Local Landscapes of Assisted Housing: Reconciling Layered and Imprecise Administrative Data for Research Purposes

Shiloh Deitz
Will B. Payne
Eric Seymour
Kathe Newman
Lauren Nolan
Rutgers University

Abstract

Understanding the stock of rental housing affordable to lower-income households is a crucial task for local governments aiming to meet rising demand and inform policy priorities. However, enumerating the number of units with public housing, Project Based Section 8, and Low-Income Housing Tax Credit (LIHTC) assistance and identifying precisely where those units are located is deceptively challenging. Although federal datasets with that information are easily accessible, development and building location information may be unavailable or imprecise. Critically, identifying units that receive more than one form of assistance is hard, especially units with LIHTC. To address these challenges in New Jersey, the authors developed a largely automated and replicable process for precisely placing subsidized housing units into tax parcels. Doing so enables linking units across federal programs and with state and local data and to more accurately aggregate counts to integrate with decennial census and American Community Survey (ACS) data from the U.S. Census Bureau. Within New Jersey, the research team re-geocoded records in three datasets using two commercial geocoding services, assigned them confidence scores, designated records for manual handling, and then assigned them to parcels. Following those steps, they identified more than 15,000 units statewide with overlapping federal subsidies, which would lead to a 12-percent overcount of subsidized units in the state if the three datasets were used as given (and up to a 40-percent overcount in individual municipalities). By reusing and reconciling those datasets at the parcel level, researchers can more accurately enumerate rental units associated with different levels of subsidy depth and duration, a crucial task for identifying housing needs within and beyond the assisted rental stock.

Introduction

Housing researchers now have access to many detailed datasets about federally assisted rental housing, including for the most widely used housing subsidy programs such as public housing, Project-Based Section 8, and Low-Income Housing Tax Credits (LIHTC). Individually, each dataset offers rich, program-specific information. To view the entire landscape of assisted housing in a given location, including units with more than one form of subsidy, the datasets must be integrated, but that task presents multiple challenges.[1] Those challenges include differences in development names and addresses across datasets, available fields, and levels of aggregation, with some information available at the contract level and other information available at the building or development level. Furthermore, some rental units appear in multiple datasets because they receive project-based subsidies from more than one federal program, an administrative data challenge that has increased as LIHTC is used to renovate or redevelop public and other federally assisted housing.

As part of the New Jersey State of Affordable Rental Housing (NJSOARH) project, a large-scale research program on rental housing in New Jersey, the research team sought to understand the landscape of federally assisted and otherwise affordable rental housing across the entire state. They collected the most granular public data on federally assisted housing and assigned developments[2] to parcels, both to precisely locate them in space and to identify overlapping subsidies. That process made it possible to estimate the share of rental housing with select federal assistance; to link assisted housing developments to other administrative data, including census data; and to visualize them within communities.

Given the scale of analysis and recent improvements in automated approaches for cleaning and linking heterogeneous datasets, the authors developed a largely automated process for placing assisted developments and buildings in tax parcels using address information contained in the source data. They then used those parcel assignments to identify potentially layered subsidies that were subsequently manually verified. The central innovation of the approach is the use of two independent geocoding engines (from Esri and Google) to triangulate the accuracy of parcel matches for federally assisted housing developments and buildings, allowing the authors to focus manual verification work on records that did not return high-quality matches to the same parcel from both geocoding engines. Recent work has highlighted the promise of using widely available geocoding engines to improve the accuracy and precision of the geographic information contained in some assisted housing datasets (Wilson et al., 2023). Although that geographic information may be suitable for the program administration purposes those datasets were designed to serve, researchers needing to identify the precise location of projects on the ground often need to make adjustments to those data.

Although the use of a single geocoding engine has been demonstrated to yield improvements from the perspective of researchers seeking to place projects in parcels, even the results reported

[1] Some project-based units also receive project- or tenant-based federal housing vouchers. These data are only available aggregated to the census tract. Although the research team was aware that overlaps exist, they were unable to untangle them using these datasets, which limits some integration with American Community Survey (ACS) and census data rental unit counts.

[2] The Low-Income Tax Credit dataset uses *developments*, whereas the Project-Based Section 8 dataset uses *projects* to refer to the same thing. Throughout the article, the authors refer to *projects* or *developments* as appropriate.

as highly accurate exhibit some degree of uncertainty due to the different underlying data that different geocoders are drawing from and the internal logics of the geocoding engines (Prener and Fox, 2021). Triangulating the results of two independent commercial geocoders and their associated metadata gave the authors an ability to see where both services agreed on the location of a given street address, allowing them to cleanly and efficiently divide data into automated and manual workflows. Specifically, the authors identified which records were easily matched to parcels based on their street addresses and which records returned less confident results that required further location verification. They supplemented that output with data fields from the corresponding tax parcels, such as property ownership, which enabled them to design a process that yielded highly accurate automated results in more than 70 percent of the input records and allowed them to focus their attention on manual checking the remaining 30 percent. The rest of this article is organized as follows: the authors describe the historical context and existing work upon which their methods are built; they then provide an overview of the stages of their method and what they revealed; finally, they discuss the importance of doing this work and the implications of the overlapping subsidies they found in New Jersey's affordable housing stock.

