Bridging the Geographic Divide: Crosswalks Across Space and Time*

Chandler Lutz†
University of North Carolina, Charlotte
Belk College of Business

February 6, 2025

Abstract

This paper presents a straightforward algorithm that leverages shapefile maps to create "crosswalks" that link data across disparate geographical delineations. Compared to prior work, our approach eases implementation, reconciles boundary differences across both space and time, and flexibly allows researchers to choose spatial weights that better capture the geographical agglomeration of economic activity. In an application, we create a crosswalk for employment data across Connecticut's new 2022 county boundaries. Accounting for the high degree of clustering in employment leads to markedly different crosswalk allocations relative to conventional approaches that use population or land area as spatial weights.

JEL Classification: R12, R30, C18, C80;

Keywords: Geographic Crosswalks, Spatial Matching, Boundary Reconciliation, Shapefiles

*Software package to create crosswalks: https://github.com/ChandlerLutz/geolinkr

†https://chandlerlutz.github.io/. chandler.lutz@charlotte.edu

Empirical comparisons across space and over time power much of modern economics. A key first step in such analyses often involves harmonizing data across disparate geographic delineations. Mismatched delineations arise from data contemporaneously collected using different boundaries (e.g., zip codes or school districts versus counties), from boundary definitions changing over time, or both. As an example of temporally shifting boundaries, consider the widely-used BLS Quarterly Census of Employment and Wages (QCEW) that Bartik (1991), for example, employs to construct his famous shift-share instrument and track local employment over time.[1] In 2022, Connecticut changed its county definitions (U.S. Census Bureau, 2022), hindering longitudinal employment or income comparisons using the QCEW. As the U.S. Bureau of Labor Statistics (2024) writes: "There is no direct mapping between the old and new [Connecticut] counties. Since QCEW is not a time series product, switching to these new county definitions creates [a] break in Connecticut data." Such boundary changes or other misaligned delineations often force researchers to pursue bespoke solutions that can impede knowledge generation, introduce errors, or limit reproducibility.

This paper introduces a straightforward algorithm to reconcile differing boundary definitions using widely available GIS shapefile maps. Relative to previous work, our approach readily scales to a variety of applications, enabling researchers to swiftly generate "crosswalks" that convert data between delineations—all while accounting for the uneven distribution of economic activity across space.[2] Our ease of implementation facilitates crosswalk creation and also allows researchers to compare crosswalks, for example, using the specification tests in Millimet (2024).

More specially, we require a "source" shapefile with the data's original delineations, a "target" shapefile with the desired delineations, and a "weights" shapefile with a variable to proxy the spatial distribution of the source data. The weights shapefile should be the smallest possible level of aggregation (e.g., Census blocks, Census tracts, or zip codes) and cover the source and target shapefiles.[3] Noting that a polygon within a shapefile represents a singular geographic unit (e.g., a county or school district), the algorithm consists of the following steps:

1. Find all source polygons completely covered by target polygons. Fully allocate these source

[1] Bartik (1991, p. 161 and 275) uses the BLS ES-202 program, now called the BLS QCEW, to measure MSA employment in aggregate as a control variable or by industry to construct his shift-share instrument. See Bureau of Labor Statistics (2001) for a historical overview of the BLS QCEW program.

[2] Software implementation at `https://github.com/ChandlerLutz/geolinkr`.

[3] We make shapefiles at the U.S. Census block, block group, or tract level, with population or household counts to be used as weights available here: `https://github.com/ChandlerLutz/census-blocks-tracks-shp`.

polygons to their corresponding target.

2. Intersect all remaining source and target polygons.
3. Find all polygons from the weights shapefile covered by an intersection from (2). Fully allocate the weighting variable from each of these covered polygons from the source to the corresponding target.
4. For any remaining polygons in the weights shapefile that intersect with two or more polygons from (2), allocate the weight variable based on the land area share according to the intersection between the weight polygons and the polygons from (2).

The algorithm accounts for the clustering of economic activity across space in steps (3) and (4) by weighting source–target intersections using the spatial distribution from the weighting variable in the weights shapefile.

These steps also highlight another advantage of our approach: Any variable can easily serve as weights. Because all crosswalks assume that weighting and source data share similar spatial distributions, the ability to select more appropriate weights allows researchers to better align these distributions, reducing errors and improving crosswalk accuracy. This becomes especially important when the geographic dispersion of traditional weighting variables—such as population, household counts, or land area—fails to reflect the source data. For instance, when constructing a crosswalk for the QCEW across Connecticut's changing boundaries below, we find more clustering in employment than population, suggesting that employment counts better capture the spatial distribution of the source data.

