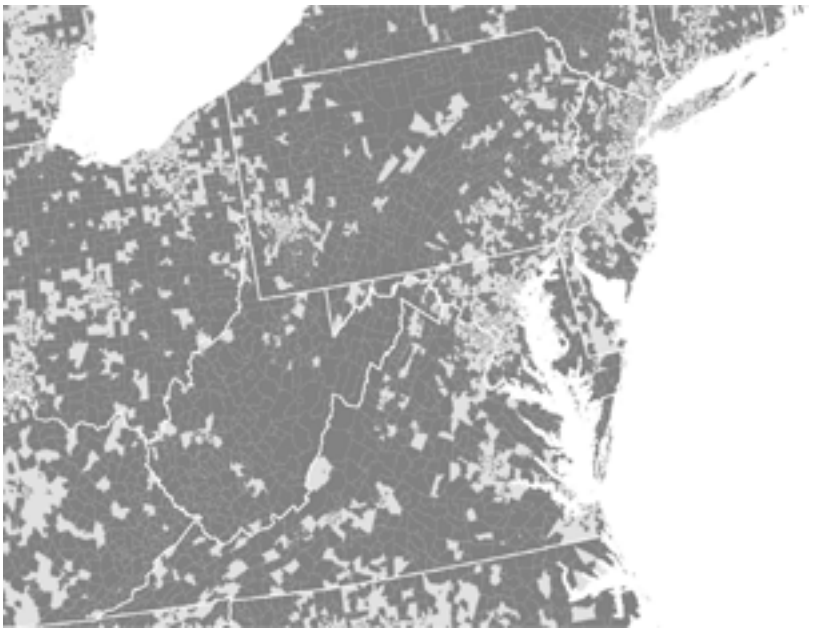


Exhibit 9

Remaining Census Tracts After ZIP Code Assignment of Highest Residential Ratio



one tract, all the adjacent tracts are omitted from an analysis, because they have no associated ZIP Code. As seen, this omission primarily occurs in urbanized areas. Unless the analytical objective is to only know which tract has the highest residential ratio, a proportion assignment solution must be found to assign a ZIP Code to all overlapping tracts with a one-to-many approach.

A potential way to approach the use of the ZIP Code to tract Crosswalk File is to consider less the idea of a one-to-one match with a census tract, but instead take a many-to-one approach that connects the ZIP Code to an associated place name, and assign it instead. This approach changes the geographic relationship from an assignment to a single overlapping geography to one of proximity to a recognized place with a name preferred by USPS and the Census Bureau. The tracts and any related socioeconomic data and all ZIP Codes associated with the place can be used in a more representative way. With that relationship, the data can be linked to the Census Bureau place-name geography and mapped, tabulated, or spatially analyzed with distance-based statistics.

Further, if the ZIP Code to census tract file is not reduced to a single one-to-one match to a tract, then the file can be linked to the ZIP Code Tabulation Area or other ZIP Code boundary file and geoprocessed to integrate the two layers and provide the proportions of addresses in each tract across all the overlapping ZIP Codes. Then, the tract Federal Information Processing Standards (FIPS) or geographic identification can be summarized to create aggregate residential ratios to the tracts, thereby transferring the residential ratios—and other ratios—to each tract.

Enhancing and Evaluating the HUD USPS ZIP Code Crosswalk Files

The Crosswalk Files contain only the most basic information about the geographies the ZIP Codes overlap, which are Census Bureau FIPS identification. Without the geography names, it is difficult to know to which of the corresponding parts of the United States the data refer. However, with a geographic identification, the Crosswalk Files can be linked to a number of other data sources to create a file that contains more robust and contextual information about the ZIP Code. As mentioned, several other data sources are available that provide additional information that can be matched to the Crosswalk Files. Companies, such as ZIP Boundary[8] and ZIP Code Database Org[9] provide information that contains additional characteristics associated for each ZIP Code. In addition, SAS® software users have access to ZIP Code data that are provided in the SAS Help library, which SAS purchases from ZIP Code Download.[10] These additional sources of information can enhance and evaluate the HUD USPS ZIP Code Crosswalk Files analysis.