Existing Data Correction and Integration Efforts

The authors are not the first to link federal housing datasets across subsidy programs. Their process draws inspiration from a set of efforts to integrate federally and other assisted housing datasets nationally and at the scale of specific states and cities, sometimes with different objectives and using different data sources. The research team also draws from recent related efforts to improve the accuracy and precision of geocoding processes and work that uses those improvements to locate assisted housing more accurately for research purposes. First, the authors discuss those existing assisted housing data integration efforts, which serve as the substantive stream of influence; second, they discuss the methodological work influencing their process.

HUD's Picture of Subsidized Households (PSH), produced by the Office of Policy Development and Research (PD&R), was first published in the 1990s, with annual releases most years since 1996. It was one of the first efforts to compile and make data from multiple HUD programs publicly available (Taghavi, 2008). This widely used data source makes information about major forms of project-based assistance available down to the level of individual projects, making it easy to combine them with other demographic datasets and create community maps. However, the PSH does not include LIHTC units, a critical source of support that is increasingly used to redevelop and preserve existing public and federally assisted housing. Simply adding LIHTC data to the PSH counts would produce an overcount of federally assisted units, because LIHTC is often used alongside other subsidy programs. The PSH also lacks data on individual building locations for scattered-site projects, which could lead to counting some housing units in the wrong place.

In addition, several related projects synthesize multiple forms of federal and other housing assistance, primarily to identify preservation needs for housing with time-delimited assistance. Those projects include the Assisted Housing Inventory (AHI) developed by the Shimberg Center at the University of Florida, the Subsidized Housing Inventory Program (SHIP) through the Furman Center at New York University, and the National Housing Preservation Database (NHPD) created by the Public and Affordable Housing Research Corporation (PAHRC) and the National Low Income Housing Coalition (NLIHC) (Reina and Williams, 2012). Both the AHI and SHIP integrate information about privately owned housing developments that receive government assistance. The NHPD also integrates information about public housing. Although the NHPD offers address-level information for New Jersey, the addresses are used as given in input datasets, not definitively placed in a geocoding and mapping workflow that can identify when different street addresses correspond to the same tax parcel. In the authors' case, organizing the data at the parcel level helped with incorporating LIHTC data supplied directly by the New Jersey Housing Finance and Mortgage Agency.

In addition to those projects directly oriented toward integrating assisted housing data, the authors draw on recent methodological work in precisely geocoding street addresses. Previous work has sought to correct the spatial coordinate information for individual assisted housing datasets to improve the quality of research involving distance-based measurements. Numerous areas of inquiry link the location of assisted housing developments to a range of localized outcomes, including property values (An et al., forthcoming; Deng, 2011) and eviction (Harrison et al., 2021; Lens et al., 2020). The precise placement of properties in space is needed to produce the best estimates of the association between assisted housing and these other contextual or distance-based processes.

Motivated by those concerns, Wilson et al. (2023) found that nearly 50 percent of the HUD-provided spatial coordinates for California LIHTC projects are outside the boundaries of the true tax parcel in which a project is located, compared with an accuracy level of 80 percent when geocoding the input address with Google's geocoding service. However, Wilson et al. (2023) also found that when Google results were incorrect, they were farther from the correct location than HUD's coordinates (derived from the freely available United States Postal Service geocoder), which, although often outside the boundary of the true parcel location, were typically near the true location.

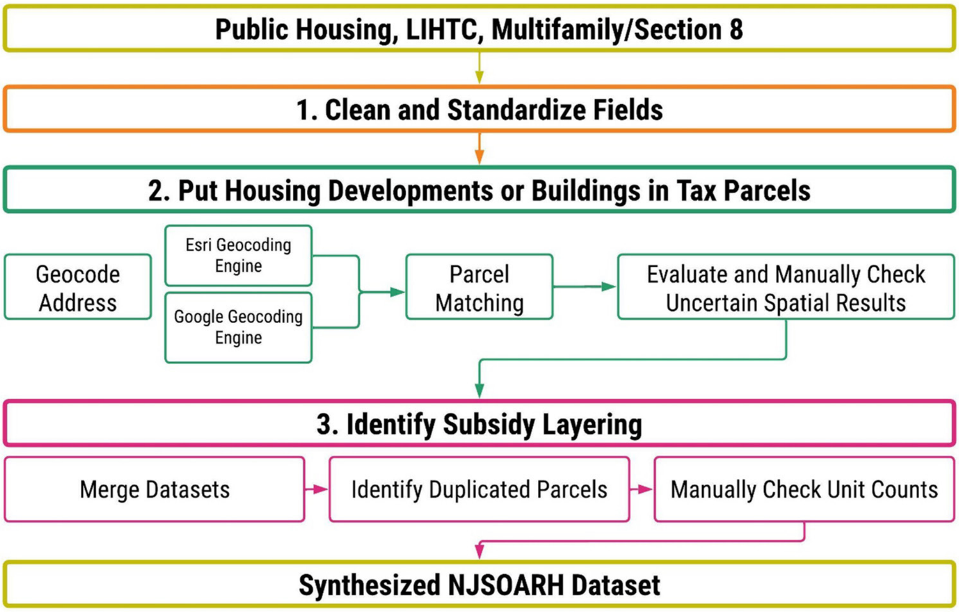
Beyond the specific application of improving the spatial accuracy of assisted housing, researchers have been developing more sophisticated methods for improving geocoding through the use of multiple services. Prener and Fox (2021) created a suite of tools for similar work in Saint Louis creating a custom composite geocoder—in their case, with access to authoritative local government positional data as a backstop. Researchers now have access to a suite of services for geocoding or converting address information to point coordinates. Although those services are generally quite accurate, the underlying reference data, normalization process, and process for matching inputs with coordinates are opaque (see Teske, 2014). The value of a good composite geocoding workflow is the triangulation between multiple sources. If geocoding results correspond across services (which have different underlying data and logics for converting address strings to spatial coordinates), they are significantly less likely to be false. Having surveyed existing efforts to integrate federally assisted housing data and the methodological literature on geocoding addresses, the authors crafted a largely automated process for integrating federally assisted housing at the parcel level using multiple geocoders in conjunction with tax parcel data. They placed assisted housing developments into parcels, similar to the SHIP project mentioned previously. Doing so enabled them to do four important things: (1) count units within the census geographies where they were actually located; (2) identify units that receive subsidies from more than one of the federal programs they are studying; (3) develop accurate narratives about housing security and neighborhood change; and (4) link together federal data, such as federal tenant-based vouchers, and local administrative data in the future.

