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.

Crosswalking ZIP Codes to Census Geographies: Geoprocessing the U.S. Department of Housing & Urban Development's ZIP Code Crosswalk Files

Alexander Din
U.S. Department of Housing and Urban Development

Ron Wilson
University of Maryland, Baltimore County

Disclaimer: The views expressed in this article are those of the author and do not represent the official positions or policies of the Office of Policy Development and Research, the U.S. Department of Housing and Urban Development, or the U.S. Government.

Allocating ZIP Code Data to Other Geographies

Although ZIP Codes are a commonly used geographic unit for mapping and spatial analysis, they frequently distort data (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). ZIP Codes are designed for efficient mail delivery, not for geographic analysis. The large area that ZIP Codes cover make them susceptible to data aggregation problems that corrupt local geographic patterns. Due to the irregular—and often contorted—shapes of ZIP Codes, smaller geographic boundaries are ignored when overlain with smaller geographies, to which population counts are disproportionally distributed between the multiple areas that are cross-cut. The U.S. Department of Housing and Urban Development (HUD) provides several crosswalk files to estimate incident counts at different geographic scales from data at the ZIP-Code level.

Our previous article offers a conceptual discussion of allocating ZIP Code data to other geographies (Wilson and Din, 2018). In this article, we explore the composition of HUD's crosswalk and demonstrate how to geoprocess the ZIP Code to census county and tract crosswalk files. Finally, we compare the estimates from the crosswalk process with the actual data to gauge the level of accuracy as an indicator of reliability.

Data

Rodents remain a widespread public health nuisance around world. They are a signal of poor environmental conditions that threaten the physical health of residents (Davis and Fales, 1949; Gardner-Santana et al., 2009). Rats pose serious health risks because they can rapidly transmit infectious diseases in dense neighborhoods. They are a substantial problem in various New York City neighborhoods and are a particular threat to restaurants and, most importantly, local supplies of pizza (Crowley, 2018).

Community associations, businesses, and city governments attempt to eradicate or manage rodent populations with a variety of techniques, sometimes collaborating with each other through a wide range of rat control programs (Biehler, 2009; Glass et al., 2009). For these groups to effectively combat rodents, they first need to identify the geographic range of the problem (Gardner-Santana et al., 2009; Getis and Boots, 1971; Lambert, et al., 2008); any rodent management program has an inherent geographic scope of their home-range. The range and density of rodent populations is difficult to measure due to the complexities of directly tracking mobile creatures[1] The use of ZIP Code-level data is not ideal for measuring the most affected neighborhoods due to the size and limited spatial range of habitats of rodents. Without more detailed geographic analysis, it is difficult for officials to effectively target mitigation efforts.

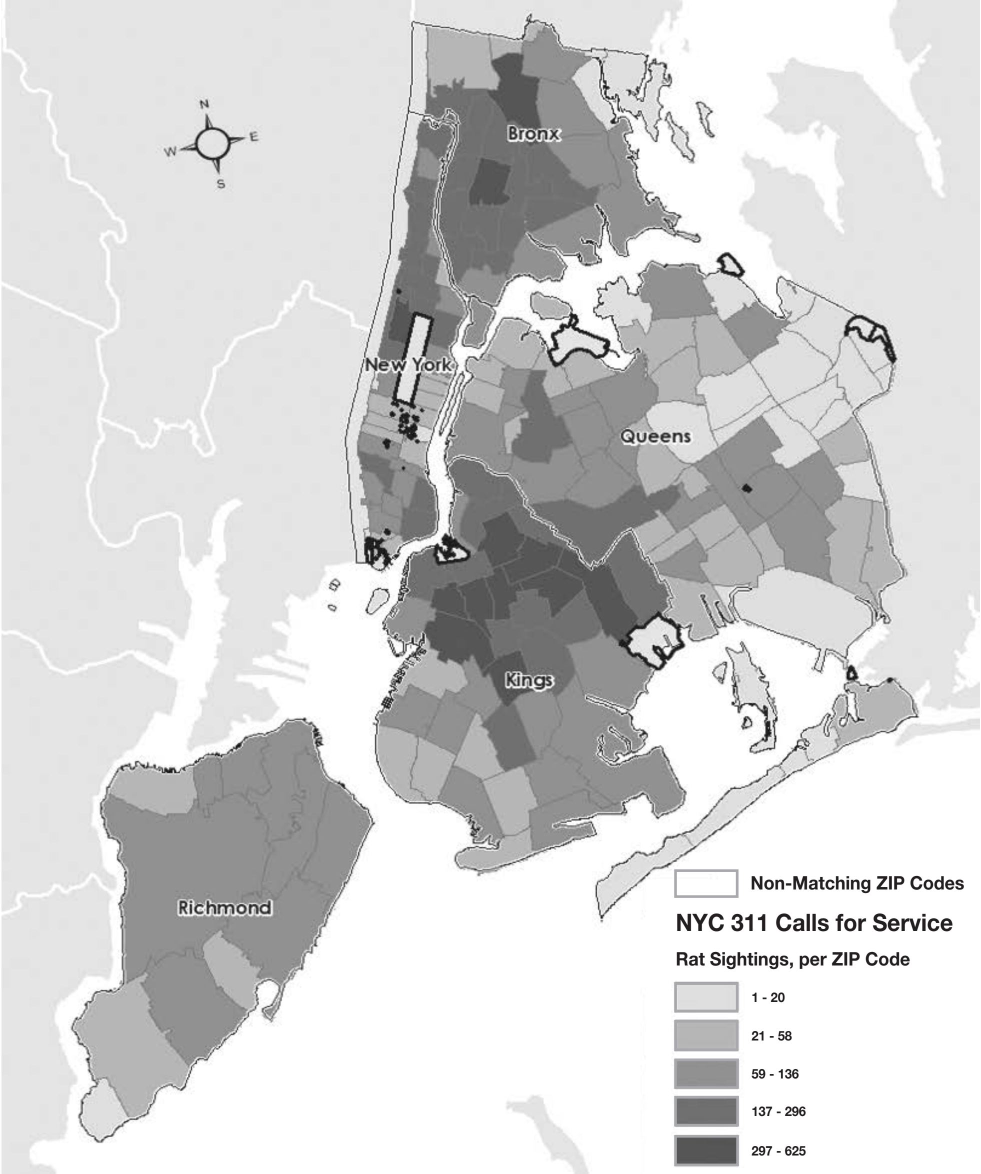
Citizen-produced 311 calls are useful for identifying the geographic patterns of many things occurring in everyday life, such as potholes to be fixed, vandalism, and public nuisances including rat sightings. These microdata are known as Volunteered Geographic Information (VGI) and provide a view of the events in everyday life that are difficult to impossible to capture otherwise (Elwood, 2008; Elwood, Goodchild, and Sui, 2012). Increasingly available 311 call data are a source of information that uses citizens as sensors for the locations of certain events. We exploit New York City (NYC) Open Data's 311 data[2] for the 2017 calendar year as a proxy to identify the geographic variation of rodents across New York City neighborhoods to determine more local patterns.

[1] Typically, rodents are tracked with radio frequency or other geographic positioning systems that require the use of delicate, complex, and expensive technology. Further, personnel and equipment are needed to capture the rodents, to which laboratories are needed to outfit the rodents with the tracking equipment and then analyze various biometric signatures to find from which colony each specimen is. Finally, and particularly in urban environments, personnel are needed on a regular basis to go and track the locations of each rodent because the dense urban cover blocks the signals when attempting to measure them from locations other than near the rodents' home-ranges.

[2] https://data.cityofnewyork.us/Social-Services/Rat-Sightings/3q43-55fe

Exhibit 1

Rat Sighting Call Frequencies by ZIP Code, 2017



NYC = New York City

The data from 311 Calls in New York City have multiple geographic identifiers that make it possible to compare the results of allocating crosswalked data to the actual counts of sightings by different geographies. We use the latitude and longitude coordinates to aggregate the sighting locations to the census tracts so that we can compare the estimated incidents with the actual count of incidents. This comparison allows us to evaluate the accuracy of county and tract estimates produced by the cross-walk process.