Enhancing a Crosswalk File

The SAS[11] ZIP Code data are used to demonstrate enhancing the HUD Crosswalk File with U.S. counties and equivalents. ZIP Code data from the companies mentioned previously provide similar

[8] http://www.zipboundary.com/.

[9] https://www.unitedstateszipcodes.org/zip-code-database/.

[10] https://zipcodedownload.com/.

[11] SAS ZIP Codes are in the SAS Help library, to which the data originate from ZIP Code Download https://zipcodedownload.com/.

variables to create a geographic context for the HUD ZIP Code Crosswalk Files. Exhibit 10 lists the additional information the SAS data provide. These enhancements offer the opportunity to geographically contextualize the HUD Crosswalk File, as well as facilitate a comparison of the accuracy with other crosswalk files. ZIP Codes and their geographies change frequently, and each company processes the updates differently, which means similar analyses will yield slightly different results.

A number of variables in the SAS file are valuable when examining the HUD United States Postal Service cross-walk files. First is the COUNTYNM variable that contains the name of the county that corresponds with county code in the HUD cross-walk file. Note that the Metropolitan Statistical Area (MSA) is contained in the SAS data, but not the Core Based Statistical Areas information. Using the county code, this data can be linked with the CBSA listing from the Census Bureau to add that information.

Exhibit 10

Additional ZIP Code Information From the ZIP Code Database

SAS ZIP Code Table Variables		
#	Variable	Type
1	ZIP	The 5-digit ZIP Code
2	X	Longitude (degrees) of the center (centroid) of ZIP Code.
3	Y	Latitude (degrees) of the center (centroid) of ZIP Code.
4	ZIP_CLASS	ZIP Code Classification:P=PO Box U=Unique zip used for large organizations
5	CITY	Name of city/org
6	STATE	Two-digit number (FIPS code) for state/territory
7	STATECODE	Two-letter abbrev. for state name.
8	STATENAME	Full name of state/territory
9	COUNTY	FIPS county code.
10	COUNTYNM	Name of county/parish.
11	MSA	Metro Statistical Area code by common pop-pre 2003; no MSA for rural
12	AREACODE	Single Area Code for ZIP Code.
13	AREACODES	Multiple Area Codes for ZIP Code.
14	TIMEZONE	Time Zone for ZIP Code.
15	GMTOFFSET	Diff (hrs) between GMT and time zone for ZIP Code
16	DST	ZIP Code obeys Daylight Savings: Y-Yes N-No
17	PONAME	USPS Post Office Name: same as City
18	ALIAS_CITY	USPS - alternate names of city separated by ||
19	ALIAS_CITYN	Local - alternate names of city separated by ||
20	CITY2	Clean CITY name for geocoding
21	STATENAME2	Clean STATENAME for geocoding

FIPS = Federal Information Processing Standards. MSA = metropolitan statistical area. USPS = United States Postal Service.

The ZIP CLASS variable identifies if a ZIP Code is (1) military,[12] (2) post office box,[13] (3) standard, or (4) unique.[14] These distinctions provide insight into understanding the types of ZIP Codes contained in the Crosswalk Files that help to understand why some ZIP Codes do not match up to a ZIP Code boundary or location, such as the ZCTA or the boundary files from private companies. Linking the HUD ZIP Code to County Crosswalk File to the SAS ZIP Code file shows that 22 percent (9,201) ZIP Codes are Post office boxes, 73 percent (29,783) are standard, and 5 percent (2,090) are unique.

ZIP Codes are often used to represent place names and linking the HUD Crosswalk File to the SAS data can help identify those places. ZIP Codes can have multiple names associated with it that reflect successful mail deliver to and address with one of several cities, towns, or subdivisions (places) within its boundaries. These names are in the CITY, PONAME, ALIAS CITY, and ALIAS CITYN variables. However, a ZIP Code will always have exactly one place that is the primary (default) city, town, or subdivision that is captured in the CITY variable. These names can be used in identifying all the places associated with a ZIP Code, including using one of them as an alternative to assigning a ZIP Code to a census tract as discussed previously.