Putting Federally Subsidized Housing in Parcels

To integrate our selected federally assisted housing datasets, the authors developed a three-step process to assign housing buildings or developments to parcels (see exhibit 1). First, they cleaned and standardized address fields from each input dataset. Second, they placed standardized addresses for buildings and projects in tax parcel polygons. Third, they used the results of the previous steps to identify developments that appeared in more than one of the assisted housing datasets and created a new composite unit field to avoid counting units more than once. Through those operations, the research team created a parcel-level dataset enabling them to integrate data across the major federal project-based subsidy programs and link it to local data sources.

Exhibit 1

Process of Integrating Federally Assisted Datasets


Public Housing, LIHTC, Multifamily/Section 8
1. Clean and Standardize Fields
2. Put Housing Developments or Buildings in Tax Parcels
Geocode Address
Esri Geocoding Engine
Google Geocoding Engine
Parcel Matching
Evaluate and Manually Check Uncertain Spatial Results
3. Identify Subsidy Layering
Merge Datasets
Identify Duplicated Parcels
Manually Check Unit Counts
Synthesized NJSOARH Dataset

LIHTC = Low-Income Housing Tax Credit. NJSOARH = New Jersey State of Affordable Rental Housing.
Source: Authors

The input datasets included a set of federal and state datasets from 2022 with information about housing developments and buildings. HUD's public housing buildings dataset records the location and tenant characteristics of public housing buildings, which may constitute one of several within a single development (HUD 2022c). HUD's multifamily assistance and Section 8 database records information about development location, size, contract origination, and contract expiration dates. It does not provide building-level information or service dates (HUD 2022b).[3] The New Jersey

[3] This dataset included a variety of programs, including the 811 Rental Assistance Demonstration (RAD), Other Section 8 (S8) New, Other S8 Rehab, Project Rental Assistance Contract (PRAC) 202/811, Pension Fund, S8 FmHA, S8 Loan Management, S8 and Section 202, S8 Preservation, S8 Property Disp., S8 RAD Mod Rehab Conversion, S8 Public Housing Conversion, S8 RAD RS/RAP Conv, S8 State Agency, and Section 202.

Housing Mortgage Finance Agency (NJHMFA) prepared a dataset of all LIHTC projects ever subsidized in New Jersey that recorded their location, number of units, and year placed in service (NJHMFA, 2022) but did not identify the location of individual buildings.[4] The research team used publicly available integrated New Jersey tax assessor and parcel location data to triangulate results using property location and ownership information (New Jersey Office of GIS, 2022).

In the next section, the authors walk through their iterative, semi-automated process to describe what they did, how they did it, and what happened as a result. They discuss the challenges and how they addressed them, where possible.

Cleaning and Standardizing Fields

The first step was to clean and standardize the address fields in each of the input datasets. The research team combined addresses into a single field and employed string substitution and related forms of text cleaning to clean and standardize, for instance, by changing all road type abbreviations to their full name (e.g., "St" to "Street"). The authors spelled out those abbreviations to improve geocoding and string matching.[5] In the case of fields containing an address range (e.g., 10–20 Chestnut St.), they split the record into two separate addresses (one for each end of the range) to geocode each and determine whether they were placed in the same parcel further in the process. Preparing addresses for geocoding revealed two main challenges to pinpointing assisted housing locations: first, that they could be scattered site (e.g., address range and multiple addresses within the same input field); and second, that addresses may refer to administrative office mailing addresses rather than actual residence locations (e.g., a P.O. Box). They flagged records with incomplete, unusual (duplicate or multiple addresses and those with unusual characters), or missing addresses.[6] Across all three datasets were 278 of those locations, or 6 percent of all records.