Note that step (4) uses areal weighting (AW) for any source–target intersections that do not fully cover polygons in the weights shapefile. Thus, using a highly disaggregated weights shapefile (e.g., Census blocks, Census tracts, or zip codes) minimizes potential misallocations from the source to the target.

The algorithm outputs a crosswalk that maps source polygons to their respective targets. Output data also include the weight variable for each source–target pair, aggregated according to their spatial intersection, and the allocation share from the source to the target. The allocation shares sum to 1 for each source polygon. We implement this algorithm and data validation checks

into a convenient software package that only requires a few lines of code to create a crosswalk.[4]

Our implementation is an extension or generalization of existing methods in economics, finance, and political science, focusing on flexibility and simplicity. Researchers can thus use our algorithm and accompanying software to readily synchronize spatial data collected at different levels of aggregation or with temporally changing boundary definitions.

A large literature aims to map data from one set of delineations to another. Millimet (2024) highlights the challenges of current approaches: Researchers often use ad hoc crosswalks or shoehorn their data into existing methods, potentially introducing "severe" errors in their estimates. At an extreme, researchers employ multiple crosswalks or eschew them altogether when existing methods prove incompatible with their data.[5]

On the practical side, popular crosswalks include MCDC Geocorr (2022), to reconcile Census boundaries, and the HUD USPS crosswalk, to map zip codes to Census tracts and counties (Din and Wilson, 2020).[6] MCDC Geocorr only produces crosswalks within the same decennial Census using land area, population, or household counts as weights. Our algorithm improves upon the MCDC Geocorr by also creating crosswalks that link boundaries over time and allowing additional weighting variables. In tests, our algorithm outputs crosswalks nearly identical to MCDC Geocorr.[7] Like the HUD USPS crosswalk, our algorithm can reconcile zip code, census tract, and county boundaries, but it can also create crosswalks for several other delineations using various weighting variables. Our approach thus extends these methods, allowing researchers to improve accuracy by creating crosswalks that better reflect their source data.

On the research side, the literature started creating crosswalks using AW and thus assuming a uniform distribution of the source data across space.[8] Yet economic data cluster (Marshall, 1920), often making AW inaccurate (Gregory, 2002; Schroeder, 2007; Ferrara et al., 2024). Nonetheless, our algorithm supports AW by setting the weighting variable to a constant.

More recent advancements link Census delineations over time. Our algorithm and these studies

[4] `https://github.com/ChandlerLutz/geolinkr`.

[5] For example, Millimet (2024) notes how Agarwal et al. (2018) drop 17% of their sample because they cannot crosswalk zip codes to congressional districts.

[6] See Millimet (2024), table I for other crosswalks.

[7] Our algorithm produces nearly identical crosswalks to MCDC Geocorr (Lutz, 2024c).

[8] See Markoff and Shapiro (1973); Goodchild and Lam (1980); Horan and Hargis (1995); Hornbeck (2010); Pebesma (2018); Prener and Revord (2019); Eckert et al. (2020).

share the same underlying idea: distribute data across boundaries by weighting source–target intersections. Yet this literature largely tackles cases where no weighting variable readily exists, such as temporal crosswalks for Census blocks, by either making additional assumptions (Howenstine, 1993; Mugglin and Carlin, 1998; Schroeder, 2007) or by including auxiliary information such as road networks or satellite imagery.[9] Our key insight relative to this work is that such additional assumptions or information are superfluous for most economic crosswalks: Block, block group, tract, and zip code shapefiles exist back to 1990 (or earlier in some cases), ready to serve as weights.[10] This allows us to simplify implementation and generate most crosswalks pertinent to social science research. We do note our algorithm is not appropriate for a Census block-to-block crosswalk, for example, where a more disaggregated shapefile does not exist to serve as weights. Yet researchers have already tabulated such crosswalks using the foregoing methods (Manson et al., 2024; NHGIS, 2024b), allowing us to focus on other use cases.