In a similar manner, if the SAS ZIP Code information is linked to the ZIP Code to County Crosswalk File, a summarization by county that includes a listing of the contained places can be conducted. Additional analyses can be done with socioeconomic data linked to the places. Additional geographic information is provided that indicates the telephone area codes covered, time-related factors, and the Cartesian X and Y coordinates that represent the geometric center of the ZIP Code area.

Evaluating A Crosswalk File

As described previously, HUD directly geocodes address data to the overlapping geographies that contain them, allowing for a specific cross-referencing between a ZIP Code and other geographies. As well, other companies do not provide information about how ZIP Codes are *crosswalked* to another geography. With the processed ZIP Code to County Crosswalk Files on the largest residential ratio, a difference comparison can be conducted between the county designations in the HUD Crosswalk Files and the SAS data. The results show that about 3 percent (909) of the ZIP Codes differ in county assignment. Exhibit 11 shows the ZIP Code areas and locations that mismatch between the two sources.

ZIP Codes overlapping multiple counties are highlighted with a black outline. The black and white points are ZIP Code locations that are either associated with another county or is a nondeliverable ZIP Code. The large ZIP Code in the center that cuts across four counties is identified in the HUD

[12] A military ZIP Code is a single entity code that represents an armed forces base or other property in which all mail is sent to a central location and distributed to the addresses within that property.

[13] A post office box ZIP Code represents multiple locations for same address, that is, delivery to same organization in multiple properties or building.

[14] Unique ZIP Codes represent a single high-volume address, for example, large organizations (government, academic, company, nonprofit). Mail delivery is to the organization that is in one large building across multiple properties or buildings or both. Some large organizations have multiple receipt centers to make mail management more efficient by delivering and collecting from multiple stops across the organization. At other times, the post office box ZIP Code is at a post office that is in another county and that has a different ZIP Code. Further, branch offices for large organization can be physically located in another county outside the ZIP Code to which it belongs.

Exhibit 11

County Assignment Mismatches With HUD ZIP Code Crosswalk Files

Crosswalk Files as belonging to Montgomery County, Maryland, which is the county with the smallest overlap, instead of Fredrick County, Maryland. However, about 47 percent of the residences are in that very small area, making it the county with which the ZIP Code becomes associated when using the highest residential ratio in the HUD Crosswalk Files. The SAS data, however, show that the ZIP Code belongs to Frederick County, which is the county that has the west section of the ZIP Code area. According to the HUD file, that part of the ZIP Code (in Frederick County) has only about 41 percent of the residential addresses, with Carroll and Howard Counties having about 11 and 1 percent, respectively.

Similarly in exhibit 11, ZIP Code locations (white points) in Montgomery County and in Prince George's County, Maryland are both federal government agencies and are associated with a Washington, D.C. ZIP Code. The other locations in the exhibit are locations that were simply in another county according to the HUD file by comparison to designation in the SAS data.

Data Limitations of the HUD USPS ZIP Crosswalk Files

Four limitations of the HUD Crosswalk data are of note. First, the Crosswalk Files do not contain the ZIP Codes for armed forces (military) and for the U.S. territories of American Samoa and the North Mariana Islands (Saipan). However, the territories of Guam, Puerto Rico, and the U.S. Virgin Islands are included.

Second, aside from the missing ZIP Codes noted previously, a small number—less than 1 percent—of addresses HUD receives from USPS cannot be geocoded due to errors in the addresses or the base maps. This means that some five-digit ZIP codes may not be in the Crosswalk Files.

Third, several other ZIP Codes across the United States are unavailable in the HUD Crosswalk Files. In some instances, this means that an undeliverable ZIP Code will be active but does not show up on the USPS list of ZIP Codes. A ZIP Code may be active, but it will not be found in the HUD Crosswalk Files, because that ZIP Code has been associated with its parent instead. In exhibit 12, more than 1,629 ZIP Code locations from the SAS dataset are shown that are not in the HUD Crosswalk Files.

