Accounting for ZIP Code Boundary Changes, 1990 - 2010

Amy Kate Bailey[1]
Allison Suppan Helmuth
September 2015

Submitted for consideration to the
Population Association of America
for its 2016 annual meeting.

[1] Draft version, not intended for public distribution. Please direct correspondence to Prof. Amy Kate Bailey, Sociology Department – M/C 312, 1007 W. Harrison St. – BSB 4112, University of Illinois at Chicago, Chicago, IL 60607, 312-996-6559, akbailey@uic.edu.

ABSTRACT: Every year the US Postal Service makes hundreds of changes to ZIP Code boundaries: dissolving existing ZIP Codes, establishing new ones, or realigning the boundaries between established ZIP Codes. These changes do not occur randomly, but are concentrated in areas experiencing both population collapse and rapid population growth. Failing to account for these boundary changes over time, then, introduces systematic measurement error in identifying the socioeconomic and demographic characteristics of geographic areas. In this paper, we describe our creation of a data tool that will allow researchers to account for changes in ZIP Code boundaries between 1990 and 2010. This enables the linkage of consistent geographic units, along with their social, economic, and demographic characteristics, over time. We believe that the existence of this data tool will allow more researchers to more effectively incorporate ZIP Codes into their research, as cases, as contextual factors for fixed effects analyses using micro-level observations as cases, or as the "Level 2" or macro level in multilevel models.

Accounting for ZIP Code Boundary Changes Over Time

Introduction

The United States Census Bureau distributes aggregated data for a number of selected geographies, including census tracts, counties, school districts, and metropolitan areas. The five-digit ZIP Code is ubiquitously available in administrative data, and has been used as a geographic reference area in 82 articles indexed in the Sociological Abstracts database since 1990[2], with heaviest concentration of ZIP Code-based research focusing on environmental inequality, health disparities, and business patterns. However, as Kreiger and colleagues (2002) warn, utilizing ZIP Codes without accounting for changes to these administrative boundaries, and most specifically the changes in the geographic locations and populations to which these numeric designations refer, is fraught with risk. For example, in 2003-2004 alone, the US Postal Service announced nearly 250 changes, affecting more than 400 ZIP Codes nationwide.[3] One in five eliminated ZIP Codes were in either Kentucky or West Virginia. More than half of the newly-established ZIP Codes were in California, Florida, or Nevada.

The Census Bureau distributed indicator data aggregated to the ZIP Code level in 1980 and 1990 (Yandle and Burton 1996). Beginning in 2000, however, it began to utilize a different geographic area associated with a ZIP Code, called the ZIP Code Tabulation Area (ZCTA). The term "ZIP Code" references administrative areas defined by the US Postal Service, while "ZCTA" defines the areas referenced by aggregated data distributed by the US Census Bureau. ZIP Codes and ZCTAs with the same numeric designation typically refer

[2] Bibliographic list available on request.

[3] Author calculations based on data created from Postal Service announcements.

to similar geographic areas. However, ZCTAs are constructed by identifying the most common ZIP Code reported by residents in a given census block, and aggregating characteristics about blocks to form a ZCTA. Note that ZIP Codes with very small population are not assigned a ZCTA (Source: https://www.census.gov/geo/reference/zctas.html, accessed 25 September 2015). Population data are aggregated to the ZCTA level for decennial census enumerations in 2000 and 2010, and for the American Community Survey 5-year estimates beginning with the 2007-2011 release.

All administrative geographic areas pose challenges in uniformity and comparability of cases, known as the Modifiable Areal Unit Problem (MAUP) (Gehlke and Biehl 1934; Openshaw 1983). MAUP indicates the lack of measurement precision that results from the use of administratively-defined geographic units, which may vary widely in population or other key metrics of interest. This concept is perhaps best exemplified by state-level analyses that include both Rhode Island and California as unique, and supposedly comparable "cases," because they are both states. This despite the fact that California's population of nearly 39 million dwarfs that of Rhode Island's – at just over one million residents, Rhode Island is smaller in population than several of California's counties (US Census Bureau 2015). Using administratively-defined geographies as "cases," as Level 2 structures in multi-level models, or to provide contextual information on individuals used as cases, is also problematic because the logic behind drawing boundaries often adheres to administrative convenience, rather than to factors such as predominant language, economic interdependence, or other social factors that indicate a relatively cohesive community.