Putting Housing Developments or Buildings in Tax Parcels

With a set of cleaned and standardized addresses, in second step, the authors employed a three-step process to assign addresses from the input data to parcels (exhibit 2). First, they geocoded

[4] HUD also provides an LIHTC building-level dataset that can help locate individual buildings in scattered-site developments. It does not include information about the unit count in each building, so the research team did not use the dataset. The dataset provided by NJHMFA included all LIHTC projects ever contracted in New Jersey. One hundred ten of these projects were put in service more than 30 years ago, and the authors could not match HUD contracts to them (HUD 2022a). Those 110 developments account for 2,183 units, which are not part of the unit tabulation. Twenty-three LIHTC projects in the NJHMFA data put in service less than 30 years ago could not be matched to a contract; the authors determined that those projects likely exist, and included them. Across those 23 developments were 2,206 units, of which 1,352 had layered subsidies (LIHTC and public housing or multifamily subsidies). Accounting for that fact, including those developments may have led to an LIHTC unit overcount of up to 854 units.

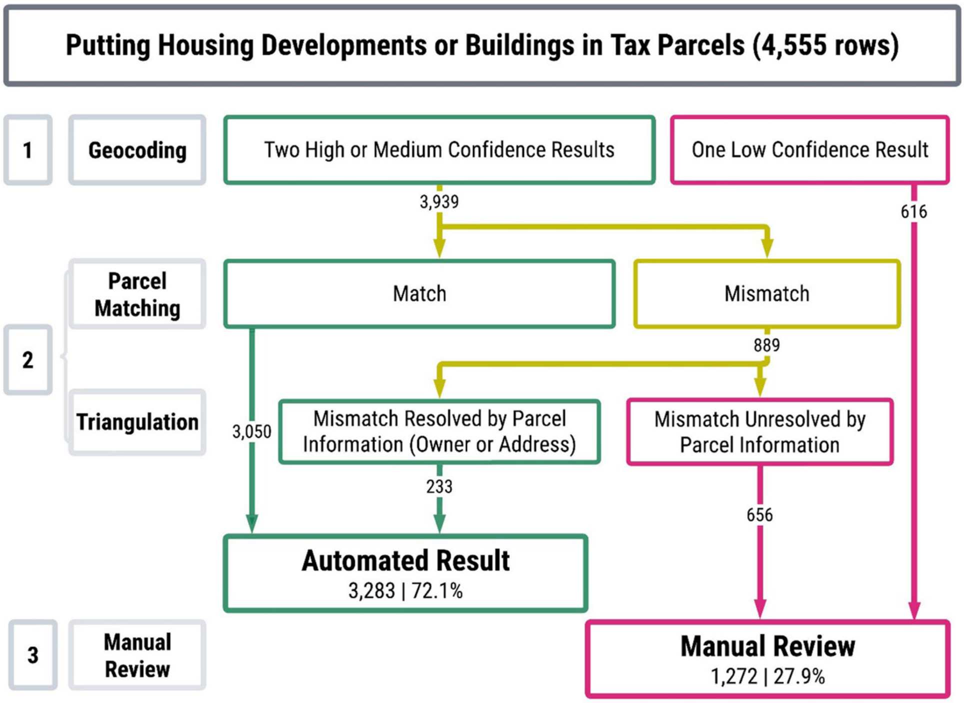
[5] The authors created their own standardization process using Python code based on common address inconsistencies in New Jersey. Standardization included the following steps: (1) capitalize the entire string; (2) fill ZIP Codes with leading zeros so that they are all 5 digits long (New Jersey ZIP Codes start with 0 and are frequently imported as integers); (3) standardize abbreviations for Avenue, Boulevard, Circle, Drive, Highway, Parkway, Place, Road, Street, Terrace, Lane, Court, 1st–10th, North, South, East, and West; and (4) remove unit or apartment numbers from addresses.

[6] These challenging addresses were more prevalent in the multifamily/Section 8 dataset than the LIHTC data. In the multifamily/Section 8 dataset, multiple addresses were often found on one line in varying ways (e.g., separated by an & symbol or a comma). This variance was not present in the LIHTC data, possibly because they were curated and cleaned by one entity: the HMFA.

addresses with two independent geocoding engines to assign point locations to developments or buildings. Second, they matched the geocode results to parcel polygons and triangulated with the parcel address and owner. Third, they manually checked the results with uncertain outcomes.

Exhibit 2

Putting Housing Developments or Buildings in Tax Parcels


Putting Housing Developments or Buildings in Tax Parcels (4,555 rows)
1
Geocoding
Two High or Medium Confidence Results
One Low Confidence Result
3,939
616
Parcel Matching
2
Triangulation
Match
Mismatch
889
3,050
Mismatch Resolved by Parcel Information (Owner or Address)
Mismatch Unresolved by Parcel Information
233
656
Automated Result
3,283 | 72.1%
3
Manual Review
Manual Review
1,272 | 27.9%

Source: Authors

Geocoding

The first step in the process to automate the placement of buildings and developments in parcels was to geocode the input data using cleaned and standardized address fields (see exhibit 2, step 1). The research team ran the cleaned addresses through Google and Esri geocoding services using each company's application programming interfaces (APIs) using Python (Esri, n.d.; Google Maps Platform, n.d.). The geocoding services provided metadata about each result, including the result address, where found, and the match type (e.g., rooftop, street, intersection, or centroid of a larger geographic area). Based on the match type and the similarity between the match address and input address,[7] the authors assigned geocode results as having high, medium, or low confidence (see exhibit 3). If either geocoder returned a low confidence result, the authors manually checked the record. This category accounted for 616 records, or 13.5 percent of the total input addresses. Those 616 records included

[7] Esri provided a score from 0 to 100 grading this correspondence. The authors calculated their own score for the Google results using the FuzzyWuzzy Python package (Cohen, 2020) to measure how closely the input address corresponded with the returned address field and its parts.