In another related study, Ferrara et al. (2024) use historical weights from power law-based extrapolations of U.S. urban area populations over time (from Fang and Jawitz (2018)) or historical urban settlements from Zillow property records data (from Leyk and Uhl (2018)). With this innovative use of weights, Ferrara et al. (2024) produce county and congressional district crosswalks back to 1790. Our algorithm easily allows for such weighting schemes but extends to other applications as well.

Finally, Millimet (2024) suggests two specification tests to assess crosswalk accuracy. Our algorithm and implementation facilitate these tests. The first test compares regression estimates using the ported data when the true crosswalk is known (due to a one-to-one correspondence between the target and the source) to those when it is not (when source delineations map to more than one target). Rejecting the null hypothesis that the two regression estimates are equal suggests crosswalk inaccuracy. Step 1 of our algorithm outputs the one-to-one correspondences to help implement this test.[11] The second tests if additional information from alternative crosswalks increases predictive power in a regression model on the ported data. Rejecting the null hypothesis that the data mapped using the alternative crosswalks have no additive predictive power again

[9] For studies that use auxiliary information, see U.S. Census (2001), Holt et al. (2004), Reibel and Bufalino (2005), ESRI (2025) and the references in Manson et al. (2024), NHGIS (2024a) and PySAL (2024).

[10] For repositories of shapefiles, see Walker (2016), U.S. Census (2024), Manson et al. (2024), and NHGIS (2024c).

[11] See Lutz (2024a) for our software implementation.

indicates crosswalk inaccuracy. Our implementation facilitates this test by allowing researchers to quickly generate crosswalks using various weights.

1 Example

To better understand the algorithm, Figure 1 demonstrates creating a crosswalk from 2023 Connecticut (CT) counties (the "source" delineations) to 2020 CT counties (the "target" delineations) to temporally link the QCEW data. The plot focuses on Fairfield County (blue lines), CT's most populous county in 2020. CT's 2022 changes split Fairfield into three new counties: Western Connecticut, Greater Bridgeport, and Naugatuck Valley (red lines).

Step (1) of the algorithm finds all 2023 counties completely covered by Fairfield; in this case, just Greater Bridgeport as seen in panel A (light gray area). Thus, the crosswalk fully allocates Greater Bridgeport to Fairfield. Western Connecticut and Naugatuck Valley only partially intersect with Fairfield, meaning that the crosswalk must allocate a share, but not all, of these counties to Fairfield. The dark gray areas in panel A identify these intersections, corresponding to step (2) of the algorithm.

Note that AW algorithms would stop at step (2) and, in this case, would just allocate Western Connecticut and Naugatuck Valley according to the land area shares of their intersections with Fairfield (dark gray area).

Yet economic activity exhibits substantial clustering across space. Figure 1, panel B adds the zip codes that intersect with Naugatuck Valley or Western Connecticut, with darker shading representing higher employment according to the Census's 2022 zip-code-level survey of businesses in the County Business Patterns (CBP) dataset. Note that this data count employment at the job location (rather than where employees live), and white areas have little employment or are unused (e.g., the southern part of Western Connecticut in the Long Island Sound). The plot shows that employment ranges widely across the plotted zip codes, with an interquartile range of 6471 workers around a median of 4112 and a standard deviation (8408.96) that is 118% of the mean (7121.86). Moreover, employment clusters in just a few zip codes: Among the 62 plotted zip codes, the top 3 account for 22.4% of all employment.

Accounting for this clustering is crucial for crosswalk creation. For example, approximately 18.4% of employment in Naugatuck Valley lies at its intersection with Fairfield relative to just

7.6% of its land area (southern part of Naugatuck Valley in panel B), documenting the drawbacks of AW and underscoring how an accurate crosswalk must reflect the uneven nature of economic activity across space.[12]

Crosswalk methods that capture agglomeration effects do so through spatial weights. Although employment (Figure 1, panel B) likely represents the most appropriate weighting variable for the QCEW, other crosswalk approaches typically use population or household counts. Panel C plots population by zip code from the 2021 IRS Statistics of Income (SOI). Compared to employment, population exhibits substantially more dispersion: Population's top 3 zip codes contain just 15.6% of population (versus 22.4% for employment), and population has a Gini coefficient across zip codes of just 0.39 (versus 0.56 for employment). Thus, the spatial distributions of employment and population differ substantially. Using population instead of employment as weights, as in previous crosswalk approaches, would likely introduce errors in crosswalk creation and ultimately lead to an incorrect representation of the source data in the target delineations.