The 311 rat sightings locations were aggregated by ZIP Code, with 184 (70 percent) of the 263 ZIP Codes having at least one reported sighting—see exhibit 1.[3] The 67 ZIP Codes with zero sightings reported—highlighted by a thick black outline—are either (1) parks, (2) industrial sites, or (3) single buildings or blocks, most of which are concentrated in Midtown and Lower Manhattan. The map illustrates that rat sightings are primarily concentrated in the Upper East Side of Manhattan, central and north Bronx, and north Brooklyn. The ZIP Code layer with the summary counts of rat sightings will be used to later verify the crosswalk results.[4]

The detailed 311 New York City rat sighting data are particularly useful for (1) testing the accuracy of the HUD crosswalk files, especially in smaller geographies, and (2) demonstrating how resulting estimates can identify highly infested areas for effective extermination. During 2017, 19,152 rat sightings were reported to 311. Records that did not have a value for the Incident ZIP Code, Latitude, or Longitude fields were removed, reducing the data to 18,990 records, a reduction of 0.8 percent. The New York City 311 data were complete.

General Approach to Crosswalking the ZIP Code Files

Linking ZIP Code data to any of the ZIP Code crosswalk data sets is not a one-to-one assignment with any of the available geographies HUD offers.[5] Rather, a many-to-many approach must be used to ensure each record's proportion in a crosswalk file is associated with the corresponding geography each ZIP Code is contained within or with which it overlaps. The larger geographies—for example, counties or Core-Based Statistical Areas (CBSAs)—will contain most ZIP Codes within their jurisdictional boundaries, with a small number cross-cutting into neighboring jurisdictions. In this instance, each jurisdiction will at least be associated with one crosswalk record.

As an example of what happens when there is a one-to-one match using the census tract file, simply linking a ZIP Code record to one of the several tracts it overlaps will leave many of the corresponding tracts without any data.[6] This occurs because tracts are typically smaller than ZIP Code areas. Exhibit 2 shows the census tracts in the five NYC counties after a one-to-one match with the ZIP Code crosswalk records. The dark gray areas are the tracts that were assigned to the ZIP Code with the highest residential ratio. The light gray areas are tracts that were not associated with a ZIP Code record, but nevertheless do overlap with a ZIP Code because they are within or cross-cut at least one other ZIP Code record. These geographic "holes" occur because the ZIP Codes in those areas are larger than all the corresponding tracts within, to which the ZIP Code only gets assigned to one tract. Consequently, when a ZIP Code is assigned to one tract, all the adjacent tracts associated with that same ZIP Code are not assigned any of the address ratios used for analysis.

[3] After the join, the frequency field sum of 311 calls were checked to ensure it matched the number of records in the filtered 2017 data.

4 The crosswalk process can be done entirely outside of the Environmental System Research Institute (ESRI) software suite or any geographic information system (GIS). While the crosswalk is a geo-process, the data processing is performed within the attribute table as opposed to checking the geometries of the involved geographies. A GIS is only needed for visualization purposes.

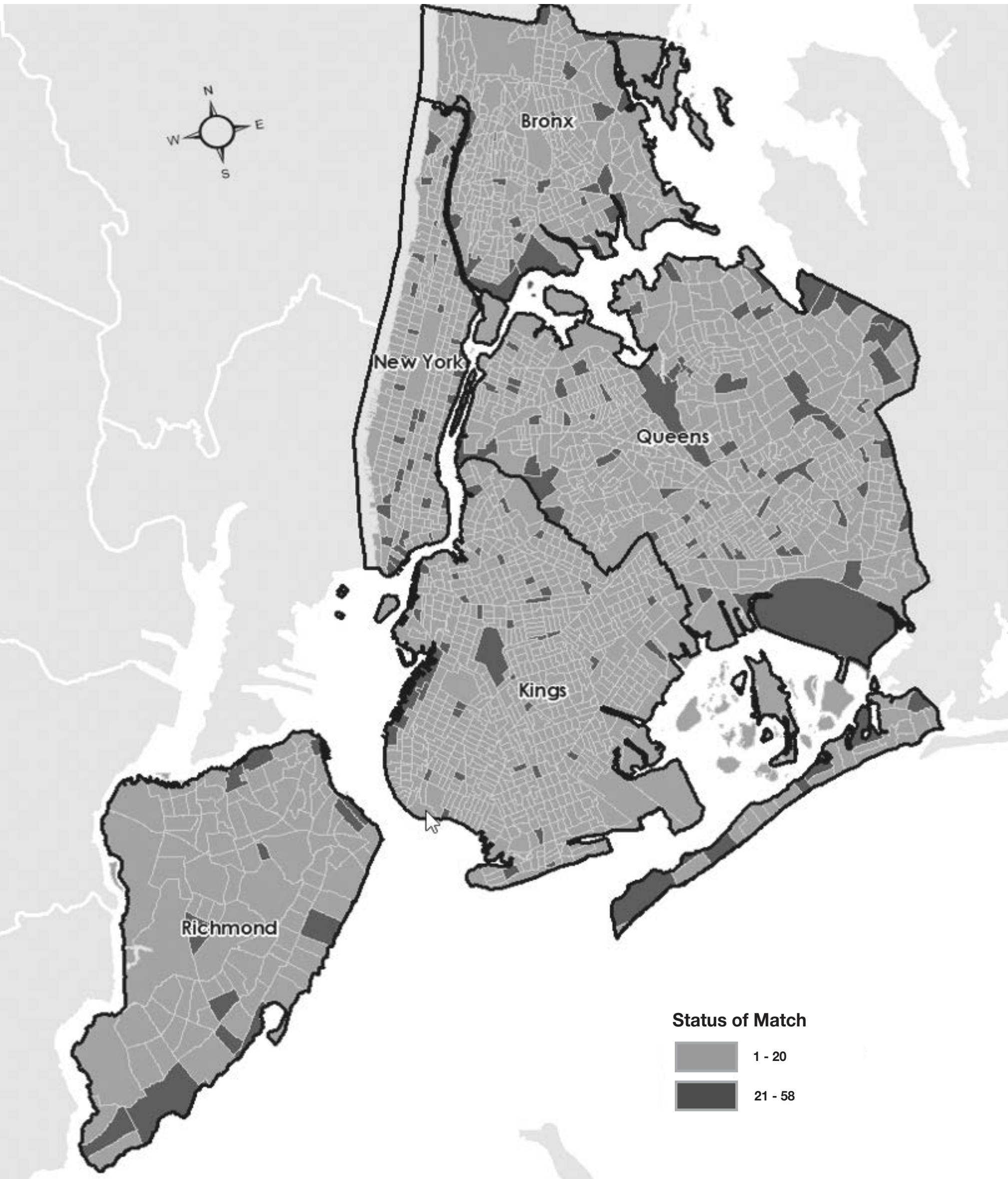
[5] HUD officers provide ZIP Code crosswalks for Census (1) tracts, (2) counties, (3) county subdivisions, (4) Core-based Statistical Areas, (5) CBSA divisions, and (6) Congressional District geographies.

[6] In dense urban areas, ZIP Codes can be associated with a large number of census tracts. For example, Brooklyn ZIP Codes 11234 and 11236 crosswalk (intersect) with 46 census tracts each.

Unless the analytical objective is to identify tracts with a particular residential ratio, a proportion assignment solution with a many-to-many match should be used to assign ZIP Code ratios to all overlapping tracts. With a one-to-many assignment, the data can be summed so that a total proportion of addresses can be calculated and used to estimate the likely number of events that occurred in each geography. The crosswalk data can then be linked to other data, such as the U.S. Census Bureau's place name geography, then easily mapped, tabulated, or spatially analyzed.

Exhibit 2

ZIP Code Assignment of Highest Residential Ratio



Crosswalking Data from ZIP Codes to County Geographies

Crosswalking ZIP Codes is a simple operation that involves (1) filtering out records for a specific geography, (2) adding additional geographic information to contextualize the data, (3) joining all relevant data sets to the crosswalk file, (4) summing the proportions for a total proportion of addresses, and (5) estimating incident counts per geography.[7] This process is the same for any of the HUD crosswalk geographies, whether census tracts, counties, Congressional Districts, or other geographies.