246 records from the public housing buildings dataset that lacked a street address and only specified the municipality; the remaining 370 low confidence results displayed no obvious pattern. The remaining 3,939 records moved to the next automated step, parcel matching.

Exhibit 3

Geocode Confidence Metrics

Confidence	Description	Esri	Google
High	A rooftop/building-level result that the authors believe to be trustworthy and think will match to the correct parcel without issues	A record that was a sub-address (an individual suite or unit in a building) *OR* A record that was a point address (building or rooftop), and the match score is at least 97.7[a]	Every record with a matching house number and ZIP Code and a full address match greater than or equal to 90 *AND* In the "geometric center" (center of a feature) and "premise" (a named location, such as a building) result *or* "rooftop" (result is accurate to the street address) *or* "range interpolated" (approximate place on street segment)
Medium	A rooftop-level result with slightly less certainty or an interpolated street result (i.e., road centerline)	All other "point address" results not labeled high confidence *OR* "Street address" with a match score of at least 99.75 *OR* "Street address ext" (match out of range based on house number) and located in New Jersey[b]	No criteria of "low" or "high" confidence geocodes met
Low	Result that matched the wrong geography (town, ZCTA, street, or intersection), no result, or result outside New Jersey	Multiple addresses (usually identified by "/" or "&") *OR* No criteria of "high" and "medium" confidence met	Result that was "approximate," *or* a "geometric center" and was not a "premise"-level result *OR* House numbers of source data and output that do not match *OR* Street address match score less than 80

ZCTA = ZIP Code Tabulation Area.

[a] *The authors selected all match score cutoffs after sampling the output and choosing values that led to no false positives (defined as geocodes flagged as correct that were in the incorrect parcel for the given input address). The suitability of those thresholds for their purposes is confirmed by the data presented in table 4.*

[b] *Although "street address ext" is less precise than "street address," the research team's explorations of the output showed that it was very uncommon for an address to be labeled that way. The few "street address ext" results in New Jersey were similar to "street address" results and fit the "medium" confidence category.*

Source: Authors

Parcel Matching

Next, the authors assigned the 3,939 high- or medium-quality geocode results to parcel polygon locations (see exhibit 2, step 2). This process added six new fields (three for each geocoder) for each record based on the results of spatially joining the Esri and Google geocodes to parcels: the parcel identifiers for each set of geocoded points, the distance between the returned point and the nearest parcel,[8] and a flag for whether the result matched to multiple parcels for each service.[9]

For 3,050 of the 3,939 records that satisfied the earlier requirement of two high- or medium-confidence results, both geocoding services returned points that fell within the same parcel. The authors accepted those parcel results without additional manual steps, leaving 889 records that matched to different parcels. For those 889 records, the research team triangulated between parcel information and the input data to determine whether either of the parcels was likely the correct match. Using the FuzzyWuzzy string matching package for Python, they evaluated whether the parcel address in New Jersey's MOD-IV parcel database corresponded to the development or building address. For the public housing data, they also checked whether a housing authority owned the parcel by looking at the owner's name and property class fields. That process resolved 233 of the 889 high- or medium-quality geocode records that matched to different parcels. Combining those 233 records with the 3,050 accepted previously, the authors determined that 72 percent of their total records did not require further attention. They manually checked the rest as described in the following section.

Manual Process

The research team's process for automating parcel assignment flagged records with missing or un-geocodable address information and low-confidence geocoding results, producing a subset of 1,272 records for manual processing. To manually check results, the authors drew on additional resources, including the New Jersey Department of Community Affairs affordable housing dataset (NJ DCA 2022); Rowan University School of Earth and Environment's interactive map of New Jersey tax parcels (2023); Google Street View imagery for house numbers, development signs, and correspondence between housing appearance and public housing or developer websites; and other online sources (housing authorities, developers, investors, news articles, property, and planning documents and websites). In many cases, Google Street View enabled the research team to find the parcel. For the remaining records, the authors used many, if not all, of the sources to triangulate and place units in the correct parcel. When they located a property and assigned it to a parcel(s), they made notes about how they made their decisions and recorded the Internet links and documents that aided their decisionmaking. In fewer than 10 cases, the authors contacted someone with local knowledge, such as a public housing authority staff member. Ultimately, they placed all but 15 projects in parcels. They assigned 5 of the remaining properties to a block or mega-parcel (set of adjacent parcels) and 11 to a municipality.

Identifying Subsidy Layering

Having placed buildings and developments into parcels, the research team used those assignments to identify parcels that appeared in more than one of the source datasets. However, shared

[8] For most records this value was zero because the geocode point location was within the parcel.

[9] Some point locations matched to multiple parcels. This was most common in condominium developments where multiple parcels are layered on top of each other. The research team flagged those cases as multiple parcel matches, dropped the duplicates, and determined the primary parcel for those developments, an underlying parcel polygon encompassing the individual condominium units, manually.

parcels are not an unambiguous indicator of shared subsidies. Multiple projects can be located on individual parcels; furthermore, individual assisted buildings or developments can include different sets of units subsidized by nonoverlapping government programs. Whereas some projects contain units carrying multiple forms of assistance, others assist different sets of units with distinct forms of assistance, so no overlap occurs at the unit level.