Using ZIP Codes and ZCTA data is doubly problematic because the US Postal Service changes ZIP Code boundaries with relative frequency, compared to the tempo of boundary changes affecting many other geographic units, and there exists no readily available data tool to help researchers track these changes over time. Because boundary changes affect the geographic area referenced by a numeric code, in the case of ZIP Codes, they typically mean that the measured variables associated with a particular numerically-identified unit now refer to a different administratively-defined geographic area. Changes to ZIP Code boundaries are publicly announced, but we are unaware of any existing mechanism allowing researchers to link consistent geographies and the characteristics of their populations or economies across time when those changes are made.

If the change in ZIP Code boundaries were merely a random process, this could potentially be ignored as only so much "noise" in the data. However, changes in ZIP Code boundaries seem to cleave to three specific social processes: 1) population growth; 2) population decline; and 3) growth in the level of socioeconomic segregation. Because these phenomena are of intrinsic value to social scientists, the loss of a large number of cases affected by these population dynamics, or the decline in measurement precision for socioeconomic and demographic indicators in areas experiencing a socially important transition, means that our estimates of the effects of these transitions, or "static" metrics associated with them, will be biased.

This poses a unique challenge when dealing with Census data prior to 2011, the year that American Community Survey 5-year estimates first reference for ZIP Codes.[4] This is

[4] Note that aggregated ZCTA data in geographies refers only to the boundaries existing during the final year of observation.

because the population Census is fielded only once each decade. Therefore, if the boundaries for a specific ZIP Code changed between 1990 and 2000, for example, the geographic areas referenced by the same numeric code are not the consistent for the two years of observation. Measures of change in key social metrics, such as racial composition or the poverty rate, then, will be skewed. Measures of the association between factors occurring after the boundary change and ZIP Code-level socioeconomic or demographic characteristics measured in the earlier time period will also include systematic bias. We are in the process of creating a "Data Crosswalk" tool that will allow scholars to correct for this source of measurement error.

The Data Crosswalk

The work that we discuss in this paper identifies our construction of a "crosswalk" data file to allow researchers to link data across time to map onto consistent geographic areas. This approach builds on the techniques developed by prior scholars. For example, the Missouri Census Data Center, housed at the University of Missouri, makes a Geographic Correspondence Engine (or "geocorr") tool available online to allow the mapping of disparate geographies *within the same time period* (Missouri Census Data Center 1990). So, for example, aggregated data from the census tract level can be combined to allow for analyses of counties or 5-digit ZIP Codes. Other, similar tools allow for the tracking of changes within a consistent geographic unit over time, such as Patrick Horan and Peggy Hargis's data file (1995) to link southern counties across several decades. Most recently, nd most directly associated with our work, the University of Minnesota's Minnesota Population Center (2011) has constructed an array of tools allowing more consistent ZIP Code referencing. We build on all of these efforts with our current work.

Our labor began by documenting every ZIP Code change listed in the "Post Office Changes" section of *The Postal Bulletin* published between 1990 and 2010. *The Postal Bulletin* is an administrative document published every two weeks by the United States Postal Service, and intended for use by USPS personnel to inform them of changes in policies, services, and products. Roughly twenty times each year, *The Postal Bulletin* includes a listing of changes affecting ZIP Codes. Many of these do not involve boundary changes – for example, a listing may announce the reassignment of a ZIP Code to a new administrative office, or the change in the classification of the post office that delivers to a particular ZIP Code. Frequently, however, changes affecting the geographic area that a particular ZIP Code references are announced. These are typically of four key types:

1) The realignment of the geographic boundaries between two or more ZIP Codes;
2) The division of an existing ZIP Code into two or more surviving geographic units;
3) The combining of two or more existing ZIP Code into a single surviving geographic unit; and
4) The establishment of a delivery area (meaning a specific geographic footprint) for a ZIP Code that had previously only included PO Box or "general delivery" mail.

Once we identified *where* and *when* changes affecting ZIP Code boundaries had occurred, we classified each of the affected ZIP Codes into a "ZIP Code cluster," with specific temporal parameters for use by other researchers. These clusters identify the smallest, consistent geographic unit that is indivisible across a decade. So, for example, if a ZIP Code that existed in 1990 was divided two years later into two smaller ZIP Codes, those two resultant ZIP Codes would be "bundled" into a cluster for the decade between 1990 and 2000. In essence, the characteristics of the ZIP Code that existed in 1990 could then be

meaningfully compared to the *combined* characteristics of the two "daughter" ZIP Codes from the 2000 census, since they reference the same geographic area. Similarly, if a ZIP Code that had existed in 1990 was subsumed under another ZIP code before the 2000 population census, the two ZIP Codes that existed in 1990 would be combined into a single ZIP Code cluster for the 1990-2000 decade. This would mean that the *combined* characteristics of the two ZIP Codes that existed in 1990 could be meaningfully compared to the characteristics of the surviving ZIP Code in 2000.