Filtering

The ZIP Code to county crosswalk file was first filtered to include only records in the state of New York, reducing the crosswalk file from 52,901 to 2,478 records. These records were then matched to ZIP Codes for all of New York State to ensure that they overlapped with the five boroughs in NYC. The 311 ZIP Code data were then joined to the crosswalk data using the ZIP Code field as the linking ID field. Once joined, a new field is calculated that is the product of the ZIP Code frequency and the tot_ratio field to produce an estimate for 311 rat calls for service.

Because ZIP Codes cross-cut multiple counties, duplicate ZIP Code records appear in the crosswalk file, indicating multiple overlaps. Most ZIP Code addresses are contained within a single county, however, there are numerous instances in which addresses with the same ZIP Codes are in several counties; the five boroughs of New York City are no exception. The ZIP Code field contains the ZIP Code, and each ZIP Code corresponds with a county Federal Information Processing Standard (FIPS) Code in the County field. The res_ratio, bus_ratio, oth_ratio, and tot_ratio fields[8] are really proportions of each address type in a ZIP Code in a county—see exhibit 3 for structure of the crosswalk file.

The proportions can be understood vertically or horizontally. For a vertical example, ZIP Code 10463 in exhibit 3 is split between Manhattan (36061) and the Bronx (36005). Among residential addresses (res_ratio), 89.7 percent of addresses in ZIP Code 10463 are located in Manhattan and 10.3 percent are located in the Bronx, adding up to 1.0 (100 percent) of addresses. The proportions in the remaining ratio fields (bus_ratio, oth_ratio, and tot_ratio) are interpreted the same way. Any ZIP Code that shows complete address proportions of 1.0 across all four address type ratio fields will have all addresses in a single county.

For a horizontal example, the Bronx (36005) contains 89.7 percent of residential, 92.0 percent of business and 90.0 percent of other addresses from the 10463 ZIP Code, with a total of 89.8 percent of the ZIP Code in this borough. For any ZIP Code that is split between larger geographies, the proportions in the ratio fields can vary between which ZIP Code part has the highest proportion. This inconsistency does not occur with ZIP Code 10463, or for any ZIP Code in the five boroughs of NYC. Note, there are instances where one ZIP Code in a larger geography will have the more

[7] The crosswalk process can be done entirely outside of the ArcGIS or any geographic information system because the data processing is performed with the attribute table by geographic name, as opposed to spatially overlaying the sighting locations with the ZIP Code boundaries. That is, no GIS is needed and can be done in any statistical, database, or spreadsheet software program.

[8] Residential, Business, Other, and Total, respectively.

residential addresses, while another ZIP Code will have more business and other addresses.

There are few instances among the five NYC boroughs where one address type is completely in one ZIP Code, but other address type proportions are split between two counties. For example, ZIP Code 11208 is in Brooklyn (36047) and Queens (36081), with residential and other addresses split between both counties but business addresses are completely in Brooklyn.

Exhibit 3

Structure of the HUD ZIP Code to County Crosswalk File

ZIP	County	Res_Ratio	Bus_Ratio	Oth_Ratio	Tot_Ratio
10458	36005	1.000000000	1.000000000	1.000000000	1.000000000
10459	36005	1.000000000	1.000000000	1.000000000	1.000000000
10460	36005	1.000000000	1.000000000	1.000000000	1.000000000
10461	36005	1.000000000	1.000000000	1.000000000	1.000000000
10462	36005	1.000000000	1.000000000	1.000000000	1.000000000
10463	36061	0.103139795	0.079906905	0.096541787	0.102085916
10463	36005	0.896860205	0.920093095	0.903458213	0.897914084
10464	36005	1.000000000	1.000000000	1.000000000	1.000000000
10465	36005	1.000000000	1.000000000	1.000000000	1.000000000
10466	36005	0.998516521	1.000000000	0.998856489	0.998607741

In this analysis, the total proportion of addresses (tot_ratio) in a ZIP Code is assigned to a county to estimate the number of 311 calls reporting a rat sighting in a census tract. The total proportion is used because rat sightings can be associated with any type of address. If the data were for home inspections, the residential proportion (res_ratio) would be more appropriate to use. These proportions represent all the individual addresses in a ZIP Code, either for a single detached unit or a building with multiple units. Because data from the USPS master address file are used for matching and standardizing addresses, all individual addresses are recorded and not just the address of the building with any number of individual addresses.

The crosswalk file should always be used, and thought of, as the "left" file in joining data. This is because ZIP Codes intersect multiple geographies for which data will be estimated, and the estimates will then be summarized by geographic place to produce the estimates.

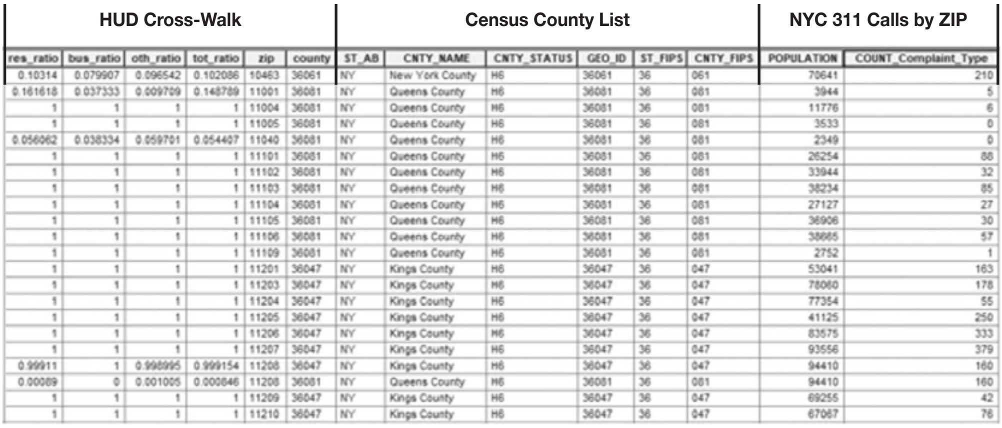
Geographically Contextualizing the HUD Crosswalk File

At this stage, using the cross-walk files to identify ZIP Codes in certain counties or states can be cumbersome because there are no place names to easily identify the places the data represent, just identifier codes that people do not easily recognize. If an entire state or subset of counties is selected for crosswalking, those records can be identified and selected using a query function to extract only those records that correspond with the study area. In our analysis, we selected those records that are within one of the five counties (boroughs) of New York City, which is more helpful when the names of the counties are known instead of the FIPS Codes when selecting those records.

To make the HUD Crosswalk file more useful, other data can be linked to the U.S. Census Bureau's County List file by the corresponding FIPS Code. Adding this data geographically "contextualizes" HUD's crosswalk file by adding the (1) individual state and county FIPS Codes, (2) county name, (3) state abbreviation, and (4) county type to the ZIP Codes and address proportions (exhibit 4). Adding this information makes the HUD crosswalk file more intuitive when geoprocessing and analyzing data, particularly by offering a method to summarize data using recognizable geographic place names.

Exhibit 4

Geographically Contextualized Crosswalk File



HUD Cross-Walk						Census County List						NYC 311 Calls by ZIP	
res_ratio	bus_ratio	oth_ratio	tot_ratio	zip	county	ST_AB	CNTY_NAME	CNTY_STATUS	GEO_ID	ST_FIPS	CNTY_FIPS	POPULATION	COUNT_Complaint_Type
0.10314	0.079907	0.096542	0.102086	10463	36061	NY	New York County	H6	36061	36	061	70641	210
0.161618	0.037333	0.009709	0.148789	11001	36081	NY	Queens County	H6	36081	36	081	3944	5
1	1	1	1	11004	36081	NY	Queens County	H6	36081	36	081	11776	6
1	1	1	1	11005	36081	NY	Queens County	H6	36081	36	081	3533	0
0.056062	0.038334	0.059701	0.054407	11040	36081	NY	Queens County	H6	36081	36	081	2349	0
1	1	1	1	11101	36081	NY	Queens County	H6	36081	36	081	26254	88
1	1	1	1	11102	36081	NY	Queens County	H6	36081	36	081	33944	32
1	1	1	1	11103	36081	NY	Queens County	H6	36081	36	081	38234	85
1	1	1	1	11104	36081	NY	Queens County	H6	36081	36	081	27127	27
1	1	1	1	11105	36081	NY	Queens County	H6	36081	36	081	36906	30
1	1	1	1	11106	36081	NY	Queens County	H6	36081	36	081	38665	57
1	1	1	1	11109	36081	NY	Queens County	H6	36081	36	081	2752	1
1	1	1	1	11201	36047	NY	Kings County	H6	36047	36	047	53041	163
1	1	1	1	11203	36047	NY	Kings County	H6	36047	36	047	78060	178
1	1	1	1	11204	36047	NY	Kings County	H6	36047	36	047	77354	55
1	1	1	1	11205	36047	NY	Kings County	H6	36047	36	047	41125	250
1	1	1	1	11206	36047	NY	Kings County	H6	36047	36	047	83575	333
1	1	1	1	11207	36047	NY	Kings County	H6	36047	36	047	93556	379
0.99911	1	0.998995	0.999154	11208	36047	NY	Kings County	H6	36047	36	047	94410	160
0.00089	0	0.001005	0.000846	11208	36081	NY	Queens County	H6	36081	36	081	94410	160
1	1	1	1	11209	36047	NY	Kings County	H6	36047	36	047	69255	42
1	1	1	1	11210	36047	NY	Kings County	H6	36047	36	047	67067	76

HUD = U.S. Department of Housing and Urban Development. NYC = New York City.