To see this, Figure 2 plots the share of each 2023 county allocated to Fairfield, weighting by employment (red), population (purple), household counts (blue), or land area (green). For population and household counts, we use Census tract-level data from the 2019–23 American Community Survey (ACS), as the ACS has better coverage across CT than the IRS SOI data.

Panel A shows that since Fairfield covers Greater Bridgeport, the allocation share equals 1 regardless of the weighting variable. This is step (1) of the algorithm that identifies instances where the target covers the source without using weights.

Differences in the allocations across weighting variables surface in panels B and C for Naugatuck Valley and Western Connecticut, 2023 source counties that partially intersect with Fairfield. For Naugatuck Valley, these differences are stark: Weighting by employment allocates 18.4% of Naugatuck Valley to Fairfield, versus just 9.2% when using either population or land area weights, or 8.6% with household counts. In other words, assuming that population, household counts, or land area reflect the geographic distribution of the QCEW data, as in other crosswalk methods, would substantially underestimate the spatial concentration of employment and introduce errors when

[12]Zip codes 06787 and 06010 have a small amount of land area outside of Naugatuck Valley, while Naugatuck Valley only intersects with a small portion of zip code 06492. For this employment calculation, we do not include zip code 06492.

mapping this data across delineations. Such errors would reverberate through a research study as boundary synchronization is typically the first step in the data analysis chain, highlighting the importance of weight variable choice in crosswalk creation.

For Western Connecticut, the differences are less extreme but still notable. With employment weights, the algorithm allocates 97.4% of Western Connecticut to Fairfield, compared to just 95.2% for population, 95.0% for household counts, and 95.9% for land area weights.

We provide the full CT crosswalks from 2023 to 2020 delineations using employment, population, and land area weights online (Lutz, 2024b). Finally, creating a crosswalk from 2020 to 2023 CT boundaries would simply reverse the steps presented here.

2 Conclusion

This paper outlines a simple algorithm that researchers can employ to create crosswalks that align data across mismatched geographic delineations. The algorithm and the corresponding software ease implementation and increase flexibility compared to existing approaches, facilitating knowledge generation and reproducibility throughout social science research.

References

S. Agarwal, G. Amromin, I. Ben-David, and S. Dinc. The politics of foreclosures. *The Journal of Finance*, 73(6):2677–2717, 2018.

T. J. Bartik. Who benefits from state and local economic development policies? 1991.

Bureau of Labor Statistics. *BLS Handbook of Methods*. 2001. URL https://www.bls.gov/cew/publications/additional-publications/archive/old-handbook-of-methods.htm. Accessed: December 31, 2024.

A. Din and R. Wilson. Crosswalking zip codes to census geographies. *Cityscape*, 22(1):293–314, 2020.

F. Eckert, A. Gvirtz, J. Liang, and M. Peters. A method to construct geographical crosswalks with an application to us counties since 1790. *Working Paper, National Bureau of Economic Research*, 2020.

ESRI. Data apportionment and layers, 2025. URL https://pro.arcgis.com/en/pro-app/latest/help/analysis/business-analyst/data-apportionment-and-layers.htm. Accessed: January 25, 2025.

Y. Fang and J. W. Jawitz. High-resolution reconstruction of the united states human population distribution, 1790 to 2010. *Scientific data*, 5(1):1–15, 2018.

A. Ferrara, P. A. Testa, and L. Zhou. New area- and population-based geographic crosswalks for u.s. counties and congressional districts, 1790–2020. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(2):67–79, 2024. doi: 10.1080/01615440.2024.2369230.

M. F. Goodchild and N. S.-N. Lam. Areal interpolation: A variant of the traditional spatial problem. *Geo-processing*, 1(3):297–312, 1980.

I. N. Gregory. The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, environment and urban systems*, 26 (4):293–314, 2002.

J. B. Holt, C. Lo, and T. W. Hodler. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31(2):103–121, 2004.

P. M. Horan and P. G. Hargis. County longitudinal template, 1840-1990. 1995.

R. Hornbeck. Barbed wire: Property rights and agricultural development. *The Quarterly Journal of Economics*, 125(2):767–810, 2010.

E. Howenstine. Measuring demographic change: The split tract problem. *The professional geographer*, 45(4):425–430, 1993.

S. Leyk and J. H. Uhl. Hisdac-us, historical settlement data compilation for the conterminous united states over 200 years. *Scientific data*, 5(1):1–14, 2018.

