National ZIP Code Crosswalk (1990 – 2020)

User Guide & Codebook

Amy Kate Bailey, PhD*
University of Illinois-Chicago

Allison Suppan Helmuth, PhD
Rice University

* Corresponding Author: akbailey@uic.edu

Contents

Summary

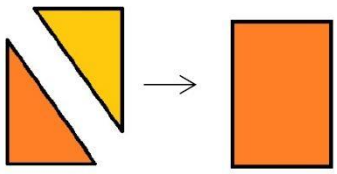
The *National ZIP Code Crosswalk (1990-2020)* was created to identify changes in U.S. Postal Service ZIP Code boundaries and mitigate data quality issues created by these changes. Since ZIP Codes are administrative tools for sorting mail, their geographic boundaries change in relationship to population growth and decline (and for a range of other reasons) resulting in measurement error (matching observations to the wrong contextual unit) or missing data (due to an observation reporting a ZIP Code that did not exist at the beginning of the observational period). The *National ZIP Code Crosswalk (1990-2020)* identifies changes in U.S. Postal Service ZIP Code boundaries between 1990 and 2020 and provides numeric codes that cluster the ZIP Codes into the smallest geographic unit, or group of ZIP Codes, that are consistent across a decade: 1990 – 2000; 2000 – 2010; and 2010 – 2020. This "crosswalk" covers the contiguous United States, Alaska, Hawaii, and the District of Columbia. The crosswalk may serve as a useful tool for the analysis of datasets where ZIP Codes are the smallest available geographic unit, such as administrative data about patients, businesses, or survey respondents.

Overview

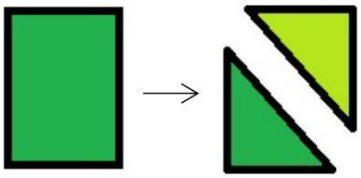
ZIP Codes are administrative codes generated by the United States Postal Service (USPS) that refer to the geographic area covered by a specific set of mail delivery routes. The U.S. Census Bureau calculates and distributes aggregated social, economic, and demographic information for the population associated with "ZIP Code Tabulation Areas" (ZCTAs), which are roughly analogous to ZIP Codes and serve as identifiers for specific neighborhoods and communities. These aggregated census data, however, are unable to account for changes in ZIP Code boundaries that occur between decennial censuses, leading to measurement error and missing data problems for scholars who attempt to use the aggregated ZCTA data. The purpose of this crosswalk file is to allow researchers to overcome this limitation, enabling them to appropriately link spatial reference information (ZIP Codes) with characteristics of the populations to which they refer.

Most ZIP Codes do not change boundaries in a decade, but a large enough percentage do as to create a problem with missing or mis-specified data. Boundary changes typically involve one or more of the following three processes, although a small number of cases do not conform to these typologies:

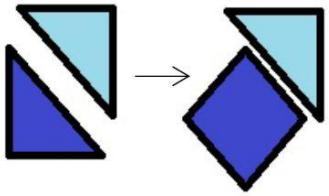
1) Two or more existing ZIP Codes are combined to create a single surviving ZIP Code.



2) An existing ZIP Code is divided into multiple resulting ZIP Codes.



3) Boundaries between two or more existing ZIP Codes are altered.



Each of these types of changes alters the geographic area that a ZIP Code refers to, and as such, the spatial unit identified by the ZIP Code includes a different population, with a different array of characteristics. By linking the spatial units associated with ZIP Codes as these boundary changes are enacted, we can both prevent the loss of observations due to missing data, and more accurately measure social, demographic, and economic characteristics associated with each ZIP Code.

Data Collection Process

Data for the *National ZIP Code Crosswalk (1990-2020)* were collected using information contained in the *The Postal Bulletin*, an administrative publication generated every two weeks by the U.S. Postal Service. It includes information on a variety of topics relevant to postal employees, including planned or accomplished modifications to ZIP Code boundaries. *The Postal Bulletin* also includes the date on which the changes took, or will take, effect.

We reviewed each issue of *The Postal Bulletin* that was published between January 1, 1990 and December 31, 2020 (a total of 781 issues), and identified all instances where ZIP Code boundaries were changed. We also identified those ZIP Code changes that signaled that a particular ZIP Code was used exclusively for Post Office Boxes. We then recorded each of these changes into an Excel spreadsheet, organized by state, and in the chronological order in which the announcements appeared.

As such, this crosswalk includes separate state-by-state master spreadsheet for the periods spanning 1990 through 2000, 2000 through 2010, and 2010 through 2020.

The information we recorded to document a ZIP Code boundary change includes: the specific ZIP Codes affected; the date on which the changes were scheduled to take effect (or had already been adopted); location information regarding the county and name of the governing Post Office; the date and issue of the *Postal Bulletin* where the announcement appeared; an indication of whether this announcement amends a prior announcement; and the specific information associated with the change (i.e., what *exactly* the boundary changes entailed).