After contextualizing the HUD Cross-walk file, the rat sighting 311 calls are linked together to complete the data set for analysis.

Estimating County Rat Calls for Service per Geography

Once the additional data are linked to the HUD ZIP to County crosswalk file, the number of 311 calls reporting rat sightings can be estimated for each ZIP Code. A new field needs to be created for this estimate, which will be the total address ratio (tot_ratio) multiplied by the number of calls within each ZIP Code (COUNT_Complaint_Type). This estimate will represent the number of calls in each county based on the addresses within each jurisdiction.

Since the crosswalk file has been matched to country level data, some of the estimates will sum to exactly the total number of rat sightings, but estimates will not add to the exact total. This occurs when all the addresses in a ZIP Code are in only one county. For example, Richmond (Staten Island) in NYC. In other counties, such as Kings (Brooklyn), New York (Manhattan), Queens, and the Bronx, estimates will sum to a value less than the total of reported rat sightings in a ZIP Code because some of the counts are in neighboring counties outside one of the five boroughs. Since ZIP Codes do not respect political boundaries, it is possible for counties outside of NYC to have estimated counts. The edge of a project area should always be considered because ZIP Codes can cross into neighboring administrative regions.

The remaining records from the HUD ZIP Code crosswalk file are summarized by county to create a set of descriptive statistics for rat sighting calls for each county. Exhibit 5 shows the results from summing the reported sightings (SUM_Complaint_Type) by ZIP Code and using the previously described estimation method (SUM_EST_Complaint_Type).

Exhibit 5

Summary Statistics of County-Level ZIP Code Counts and Estimates

Cnty_Name	Cnty_ Count	SUM_ Complaint_ Type	SUM_EST_ Complaint_Type	MEAN_EST_ Complaint_Type	STD_EST_ Complaint_Type
Bronx County	25	3735	3713.045	148.52182	91.78859
Kings County	43	7359	7204.375	167.54360	153.98840
New York County	95	4640	4451.438	46.85724	78.34508
Queens County	82	2998	2833.369	34.55328	44.50728
Richmond County	14	783	783.000	55.92857	37.31254

The estimates are very similar to the actual counts because the geographic unit used to estimate the number of rat sightings is at an aggregate to a larger level geography. The CNTY_COUNT field shows the number of ZIP Codes that fall within each county boundary. The SUM_, MEAN_ and STD_EST_Complaint_Type fields provide information on the variation of reported rat sightings between counties. The difference between the county count and estimated sum is due to a small percentage of 311 reports coming from ZIP Code addresses outside the New York City boroughs and NYC's 311 service does not respond to requests made outside of the City.

These summary statistics communicate the spread of estimated calls across each county and provide additional geographic information that the counts alone cannot provide. Without these statistics, there is no information about the distribution of rat calls across each county, minimizing the necessary assumption that calls are evenly distributed across a county. For example, among the counties in exhibit 5, the estimate's standard deviation indicate a wide distribution of calls around the mean. This finding implies that counties contain ZIP Codes with both very many and very few reported sightings.

Crosswalking Data from ZIP Codes to Census Tracts

The crosswalking process to census tracts is the same as crosswalking to counties, but in this case, ZIP Codes are likely to intersect a greater number of geographies. Because census tracts intersect with multiple ZIP Codes, no single tract represents a single ZIP Code. As such, the address proportions need to be summed to provide a complete proportion of addresses for this level of analysis. The structure of the ZIP Code to census tract crosswalk file is the same as the HUD ZIP to county crosswalk file. Because the census tract crosswalk file details multiple tracts to multiple ZIP Codes, more complex relationships are revealed at this geographic level (exhibit 6).

Exhibit 6

Structure of the HUD ZIP Code to Census Tract Crosswalk File

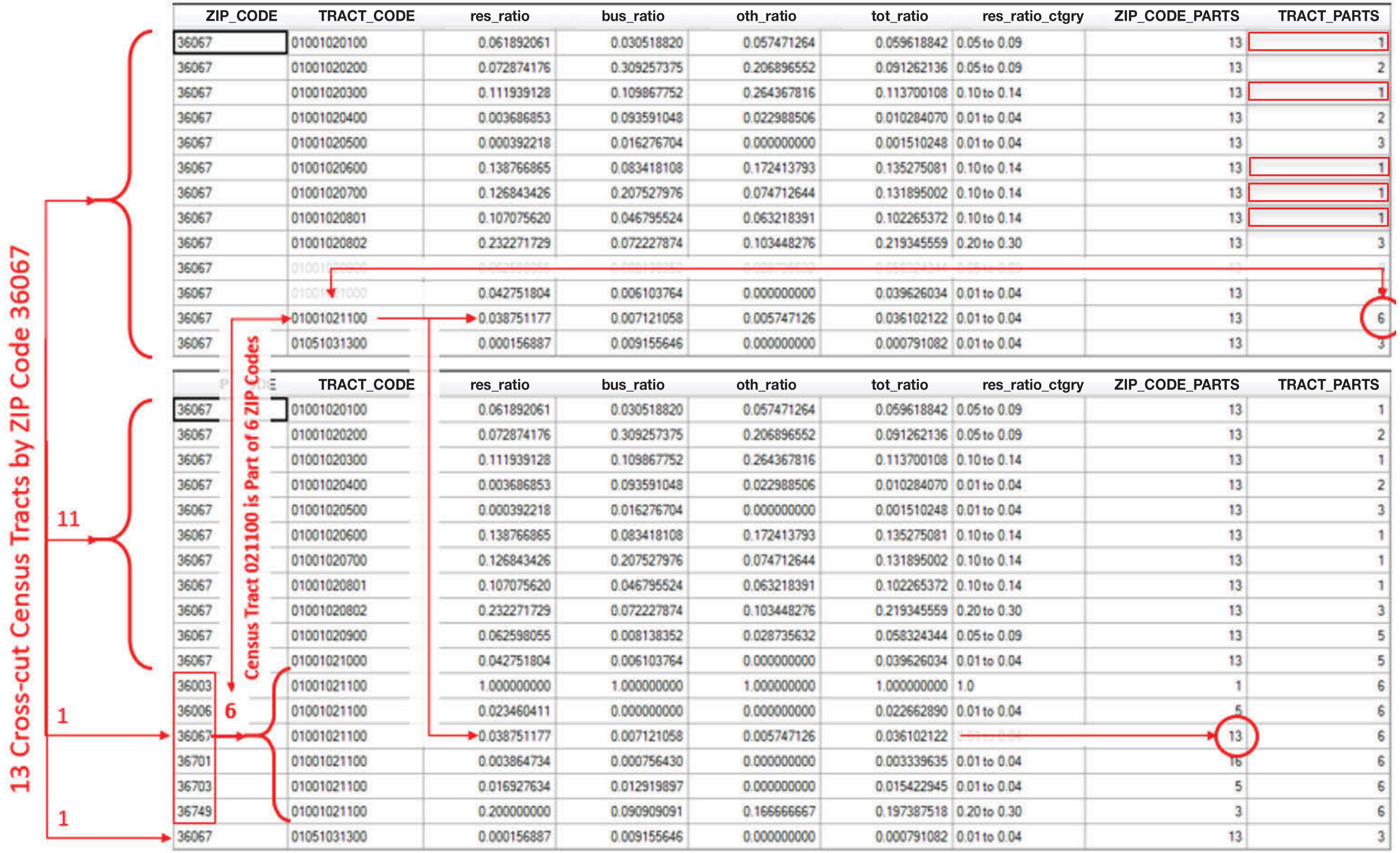
ZIP_CODE	TRACT_CODE	Res_Ratio	Bus_Ratio	Oth_Ratio	Tot_Ratio
11697	36081091601	0.987118644	0.850000000	1.000000000	0.985299031
11697	36081091602	0.012881356	0.100000000	0.000000000	0.014032743
11694	36081091800	0.001121076	0.000000000	0.000000000	0.001018330
11694	36081092200	0.093497758	0.002040816	0.024390244	0.086048880
11694	36081092800	0.129596413	0.024489796	0.039024390	0.120570265
11694	36081093401	0.152466368	0.130612245	0.165853659	0.151934827
11694	36081093402	0.189237668	0.175510204	0.314634146	0.193788187
11694	36081093800	0.329484305	0.573469388	0.321951220	0.341344196
11693	36081094201	0.076448157	0.091269841	0.168141593	0.080534670
11693	36081107201	0.174323588	0.194444444	0.039823009	0.170091896
11693	36081107202	0.000181587	0.011904762	0.000000000	0.000668338
11692	36081094202	0.031427832	0.088435374	0.000000000	0.031624460
11692	36081094203	0.072038282	0.013605442	0.018518519	0.069549104
11692	36081095400	0.204604242	0.360544218	0.166666667	0.206423718
11692	36081096400	0.319322297	0.319727891	0.092592593	0.313279802
11692	36081097202	0.141360579	0.061224490	0.314814815	0.144533663
11691	36081097203	0.069752432	0.011688312	0.056316591	0.067436629
11691	36081097204	0.013738417	0.037662338	0.019786910	0.014707155
11691	36081099200	0.058226914	0.038961039	0.057838661	0.057574185