C. Lutz. Generate one-to-one correspondence weights between spatial datasets, 2024a. URL https://github.com/ChandlerLutz/geolinkr/blob/1d62181c782918eae296304d506df0fc69b4509c/R/get_one_to_one_cw.R. Accessed: January 27, 2025.

C. Lutz. Appendix Data for "Bridging the Geographic Divide: Crosswalks Across Space and Time", 2024b. URL https://github.com/ChandlerLutz/geolinkr_app_data. Accessed: January 27, 2025.

C. Lutz. Test geolinkr vs Missouri Geocorr, 2024c. URL https://github.com/ChandlerLutz/geolinkr/blob/main/tests/slow-tests/test-create_cw_worker_vs_missouri_geocorr.R. Accessed: February 6, 2025.

S. Manson, J. Schroeder, D. Van Riper, K. Knowles, T. Kugler, F. Roberts, and S. Ruggles. Ipums national historical geographic information system: Version 19.0 [dataset]. dataset, 2024. URL http://doi.org/10.18128/D050.V19.0.

J. Markoff and G. Shapiro. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter*, 7(1):34–46, 1973.

A. Marshall. Principles of economics, 8e éd. *Londres, Macmillan [trad. cast. Aguilar].(1961): Principles of economics*, 9:1867–1890, 1920.

MCDC Geocorr. Missouri census data center. geocorr 2022: Geographic correspondence engine, 2022. URL https://mcdc.missouri.edu/applications/geocorr.html. Accessed: January 25, 2025.

D. L. Millimet. (don't) walk this way: The econometrics of crosswalks. 2024.

A. S. Mugglin and B. P. Carlin. Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 111–130, 1998.

NHGIS. 2000 block data standardized to 2010 geography, 2024a. URL https://www.nhgis.org/2000-block-data-standardized-2010-geography. Accessed: December 31, 2024.

NHGIS. https://www.nhgis.org/geographic-crosswalks, 2024b. URL https://www.nhgis.org/geographic-crosswalks. Accessed: January 27, 2025.

NHGIS. Gis files, 2024c. URL https://www.nhgis.org/gis-files. Accessed: January 27, 2025.

E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

C. G. Prener and C. K. Revord. areal: An r package for areal weighted interpolation. *Journal of Open Source Software*, 4(37):1221, 2019.

PySAL. Tobler references, 2024. URL https://pysal.org/tobler/references.html. Accessed: January, 28, 2025.

M. Reibel and M. E. Bufalino. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1):127–139, 2005.

J. P. Schroeder. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis*, 39(3):311–335, 2007.

U.S. Bureau of Labor Statistics. New 2024 connecticut counties, 2024. URL https://www.bls.gov/cew/classifications/areas/new-2024-connecticut-counties.htm. Accessed: December 31, 2024.