Based on this master list of all modifications to ZIP Code boundaries, we identified clusters of ZIP Codes for each decade that need to be considered together to guarantee a consistent geographic spatial reference. In many instances, a cluster consists of only two ZIP Codes that were merged together, were created by the division of an existing ZIP Code, or that experienced a change in their shared boundary. Sometimes, however, a cluster may involve many more ZIP Codes. This was particularly common if a geographic area experienced rapid population growth or decline which could have necessitated the expansion or decline of mail delivery routes in that geographic area, leading to a change in the shape or geographic size of the ZIP Code itself. In these cases, all of the ZIP Codes that cover the same geographic reference area are clustered together.

Materials Included in this Collection

This collection includes two key types of files. The first are Stata data files that provide a crosswalk to accurately assign ZIP Codes that have experienced boundary changes to a cluster of ZIP Codes. The second are Stata .do files that enable users to add a column of ZIP Code clusters to their existing administrative data. These files are discussed in greater detail below. We also include a series of Excel files to provide additional documentation for the ZIP Code boundary changes underlying these data files.

Stata Data Files

We have created a Stata data file for each decade (1990 – 2000, 2000 – 2010, and 2010 – 2020) including five variables related to each change affecting ZIP Code boundaries: 1) The State FIPS Code; 2) The two-letter state postal abbreviation; 3) The cluster to which a ZIP Code should be assigned; 4) The ZIP Code that was affected by the change; 5) The year in which the boundary change took place. If your data only reference a particular bandwidth of years within a decade, or a particular subset of states, you will want to restrict by that time period or spatial area.

We suggest using these files as follows:

1) Sort your analytic data set by ZIP Code;
2) Sort the boundary change file by ZIP Code;
3) Merge these two data sets by ZIP Code. This should create quite a large number of observations with missing values for the ZIP Code cluster variable. This is because most ZIP Codes do not change their boundaries in a given decade, and therefore do not need to be bundled with other ZIP Codes;
4) Replace missing ZIP Code cluster values with the observation's ZIP Code;

5) Collapse across observations by ZIP Code cluster to generate new values – for example, total population.

Stata .do Files

We created a Stata .do file for each decade (1990 – 2000, 2000 – 2010, 2010 – 2020). These files are designed to assign ZIP Code cluster codes with specific dates of transition. So, for example, if your records relate to business dissolution or hospital admission and you have the exact date of the event, using this file will enable you to cluster ZIP Codes for these events. Using the date of the ZIP Code boundary change allows to ensure that your geographic clustering only affects observations where the event of interest occurred after the ZIP Code boundary changes were enacted. In order to successfully use these files, you must have a date for each observation, formatted as follows: YYYYMMDD.

To use these files, check that the variable in your data indicating the date when the event of interest occurred, is called "date" and is correctly formatted. Also confirm that the variable for the ZIP Code is titled "zipcode". You should then be able to run this file. It will first create a new variable called "zipclust" – initially set at the same value as the "zipcode" variable. It will then replace the value for "zipclust" only for those observations where the event of interest from your data occurred on or after the date when the ZIP Code boundary change was implemented.

Excel Documentation Files

These files are designed to document each change to ZIP Code boundaries, as published in the *Postal Bulletin*. There are three decade-specific workbooks: 1990 – 2000; 2000 – 2010; 2010 - 2020. Each workbook contains 51 spreadsheets – one for each of the 50 states plus the District of Columbia. It lists the following information for each announced change affecting ZIP Code boundaries:

- Old ZIP Code
- New ZIP Code
- New Post Office station name
- New County
- Effective Date
- Wording that describes the change (a direct transcription)
- *Postal Bulletin* issue when then change was announced.
- Date of publication for the *Postal Bulletin* that announced the change
- Old Post Office station name
- Old County

Bibliographic Citation

Publications that utilize this crosswalk in data preparation or analyses for published work, or in public presentations, should include the following citation in their list of bibliographic references or footnotes. The preferred bibliographic citation for the current version of this data tool is:

Bailey, Amy Kate and Allison Suppan Helmuth. 2023. *National ZIP Code Crosswalk, 1990-2020.*

Funding Acknowledgement

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, to the Center for Studies in Demography & Ecology at the University of Washington; the University of Illinois at Chicago's Office of Social Science Research; and Grant #UTA01060 from the Utah Agricultural Experiment Station.

Data Disclaimer

These data are complete and accurate to the best knowledge of its producers. To the extent that they contain omissions or errors, the authors encourage users to send a notification email to be incorporated into subsequent releases and updates of the data.