Exhibit 6 illustrates the complex relationship between one ZIP Code and the census tracts it cross-cuts. The top table in exhibit 7 highlights ZIP Code 36067 and its corresponding census tract crosswalk records. The bottom table highlights the 13 census tracts associated with ZIP Code 36067, and the all other ZIP Codes associated with any of the tracts. Both tables have two additional fields, one showing the number of parts per ZIP Code (ZIP_CODE_PARTS) and the other with the total census tracts associated with each crosswalk record (TRACT_PARTS).

In the top table, ZIP Code 36067 is associated with a total of 13 census tracts, each of which are either completely contained within the ZIP Code, or cross-cut and are shared with neighboring ZIP Codes. The TRACT_PARTS field shows the number of times each of those tracts is cross-cut by any ZIP Code that it overlaps. Those records with 1 in the TRACT_PARTS field indicate that the tract is completely contained within ZIP Code 36067. The rectangles in the last column (TRACT_PARTS) are five tracts that are completely contained in ZIP Code 36076. The remaining tracts cross cut between two to six other ZIP Codes.

Exhibit 7

Linkages Between ZIP Codes and Cross-Cut Census Tracts



ZIP_CODE	TRACT_CODE	res_ratio	bus_ratio	oth_ratio	tot_ratio	res_ratio_ctgry	ZIP_CODE_PARTS	TRACT_PARTS
36067	01001020100	0.061892061	0.030518820	0.057471264	0.059618842	0.05 to 0.09	13	1
36067	01001020200	0.072874176	0.309257375	0.206896552	0.091262136	0.05 to 0.09	13	2
36067	01001020300	0.111939128	0.109867752	0.264367816	0.113700108	0.10 to 0.14	13	1
36067	01001020400	0.003686853	0.093591048	0.022988506	0.010284070	0.01 to 0.04	13	2
36067	01001020500	0.000392218	0.016276704	0.000000000	0.001510248	0.01 to 0.04	13	3
36067	01001020600	0.138766865	0.083418108	0.172413793	0.135275081	0.10 to 0.14	13	1
36067	01001020700	0.126843426	0.207527976	0.074712644	0.131895002	0.10 to 0.14	13	1
36067	01001020801	0.107075620	0.046795524	0.063218391	0.102265372	0.10 to 0.14	13	1
36067	01001020802	0.232271729	0.072227874	0.103448276	0.219345559	0.20 to 0.30	13	3
36067	01001020900	0.062598055	0.008138352	0.028735632	0.058324344	0.05 to 0.09	13	5
36067	01001021000	0.042751804	0.006103764	0.000000000	0.039626034	0.01 to 0.04	13	5
36067	01001021100	0.038751177	0.007121058	0.005747126	0.036102122	0.01 to 0.04	13	6
36067	01051031300	0.000156887	0.009155646	0.000000000	0.000791082	0.01 to 0.04	13	3

ZIP_CODE	TRACT_CODE	res_ratio	bus_ratio	oth_ratio	tot_ratio	res_ratio_ctgry	ZIP_CODE_PARTS	TRACT_PARTS
36067	01001020100	0.061892061	0.030518820	0.057471264	0.059618842	0.05 to 0.09	13	1
36067	01001020200	0.072874176	0.309257375	0.206896552	0.091262136	0.05 to 0.09	13	2
36067	01001020300	0.111939128	0.109867752	0.264367816	0.113700108	0.10 to 0.14	13	1
36067	01001020400	0.003686853	0.093591048	0.022988506	0.010284070	0.01 to 0.04	13	2
36067	01001020500	0.000392218	0.016276704	0.000000000	0.001510248	0.01 to 0.04	13	3
36067	01001020600	0.138766865	0.083418108	0.172413793	0.135275081	0.10 to 0.14	13	1
36067	01001020700	0.126843426	0.207527976	0.074712644	0.131895002	0.10 to 0.14	13	1
36067	01001020801	0.107075620	0.046795524	0.063218391	0.102265372	0.10 to 0.14	13	1
36067	01001020802	0.232271729	0.072227874	0.103448276	0.219345559	0.20 to 0.30	13	3
36067	01001020900	0.062598055	0.008138352	0.028735632	0.058324344	0.05 to 0.09	13	5
36067	01001021000	0.042751804	0.006103764	0.000000000	0.039626034	0.01 to 0.04	13	5
36003	01001021100	1.000000000	1.000000000	1.000000000	1.000000000	1.0	1	6
36006	01001021100	0.023460411	0.000000000	0.000000000	0.022662890	0.01 to 0.04	5	6
36067	01001021100	0.038751177	0.007121058	0.005747126	0.036102122	0.01 to 0.04	13	6
36701	01001021100	0.003864734	0.000756430	0.000000000	0.003339635	0.01 to 0.04	16	6
36703	01001021100	0.016927634	0.012919897	0.000000000	0.015422945	0.01 to 0.04	5	6
36749	01001021100	0.200000000	0.090909091	0.166666667	0.197387518	0.20 to 0.30	3	6
36067	01051031300	0.000156887	0.009155646	0.000000000	0.000791082	0.01 to 0.04	13	3

13 Cross-cut Census Tracts by ZIP Code 36067
11
1
1
Census Tract 021100 is Part of 6 ZIP Codes
6

The bracket at the top left identifies the corresponding 13 census tracts (TRACT_CODE) among the 18 listed in the bottom table. The other brackets, lines, and circles highlight the complete relationship between the ZIP Code 36067 and its corresponding tracts from one table to the other. The bottom table also show that there are five other ZIP Codes that cross-cut at least one of the same tracts cross-cut by ZIP Code 36067. The bottom table—like the top table—shows which tracts are completely contained in the ZIP Code with a value of 1 in the TRACT_PARTS field. Unlike the top table, the bottom table shows which tracts in ZIP Code 36067 are cross-cut by other ZIP Codes. For example, tract 010012021100 in ZIP Code 36067 is cross-cut by five other ZIP Codes (detailed in rows 12 to 17). As shown in both the top and bottom tables, one tract (01001021100) is cross-cut six times by one ZIP Code.