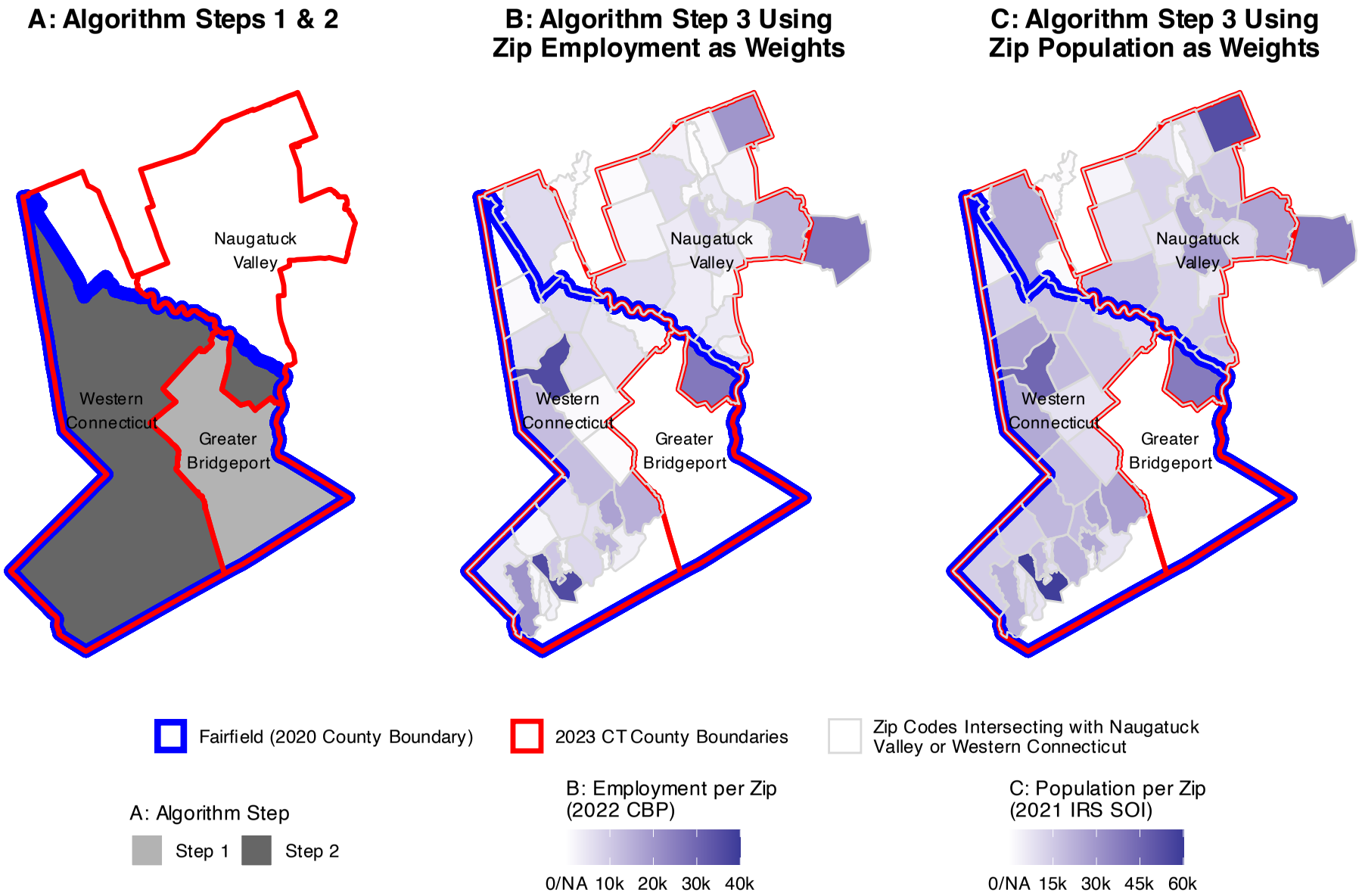
U.S. Census. Census 2000. census tract relationship files, 2001. URL http://web.archive.org/web/20020413105513/ttp://www.census.gov/geo/www/relate/rel_tract.html. Accessed: January 27, 2025.

U.S. Census. Census 2000. census tract relationship files, 2024. URL https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. Accessed: January 27, 2025.

U.S. Census Bureau. Change to county equivalents in the state of connecticut. *Federal Register*, 87 (107):33379–33380, June 2022. URL https://www.federalregister.gov/documents/2022/06/06/2022-12063/change-to-county-equivalents-in-the-state-of-connecticut.

K. Walker. tigris: An R Package to Access and Work with Geographic Data from the US Census Bureau. *The R Journal*, 8(2):231–242, 2016. doi: 10.32614/RJ-2016-043. URL https://doi.org/10.32614/RJ-2016-043.

Figure 1: Crosswalking 2023 CT County Delineations to Fairfield County, CT (2020)



Notes: Panel A shows steps (1) and (2) of our algorithm when creating a crosswalk for 2023 CT counties (red lines; the "source" delineations) to Fairfield County (blue lines; the 2020 "target" delineation). Fairfield County covers Greater Bridgeport (light gray area in panel A), meaning the algorithm fully allocates Greater Bridgeport to Fairfield in step (1). In step (2), the algorithm identifies the partial intersections between Fairfield and Western Connecticut and Naugatuck Valley, the dark gray areas in panel A. Panel B visualizes step (3) by layering the weighting variable, employment by zip code from the 2022 CBP, onto the county delineations. White areas in panel B are either uninhabited (e.g., the southern part of Western Connecticut in the Long Island Sound) or belong to Greater Bridgeport (accounted for in step (1) in panel A). Panel C repeats panel B but uses population by zip code from the 2021 IRS SOI.

Figure 2: Allocation Shares of 2023 CT Counties to Fairfield by Weight Variable