Not all the unmatched ZIP Codes are post office boxes. Of these nonmatches, 48 (3 percent) are standard, 889 (55 percent) are post office boxes, and 692 (42 percent) are unique. About 54 percent of the nonmatches are for large private, government, or academic organizations with their own ZIP Codes that have mail delivered or collected from a central location within the organization. The remaining 46 percent have no name associated with the code and likely do not have mail directly delivered to that location, but from another mail-receiving ZIP Code.

Finally, HUD has not yet tabulated a crosswalk for CBSAs that solely use the CBSA code for every metropolitan area. This limitation means that in the ZIP Code to CBSA Crosswalk File, the metropolitan division code is used in place of the actual CBSA code, which does not always correspond with the Census Bureau-assigned CBSA code. To remedy, a separate file matching the Census

Exhibit 12

ZIP Codes Not in HUD ZIP Code Crosswalk Files

Bureau CBSA delineation files and the CBSA Crosswalk Files must be made on the metropolitan division codes to obtain the correct CBSA codes. However, at the time of this article, HUD is correcting the previously discussed limitation. By the fourth quarter of fiscal year 2017, HUD will split the ZIP Code-CBSA Crosswalk File into two products. One file will contain the CBSA codes, and the second file will contain metropolitan division codes. Users who need the CBSA Crosswalk Files prior to the fourth quarter of 2017 must use the Census Bureau CBSA delineation files to identify the CBSAs. After the fourth quarter of 2017, users can select the appropriate Crosswalk Files for their needs.

Summary

The HUD Crosswalk Files are one valuable piece of the geographic context puzzle when making use of ZIP Codes. The files contain no geographic information other than ZIP Codes and the corresponding identifications of the geographies that they cross-reference. This relationship enables the files to easily expand an address dataset to another geography that can connect to a wealth of other geographic information and socioeconomic data. Because the Crosswalk Files contain the ratios of address types, the most important aspect of the data is that users are empowered in making their own decisions about assigning ZIP Codes to other geographies.

Authors

Ron Wilson is an adjunct faculty member of the Geographic Information Systems Program at the University of Maryland, Baltimore County.

Alexander Din is a housing research and GIS analyst with the Maryland Department of Housing and Community Development.

References

Beyer, Kirsten M.M., Alan F. Schultz, and Gerard Rushton. 2007. "Using ZIP® Codes as Geocodes in Cancer Research." In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, edited by Gerard Rushton, Marc Armstrong, Josephine Gittler, Barry Greene, Claire Pavlik, Michele West, and Dale Zimmerman. New York: CRC Press: 37–67.

Cudnik, Michael T., Jing Yao, Dana Zive, Craig Newgard, and Alan T. Murray. 2012. "Surrogate Markers of Transport Distance for Out-of-Hospital Cardiac Arrest Patients," *Prehospital Emergency Care* 16 (2): 266–272.

Dai, Dajun. 2010. "Black Residential Segregation, Disparities in Spatial Access to Health Care Facilities, and Late-Stage Breast Cancer Diagnosis in Metropolitan Detroit," *Health & Place* 16 (5): 1038–1052.

Grubesic, Tony H., and Timothy C. Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data," *International Journal of Health Geographics* 5: 58.

Hipp, John R. 2007. "Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point," *American Sociological Review* 79: 659–680.

Krieger, N., P. Waterman, J.T. Chen, M.J. Soobader, S.V. Subramanian, and R. Carson. 2002. "ZIP Code Caveat: Bias due to Spatiotemporal Mismatches between ZIP Codes and US Census-defined Geographic Areas—The Public Health Disparities Geocoding Project," *American Journal of Public Health* 92 (7): 1100–1102.

Montalvo, Jose G., and Marta Reynal-Querol. 2017. Ethnic Diversity and Growth: Revisiting the Evidence. Working paper. Barcelona Graduate School of Economics.

Wilson, Ronald E. 2015. "The Neighborhood Context of Foreclosures and Crime," *Cartography and Geographic Information Science* 4 (2): 162–177.

———. 2013 "Changing Geographic Units and the Analytical Consequences: An Example of Simpson's Paradox," *Cityscape* 15 (2): 289–304.

———. 2011. "Visualizing Racial Segregation Differently—Exploring the Changing Patterns From the Effect of Underlying Geographic Distributions," *Cityscape* 13 (2): 163–174.