Preparing the ZIP Code to Census Tract Crosswalk File

Even though a ZIP Code intersects multiple census tracts, the records in the HUD Crosswalk file are summarized by tract ID. The geoprocessing remains the same regardless of geographic level. The total proportion of addresses in tract in each ZIP Code can be calculated by summing the proportions from all cross-cut addresses.

There are 18,990 rat sighting calls across 184 ZIP Codes in the five boroughs of NYC. Of the 3,191 tracts in the NYC area, 3,033 (95 percent) ZIP Codes match the HUD crosswalk file.

Census tracts that are outside the city limits and that have no reported sightings at the ZIP Code-level were removed from the analysis. Note that 60 of 153 ZIP Code tracts that were outside the city limits were from either Nassau or Westchester Counties. After removing these tracts, 3,131 remain and were matched with data from the HUD crosswalk file. Of the remaining crosswalk tracts that do not match to a ZIP Code, 66 are in New York County, 17 are in Queens, 13 are in Kings, and 2 are in Richmond, with 0 in the Bronx. After these 93 unmatched records were removed, a total of 3,033 crosswalk file tracts remained for matching with census tract data.

Of the 67 ZIP Codes that did not match in exhibit 7, 49 of the ZIP Codes from the crosswalk tracks also did not match and were removed.

In 2017, a total of 19,580 rat sightings were reported among ZIP Codes within one of five counties of NYC, with the estimates totaling to 18,841—739 less the number reported. Again, the lower estimate is due to a reduction in calls from tracts that do not correspond with an NYC ZIP Code. Of the 739 calls that were not included in the estimate, 151 were dropped due to ZIP Codes not matching to the ZIP Code to Tract HUD Crosswalk file, leaving 588.

Estimating Tract Rat Calls for Service

Exhibits 8 and 9 demonstrate how to calculate the census tract estimates using HUD's crosswalk files. Exhibit 8 shows the calculations for the 11415 ZIP Code, which serves a portion of southern Kings County (Brooklyn). In 2017, there were 21 total sightings reported in the 11415 ZIP Code. The frequency of sightings was then multiplied by the tot_ratio of each census tract. The result is a set of estimated rat sightings for each of the eight tracts that intersect with the 11415 ZIP Code. The estimates range from 3.50 to 0.38 sightings and total to 21, which is the ZIP Code frequency.

Exhibit 8

Rat-Call Estimates for ZIP Code 11415 by Census Tract

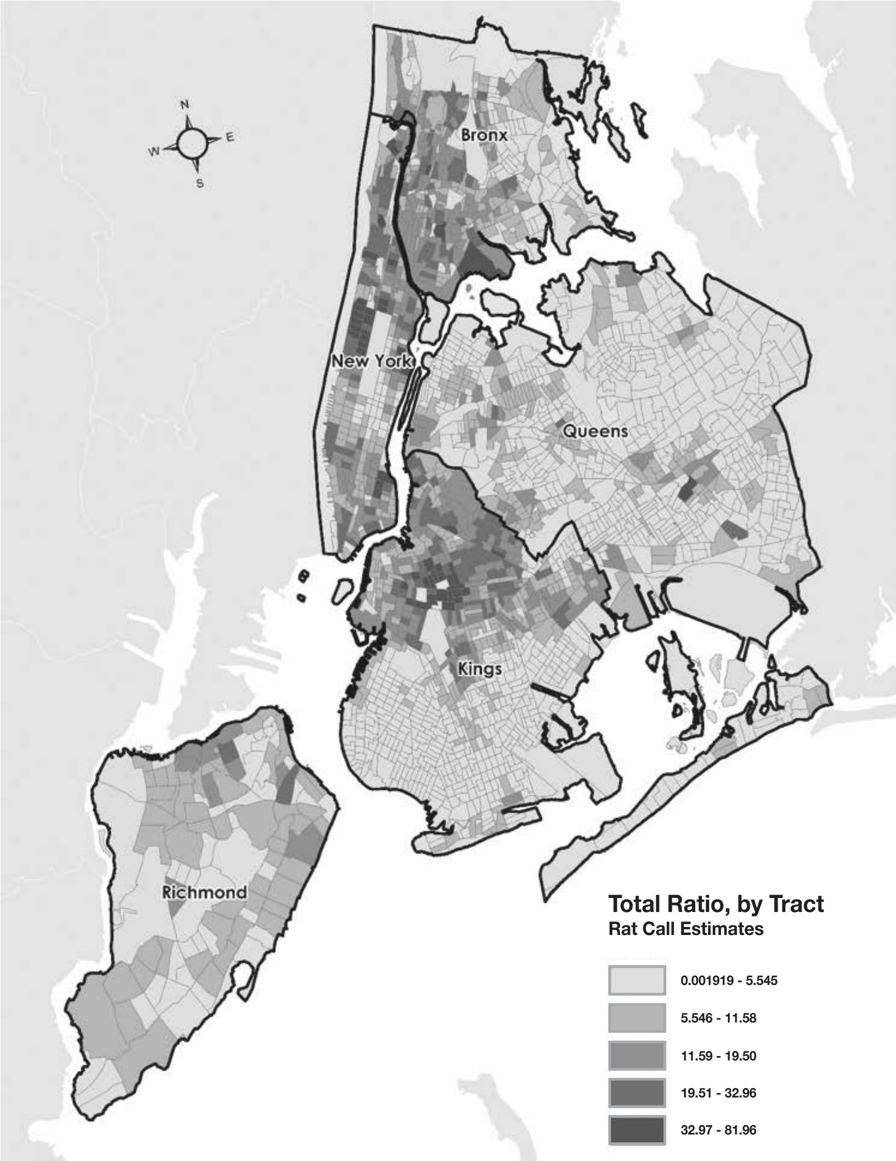
ZIP	TRACT	Tot_Ratio	City	Alias_City	County_Name	EST_Call_Cnt	ZIP_Call_Cnt
11415	36081021600	0.16773620	Kew Gardens	Jamaica	Queens	3.5224603	21
11415	36081014000	0.13861191	Kew Gardens	Jamaica	Queens	2.9108500	21
11415	36081013200	0.01806694	Kew Gardens	Jamaica	Queens	0.3794057	21
11415	36081077500	0.09951624	Kew Gardens	Jamaica	Queens	2.0898411	21
11415	36081013800	0.14986672	Kew Gardens	Jamaica	Queens	3.1472011	21
11415	36081013400	0.19616941	Kew Gardens	Jamaica	Queens	4.1195577	21
11415	36081013600	0.13012143	Kew Gardens	Jamaica	Queens	2.7325501	21
11415	36081077300	0.09991115	Kew Gardens	Jamaica	Queens	2.0981341	21

Because census tracts intersect with multiple ZIP Codes, the resulting table will have multiple records for the same census tract. The calculated estimates are summed by a census tract identifier using a summing function. Exhibit 8 displays the calculations for the 36081021600 census tract which intersects with two ZIP Codes. Exhibit 9 illustrates the mapped results for the estimated rat sightings at the census tract level for the 11415 ZIP Code. Note that in Exhibit 8, while the portion of total addresses within a census tract is likely proportional to the total area within a ZIP Code, the two may differ due to an uneven distribution of addresses.

The resulting estimated rat sightings at the census tract-level are shown in exhibit 9 and are mapped later. Note that the mapped results at the tract-level are comparable to those at the ZIP Code level. The crosswalk file assumes that the reports are evenly distributed across the ZIP Code, thus the results look similar at the tract and ZIP Code level.

Exhibit 9

Rat-Call Estimates, by Census Tract



Assessing the Crosswalk Estimations

With the New York City 311 data containing the coordinates of the call for service locations, we used the geocoded locations of the reported rat sighting to assess the accuracy of the ZIP Code to census tract crosswalk, apportioning to smaller geographies to create the estimates. We aggregated the locations to the counties and census tracts using a spatial join to directly compare the actual count of rat sighting counts with the estimates in both geographies. We summarized the counts and estimates by counties and tracts to compare the differences between the actual vs. the estimates.

Crosswalking data from ZIP Codes to counties yielded highly accurate results, which is because most ZIP Codes are contained with the county boundary with only a small amount of area falling in neighboring Nassau and Westchester counties. When actualized as geographic boundaries, ZIP Codes often follow closely to the larger jurisdiction's (for example, county or city) administrative boundary. With counties, then, it is expected that the differences between the actual number of calls and the estimates are not very different.