We have thus far completed the process of identifying ZIP Codes that have experienced boundary changes for all but four years between 1990 and 2010, and have hired an undergraduate student to complete data entry for the remaining years. Once we identified *which* ZIP Codes had been affected by boundary changes, we also linked them between Census years by assigning them to a ZIP Code cluster. Note that the vast majority of ZIP Codes do not experience a boundary change in a given year. We have begun coding for a data program file (referred to colloquially as a "cross walk") to link all consistent ZIP Code clusters between 2000 and 2010, and will similarly complete a "cross walk" for the years between 1990 and 2000.

ZIP Codes Referencing Only Post Office Boxes

As a separate step, we identified ZIP Codes that were newly created, or converted from existing ZIP Codes, to service only PO Boxes. These ZIP Codes do not correspond to a particular geographic area, and therefore are not able to be linked to aggregated Census data on ZCTAs. We further identified ZIP Codes that provided only PO Box and "general delivery" services that subsequently added delivery to a specific geographic area. In 2003 and 2004, the US Postal Service announced 97 changes that involved the creation or

elimination of a ZIP Code dedicated exclusively to PO Boxes, or that expanded the function of a ZIP Code to include delivery to a specific geographic area. We created a dummy variable, coded as 0 if the ZIP Code was not a PO Box-only ZIP Code, and 1 if it referenced exclusively PO Boxes. Note that this coding scheme also specifically references the dates during which a specific ZIP Code was a PO Box ZIP Code, and utilizes information on when delivery was launched within the ZIP Code. To provide some concrete examples:

- The ZIP Code 30046, in Gwinnett County, Georgia, only began offering home delivery on 1 July 2009. Prior to that date, it was coded "1" to indicate a "PO Box only" ZIP Code.
- ZIP Code 50295, in Louisville, Kentucky, was an existing ZIP Code that restricted itself to offering only PO Boxes on 1 December 2005.
- On 27 July, 2001, a new ZIP Code, 98162, was created specifically for the PO Boxes that had previously been associated with a "multi-use" ZIP Code 98199 (meaning it provided delivery services as well as PO Boxes).

Future Steps

We expect to have completed the construction and testing of the "crosswalk" file by early January. We also plan to estimate the error resulting failing to account for changes in ZIP Code boundaries for two metrics related to our own research interests: the degree of residential segregation, as measured by a ZIP Code-level entropy score, and the rate of military enlistment.

BIBLIOGRAPHY

Gehlke, C.E., and Katherine Biehl. 1934. "Certain Effects of Grouping on the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association* 29 (185): 169-170.

Horan, Patrick M., and Peggy G. Hargis. 1995. "County Longitudinal Template, 1840–1990." Computer file. Inter-university Consortium for Political and Social Research, Ann Arbor, Mich.

Kreiger, Nancy, Jarvis T. Chen, Pamela D. Waterman, David H. Rejkopf, and S. V. Subramanian. 2003. "Race/Ethnicity, Gender, and Monitoring Socioeconomic Gradients in Health: A Comparison of Area-Based Socioeconomic Measures – The Public Health Disparities Geocoding Project." *American Journal of Public Health* 93 (10): 1655 – 1671.

Kreiger, 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* 97 (2): 1100-1102.

Minnesota Population Center. 2011. *National Historical Geographic Information System: Version 2.0*. Minneapolis, MN: University of Minnesota

Missouri Census Data Center. 1990. MABLE/Geocorr90 Geographic Correspondence Engine. .accessed from http://mcdc2.missouri.edu/websas/geocorr90.shtml on 14 July 2015.

Openshaw, Stan. 1983. *The Modifiable Areal Unit Problem*. Norfolk, VA: GeoBooks.

US Census Bureau. 2015. “State & County Quick Facts.” from http://quickfacts.census.gov /qfd/states/06000.html and http://quickfacts.census.gov/qfd/states/44000.html, accessed 6 September 2015.

Williams, Robert W. 1999. “The Contested Terrain of Environmental Justice Research: Community as Unit of Analysis.” *The Social Science Journal* 36 (2): 313-328.

Yandle, Tracy, and Dudley Burton. 1996. “Methodological Approaches to Environmental Justice: A Rejoinder.” *Social Science Quarterly* 77 (3): 520 – 527.