The authors checked records with shared parcels using the resources from the manual parcel matching process described previously. However, this time, first they ensured that two separate developments with different subsidies were not on the same parcel; and second, they identified where subsidies applied to the same units. The authors triangulated that information in the assisted and total unit fields in each dataset, along with the manual sources, as part of a qualitative process to make a determination. Using that process, they estimated that 12 percent, or 15,529, of the total units across the datasets appeared in more than one dataset (see exhibit 4). Many of the overlaps are the result of LIHTC being used to renovate public housing or Multifamily/Section 8 developments. For example, LIHTC was used alongside federal multifamily assistance in 13,549 units. If the authors had not carried out the process previously described to reconcile the three input datasets and simply aggregated the unit counts across all three without looking for subsidy overlaps, they would have overcounted subsidized units in the state by 12 percent.

Exhibit 4

Total Units by Federal Program Before and After Identifying Layered Subsidies

Unit Count	Units in Source Data[a]	NJSOARH Output Data	Difference[b]
Total Subsidized	144,411	128,882	–15,529 (12.0%)
Only LIHTC Affordable	61,269	45,740	–15,529 (34.0%)
Only Public Housing	29,651	27,671	–1,980 (7.2%)
Only Multifamily/Section 8 (MF)	53,491	39,942	–13,549 (33.9%)
MF & LIHTC Affordable	0	13,549	+13,549 (100%)
Public Housing & LIHTC Affordable	0	1,980	+1,980 (100%)

LIHTC = Low-Income Housing Tax Credit. MF = multifamily. NJSOARH = New Jersey State of Affordable Rental Housing.

[a] In the process of cleaning the data, the research team sometimes modified unit counts for reasons other than subsidy overlap. For example, they found that 502 units of public housing had been demolished and 2,366 units of LIHTC housing were more than 30 years old and missing a contract. Those units are excluded from the totals.

[b] These values and percentages are the number of units that would have been overcounted, meaning the difference between the two counts divided by the cleaned data count.

Source: Authors

Through this process, the authors noticed that the extent of housing subsidy overlap varies widely by geography, in large part because of the clustering of multifamily developments in space. Thus, the impact of overcounting units due to subsidy layering depends on the prevalence of LIHTC in different communities and, in some communities, would affect a far greater proportion of units than the statewide average of 12 percent. Without this process, the research team would have double-counted 4,624 units of federally subsidized housing in Newark, or 22 percent of the city's total subsidized stock (see exhibit 5). An overcount of 25 percent or more would have occurred in Atlantic City, East Orange, Elizabeth, Pennsauken, and Trenton. Many of those municipalities have large numbers of older assisted housing that is being redeveloped using LIHTC. Those cities are also some of the most populous in the state, with acute housing needs and entrenched social disparities; working from accurate understandings of the existing affordable housing landscape in these places is crucial.

Exhibit 5

Top 10 New Jersey Municipalities for Most Overcounted Units[10]

Municipality (County)	Units in Source Data	NJSOARH Output Data	Difference[a]
Newark (Essex)	25,686	21,062	4,624 (22.0%)
Trenton (Mercer)	6,888	5,481	1,407 (25.7%)
East Orange (Essex)	4,567	3,357	1,210 (36.0%)
Atlantic City (Atlantic)	5,587	4,408	1,179 (26.7%)
Jersey City (Hudson)	10,319	9,319	1,000 (10.7%)
Elizabeth (Union)	3,749	2,770	979 (35.3%)
Camden (Camden)	7,276	6,375	901 (14.1%)
Paterson (Passaic)	5,166	4,284	882 (20.6%)
Orange (Essex)	2,961	2,429	532 (21.9%)
Pennsauken (Camden)	1,065	762	303 (39.8%)

NJSOARH = New Jersey State of Affordable Rental Housing.

[a] These values and percentages are the number of units that would have been overcounted, meaning the difference between the two counts divided by the cleaned data count.

Source: Authors

Process Evaluation

At the end of the parcel assignment and deduplication process, the research team assessed the results of their workflow to better understand the impact of key components of the methodology and thresholds that were set along the way. They explored two questions. First, was the automated process effective at placing buildings or developments in the correct parcels? Second, were the criteria too conservative, and could the authors have automated more of the process without sacrificing accuracy at the parcel level?

Results of Manual Validation of Select Automated Results

In this section, the authors evaluate whether the automated process was effective at placing buildings or projects in the correct parcels. They conducted this evaluation by comparing the parcel match results from their automatic process to the final results that they later determined manually. This sample included the 898 records (29 percent of automated results) that were accepted automatically but also manually checked, either because they were flagged for possible subsidy overlap or they were located in one of seven communities in which the authors were doing more in-depth research.

Nearly all of the automated results (98 percent) either correctly matched to only one parcel (86 percent) or matched to one of the correct parcels in a multiparcel development (12 percent, see exhibit 6. For example, the LIHTC and multifamily datasets are aggregated at the development

[10] Future work could attempt to better understand the contextual and demographic drivers of redevelopment in those communities.

level and typically provide only a single address per development, even in cases in which developments contain multiple buildings that straddle several tax parcels. Because this sample includes all records likely having layered subsidies, it contains a large share of developments likely sitting on multiple parcels, typically because of redevelopment.