Exhibit 10

Actual Versus Estimated Rat Calls for Service, by County

Cnty_Name	Cnty_Count	SUM_ Complaint_ Type	SUM_EST_ Complaint_ Type	DIF_ Complaint_ Type	PCT_DIF_ Complaint_ Type
Bronx County	25	3735	3713.045	-21.95455	0.5878058
Kings County	43	7359	7204.375	-154.62516	2.1011708
New York County	95	4640	4451.438	-188.56196	4.0638353
Queens County	82	2998	2833.369	-164.63089	5.4913573
Richmond County	14	783	783.000	0.00000	0.0000000

Notes: CNTY_NAME is the name of the county, not borough. SUM_Complaint_Type describes the actual number of complaints, whereas SUM_EST_Complaint_Type describes the number of estimated complaints. DIF_Complaint_Type is the count difference between actual and estimated complaints, and PCT_DIF_Complaint_Type is the percentage difference between actual and estimated complaints.

The differences are shown in exhibit 10 for each of the counties. Richmond (Staten Island) and the Bronx show very little difference, with no difference in Richmond and about 22 in the Bronx. Kings (Brooklyn), New York (Manhattan), and Queens show a difference between about 154 to 188, no more than 6-percent difference.

For census tracts, however, the results are less accurate because ZIP Codes cross-cut numerous tracts because they do not follow the boundaries closely, and as many as 46 census tracts intersect with a single ZIP Code in NYC. Exhibit 11 shows multiple ZIP Codes cross-cutting the census tracts in Bronx and New York (Manhattan) Counties. ZIP Code 10457, as one example, in the center of exhibit 11 contains 17 tracts and cross-cuts 11 tracts shared with three other ZIP Codes. When the cross-cuts occur, a sizeable number of the tracts are often in multiple ZIP Codes. This means that there is a more even proportion of addresses in each tract, which corresponds with a more equal likelihood of a rat call for service coming from each part of the tract.

Exhibit 11

ZIP Codes Cross-Cutting Multiple Census Tracts

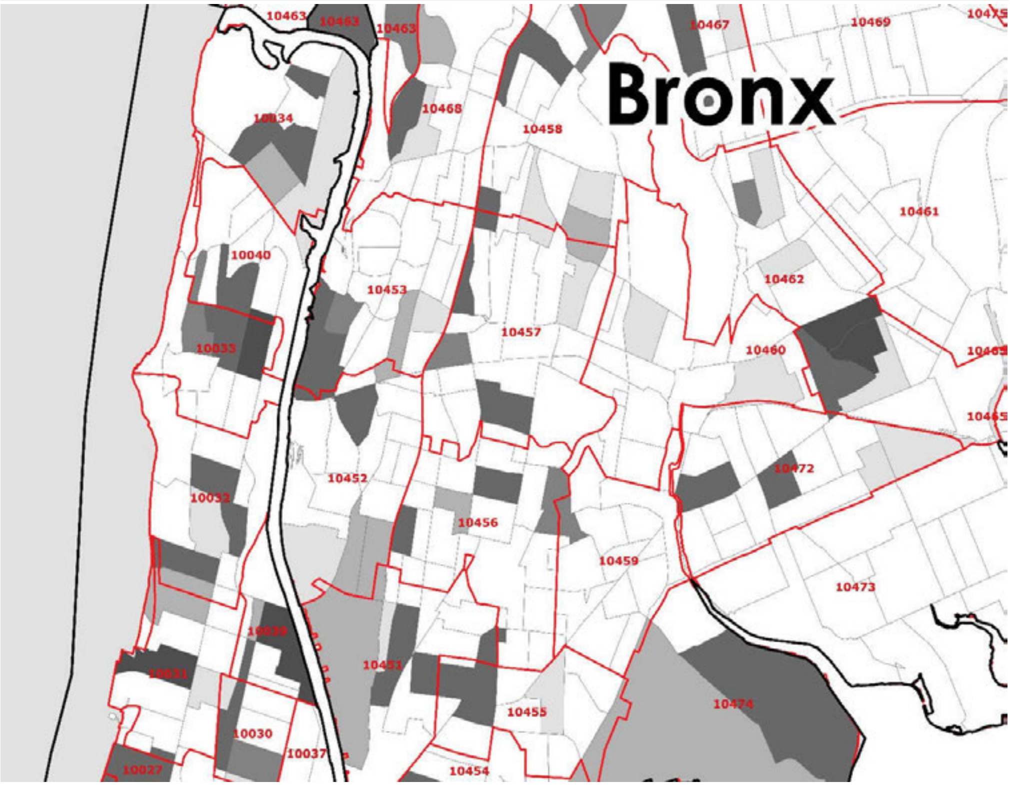


Exhibit 12 shows a frequency distribution of differences between the actual counties and estimates for each of the 2,164 tracts across the New York City area. The distribution shows to have a high kurtosis normal distribution, which means the differences for most tracts are not that wide.[9] In this aspect the histogram indicates that there is little bias in the estimate errors because the differences are distributed in similar way as to be expected under a random process.

The three middle bars at the average of the histogram make of up 40 percent (866) of all tracts, to which the difference is between the actual number of calls and estimated number of calls is -1.29 and 2.54. The largest group of census tracts is within approximately 2.5 estimated to actual calls, a relatively small difference. The remaining 28 percent of tracts within the first standard deviation extends the difference range from -8.18 to 8.54, that is, about eight calls below or above the actual number of rat calls for service.

Beyond the first standard deviation, the range extends from about 9 to 58 calls, indicating that some tracts are very different between the actual and estimated number.

[9] Having a high kurtosis is an ideal characteristic of a difference distribution because it indicates that most of the differences are not large and are similar to the average.

Exhibit 12

Frequency Distribution of Actual Versus Estimated Rat Calls for Service, Census Tracts