Even when the result of the automated process was not in the right parcel (about 2 percent of the cases), it was for reasons largely out of the research team's control that no automated process could hope to address. For about one-half of those cases, the input address provided in the dataset was not the actual location of the housing, and for the second half, the result was clearly wrong (e.g., the address points to the middle of a field or a shopping mall), but the authors could not ascertain where the housing was even after extensive manual research. In those cases, the address may have been inputted incorrectly by the initial compiler of the administrative datasets. That result underscores the robustness of our conservative process for automating results. Based on those results, the thresholds for accepting results automatically were sufficiently stringent.

Exhibit 6

Results of Manual Validation of Select Automated Results

Result	%	Count
Automated and manual checks correspond	98.0	880
– Matched to a unique parcel	86.3	775
– Matched to one parcel in a multiparcel development	11.7	105
Automated and manual checks do not correspond	2.0	18
– Input address from administrative dataset is incorrect	0.9	8
– Result is wrong; housing location still unclear	1.1	10
Total	**100**	**898**

Source: Authors

Results of Manual Validation of Low-Quality Matches

In this section, the authors evaluate whether they could have accepted more geocoding results programmatically and thus manually checked fewer records. They conducted this evaluation using a sample of 994 records that had valid, geocodable addresses[11] but were manually checked because the research team could not ascertain their location through geocoding and parcel matching. Using that sample, the authors examined the relationship between the different possible sources of uncertainty that led to manual review (see exhibit 7, which shows low-quality matches from one or both geocoders or conflicting parcel matches between geocoders) and the ultimate accuracy of each of the two geocoders after manual validation. That examination enabled the authors to see whether, for instance, an address that returned a low-quality Esri geocoding match but a medium- or high-quality Google match could have been assigned to the Google result with confidence. If all or nearly all records would have been assigned to the parcel that the authors ultimately accepted on the basis of a single medium- or high-quality match, then the process of requiring parcel matches from both geocoders or independent verification from tax parcel data would have been unnecessarily restrictive. However, if the research team loosened the process to allow a

[11] Of the total records, 278 had missing or multiple addresses.

single medium- or high-quality geocoding match to determine the correct parcel, and a nontrivial number of those records would have led them to accept inaccurate parcel assignment, that outcome would have contravened the objective of deliberately minimizing false positives, ensuring reliable results and an efficient use of the team's time for manually checking disputed results.

Exhibit 7

Results of Manual Validation of Low-Quality Matches

Reason for Manual Check	Records by Reason		Correct Results After Manual Check					
			Google		Esri		Neither	
	#	%	#	%	#	%	#	%
Both Low Quality	47	4.7	14	29.8	11	23.4	29	**61.7**
Esri Low Quality	143	14.4	112	**78.3**	41	28.7	28	19.6
Google Low Quality	175	17.6	85	48.6	123	**70.3**	39	22.3
Parcel Mismatch	629	63.3	368	**58.5**	137	21.8	124	19.7
Total	994	100	579	**58.2**	312	31.4	220	22.1

Note: The bolded value in each row shows the highest percentage of ultimate accuracy for each type of low-quality geocode: Google, Esri, or neither.
Source: Authors

The results of this manual validation align with the goals and expectations the research team had when setting up their process. Looking across the possible reasons for manually checking and the parcel-level accuracy of each geocoder, no clear patterns emerge that might have been used as rules to modify the automated process. In cases in which one geocoder has low-quality results, the other geocoder is more likely to be correct but not at a high enough rate to simply accept them automatically: 78 percent of Google results were correct when only Esri had a low-quality match, and 70 percent of Esri results were correct when only Google had a low-quality match. The majority of the results that the authors manually checked were checked because both geocodes matched to different parcels despite both geocode results exhibiting high or medium quality (629, or 63 percent of the cases). In those cases, the Google geocode corresponded with the ultimate decision more than twice as often as the Esri one (368, or 59 percent, compared with 137, or 22 percent). Because parcel mismatches account for the majority of the records in this sample, Google has the overall edge in accuracy. Across all reasons for manually checking, the Google geocode result was in the parcel that the authors determined to be true 58 percent of the time (compared with Esri 31 percent of the time). But the 532 records (42 percent) that the Google geocoder incorrectly located are dispersed across all the reasons for manually checking in exhibit 7 and show no clear pattern that would have allowed the authors to identify them and handle more records procedurally without resulting in false positives.

The authors explored one other potential avenue for additional efficiency gains in the process: low-confidence results that still led to parcel matches. In 147 cases, the Google and Esri results matched to the same parcel, but at least one of the two geocoders had low match quality; 117 of those results ultimately corresponded with the manual decision (79.6 percent). However, given the goal of having parcel-level certainty whenever possible, having no way to know which 20 percent of the results were incorrect was not acceptable. The research team concluded that they could not wring additional gains out of the automated portion of the workflow.

Discussion

HUD publishes a wealth of easily accessible and richly detailed administrative housing datasets. Improving capacity to link those datasets to each other and to local data holds much promise for performing increasingly sophisticated housing analysis to aid public policy decisions. The effort outlined in this article sought to link federal datasets at the parcel level both because the authors needed to identify units with layered federal subsidies and because they wanted to situate those data more precisely in their community context—for instance, by placing properties in parcel maps, which better reveals their community presence. This activity is particularly important with the presence of processes such as gentrification and redevelopment, which often occur block by block rather than within neat categories such as ZIP Codes or census tracts. To achieve that goal, the authors developed a largely automated process that linked federal housing datasets and the state LIHTC dataset. The process was efficient and effective. It gave them high confidence in their automated results and focused their energy on necessary manual checks.