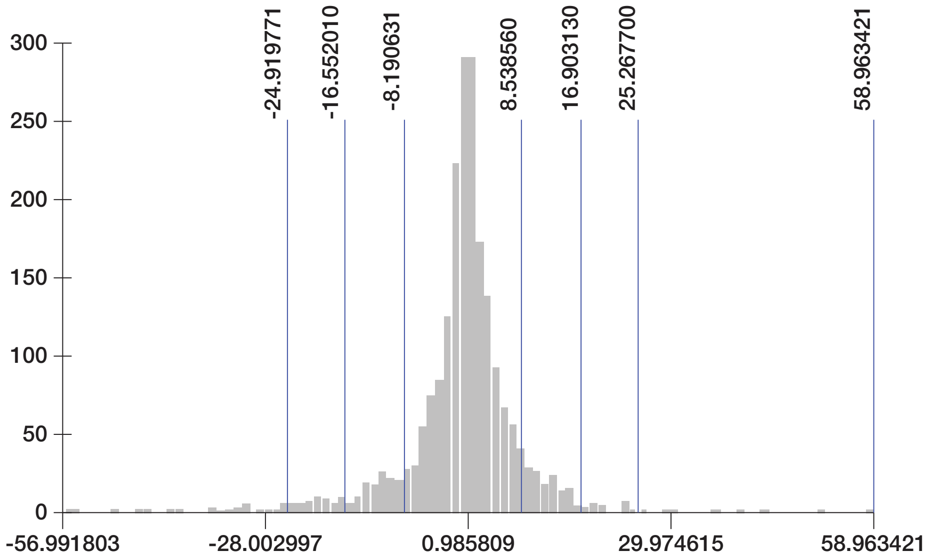
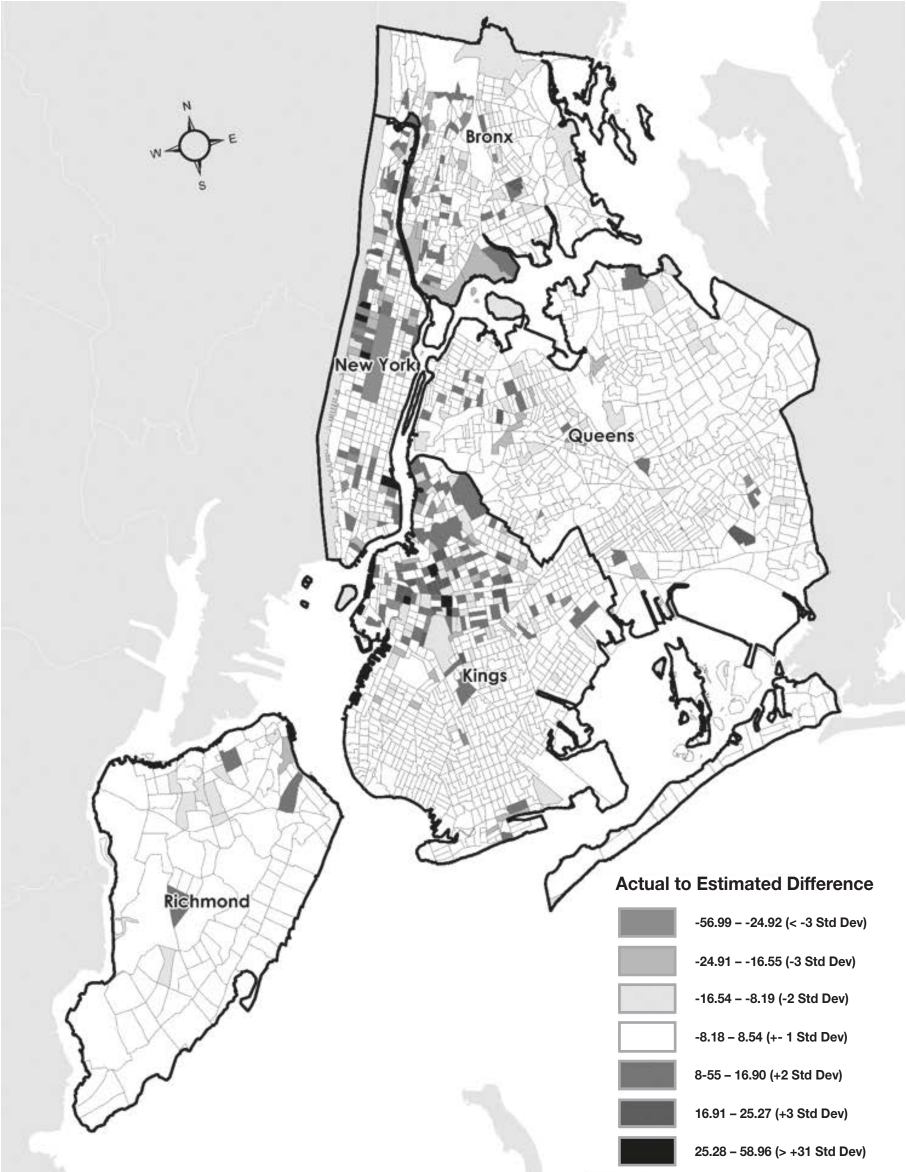


Exhibit 13 shows the geographic distribution of the difference across the NYC area. Unlike the histogram of actual counts vs estimate differences, the geographic distribution reveals a bias in the differences through the clustering of higher and lower differences. The clusters are largely in north Brooklyn and the Upper West Side in New York (Manhattan). This finding indicates that using addresses may not be the best for estimating calls for service in these areas, which for Brooklyn is quite sizeable. A scattered number of tracts with large differences can be seen in the Lower East in Manhattan, in east and south Bronx, and in west Queens.

It is unlikely that when crosswalking data that one will have the point data to check the accuracy of the estimate or there would not be a need to perform the crosswalk data processing. In identifying these discrepancies between actual and estimated counts for whatever the data, however, additional patterns of human activity may be identified. These patterns and trends may be further described through additional spatial analysis.

Exhibit 13

Geographic Distribution of Actual Versus Estimated Rat Calls for Service, Census Tracts


N
W
E
S
Bronx
New York
Queens
Kings
Richmond
Actual to Estimated Difference
-56.99 – -24.92 (< -3 Std Dev)
-24.91 – -16.55 (-3 Std Dev)
-16.54 – -8.19 (-2 Std Dev)
-8.18 – 8.54 (+- 1 Std Dev)
8-55 – 16.90 (+2 Std Dev)
16.91 – 25.27 (+3 Std Dev)
25.28 – 58.96 (> +31 Std Dev)

Summary

The HUD ZIP Code crosswalk files are one of the best available data sets for allocating address locations from ZIP Codes to other geographies or vice versa. Because the crosswalk files are updated quarterly, they are a timely data set that keeps pace with the frequent changes in ZIP Code areas. The crosswalk files are also available from the first quarter of 2010 which means the most appropriate file can be applied to the data if it occurred within the timeframe. These data sets allow analysts, policy-drafters, and decisionmakers—to name a few—to integrate data that is aggregated at the ZIP Code level with existing data sets such as census data or the American Community Survey (ACS) and make the data more useful for analysis. The availability of previous quarters also allows for a relatively granular temporal analysis for changes across time.

It should be noted that there are considerations when crosswalking data to smaller geographies, particularly census tracts, however, because data on the resulting estimates are allocated as proportionally as to how the addresses are located between the ZIP Codes and the recipient geography. When using the files, there are a number of data components of which one needs to be mindful. With the exception of Congressional Districts crosswalk file, the crosswalk files are produced using the 2010 Census geographies and do not correspond to geographies that changed between censuses.[10] This can cause confusion to those using the files. The crosswalk files are produced using ZIP Codes that have associated address counts; this does not include ZIP Codes that are only associated with postal office boxes. Still, even with these considerations, the crosswalk files are a popular and robust method to reapportion data from ZIP Codes to other administrative geographies, or vice versa, in the health, policy, finance, marketing, and other fields.

[10] https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html

Acknowledgments

We thank Danielle M. Wilson from American University for her help in editing this paper.

Author

Alexander Din is a social science analyst in the Office of Policy Development & Research at the U.S. Department of Housing and Urban Development.

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

References

Biehler, Dawn Day. 2009. "Permeable Homes: A historical political ecology of insects and pesticides in US public housing," *Geoforum* 40: 1014–1023.

Beyer, Kirsten, Alan 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 G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West, and D. L. Zimmerman. Boca Raton, FL: CRC PRESS: 37–64.

Crowley, Chris. 2018. "Pizza Rat Made Its Triumphant Return to New York This Weekend." *The New York Times*. April 30. http://www.grubstreet.com/2018/04/pizza-rat-makes-triumphant-return-to-nyc.html.

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.

Davis, DE., and WT. Fales. 1949. "The Distribution of Rats in Baltimore, Maryland," *American Journal of Hygiene* 49: 247-254.

Elwood, Sarah. 2008. "Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS," *Geojournal* 72: 173–183.

Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. 2012. "Researching Volunteered Geographic Information: Spatial Data, Geographic Research and New Social Practice," *Annals of the Association of American Geographers* 102: 571–590.

Gardner-Santana, Lynne C., Gregory Glass, Douglas E. Norris, Christen M. Fornadel, Ella R. Hinson, and Sabra L. Klein. 2009. "Commensal Ecology, Urban Landscapes, and Their Influence on the Genetic Characteristics of City-Dwelling Norway Rats," *Molecular Ecology* 18: 2766–2778.

Getis, Arthur., and B. Boots. 1971. "Spatial Behavior: Rats and Man," *The Professional Geographer* 23: 1, 11–14.

Glass, Gregory E., Lynne C. Gardner-Santana, Robert D. Holt, Jessica Chen, Timothy M. Shields, Manojit Roy, Stephen Schachterle, and Sabra L. Klein. 2009. "Trophic Garnishes: Cat-Rat Interactions in an Urban Environment," *PLoS One* 4 (6): e5794.

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, Nancy, Pamela Waterman, Jarvis T. Chen, Mah-Jabeen Soobader, S.V. Subramanian, and Rosa 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.

Lambert, Mark S., Roger J. Quy, Rob Smith, and Dave P. Cowan. 2008. "The Effect of Habitat Management on Home-Range and Survival of Rural Norway Rat Populations." *Journal of Applied Ecology* 45: 1753–1761.

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

Wilson, Ron, and Alexander Din. 2018. "Understanding and Enhancing the U.S Department of Housing and Urban Development's ZIP Code Crosswalk files," *Cityscape* 20 (2): 277–294.