As large administrative datasets and the tools to analyze them are increasingly available to wider audiences, not taking care to understand the limitations of those datasets may introduce biased results, generating considerable error at scale. Although the results presented here have limitations, the authors developed a robust process that employed a conservative approach to accepting geocoding matches that gave them high confidence in the automated matches. Rather than accepting results for which either of the geocoding results were high when the other was low, the authors accepted only results for which both geocodes were high or medium quality to severely limit false positives.

This methodological conservatism increased the amount of manual work, but the manual work helped the research team to better understand both the data and the choices communities are making to renovate, preserve, and create new affordable housing with the subsidies. In fact, for the authors purposes, the manual work was invaluable for reasons beyond confirming or fixing the results of the automated workflow. It revealed the extent of redevelopment and renovation of the older stock of public and federally assisted housing. It also highlighted that many LIHTC and federally assisted housing projects include multiple buildings that are not revealed through address point matching. And by manually checking data, the authors were able to see the extensive redevelopment of existing public and federally assisted housing occurring throughout New Jersey. Although many people have a general understanding of redevelopment initiatives—including Housing Opportunities for People Everywhere (HOPE) VI, Choice Neighborhoods, and the Rental Assistance Demonstration (RAD)—seeing how and where they are implemented is important to grasp the implications. The manual process gave the authors insight into those trends.

Limitations of the Present Work

Although this process enabled the research team to do much of what they intended, there were several limitations. First, only the public housing dataset included building-level data; the other two sources provided development-level information. Thus, the authors were not able to place all buildings in the right location; they may not have correctly placed some sprawling and scattered-site developments. In addition, because voucher information is publicly available only at the tract level, the authors did not explore project- or tenant-based housing choice vouchers in this project, which they expect to overlap with many project-based units. Finally, the ability to replicate this process is dependent on access to two private geocoding services, which may charge rates for the use of their service that some data users would find prohibitive and which may cease to be supported in the future if the companies who created and maintain them change their business objectives, unlike public alternatives such as the United States Postal Service geocoder used by HUD in its public housing buildings dataset.

Conclusion

By reusing and reconciling subsidized housing datasets, researchers can more accurately enumerate rental units associated with particular levels of subsidy depth and duration, which are crucial for identifying housing needs within and beyond the assisted rental stock. For the authors' purposes, understanding what exists in subsidized housing projects and how the subsidies are layered to make housing affordable allowed them to bring those data into conversation with other datasets and to more accurately understand the landscape of affordable rental housing in New Jersey. The research team's aim is that other researchers and practitioners can apply the methods and lessons learned in their process to better understand the federally subsidized housing landscape in other parts of the country and to consider some of the tradeoffs between efficiency and accuracy inherent in using off-the-shelf geospatial information in administrative datasets for research purposes.

Acknowledgments

Support for this article was provided by the Robert Wood Johnson Foundation with the assistance of the New Jersey Housing Mortgage Finance Agency, which provided the LIHTC dataset. The views expressed here do not necessarily reflect the views of the Foundation.

Authors

Shiloh Deitz is a postdoctoral associate in the Edward J. Bloustein School of Planning and Public Policy at Rutgers University. Will B. Payne is an assistant professor in the Edward J. Bloustein School of Planning and Public Policy at Rutgers University. Eric Seymour is an assistant professor in the Edward J. Bloustein School of Planning and Public Policy at Rutgers University. Kathe Newman is a professor in the Edward J. Bloustein School of Planning and Public Policy and director of the Ralph W. Voorhees Center for Civic Engagement at Rutgers University. Lauren Nolan is a Ph.D. candidate in the Edward J. Bloustein School of Planning and Public Policy at Rutgers University.

Appendix A: Findings From Research in Seven Communities

The research team is also conducting in-depth research in seven communities and have manually checked and mapped every federally assisted development with more care. Through this research, they have a further understanding of the possible limitations of our workflow. Following are their findings from those seven communities.

- In **Asbury Park**, one development—Asbury Park Village—was correctly placed in the wrong address. The public housing buildings data put it in Trenton, but it was an error in input data, and no amount of automation could have identified it. The research team also found that one development (Vita Gardens) covered two parcels rather than one.
- In **Millville**, Maurice View Plaza is a development, but the name was also used for several scattered sites. That fact poses no problem to locating the units in parcels because the data were at the building level. Holly Berry Court was on a nearby parcel but not the location of the housing units (as verified on Google Street View). The address given, 1100 Holly Berry Lane, is the location where the geocoders initially placed the units, but it was not the correct place. It is also a match in the parcel data. This discrepancy, like Asbury Park Village, is a case of a correct geocoder on an incorrect address.
- In **Montclair**, one sprawling (multiparcel) development had already been identified in the research team's previous processes.
- In **Passaic**, the research team found that one development (Chestnut Homes) covered two parcels rather than one.
- In **Phillipsburg**, what was called Heckman House was actually a number of developments with different names. This fact had no effect on placing the units in parcels because they were public housing and building-level data, but it would cause problems if someone wanted to identify overlap on the basis of the development name.
- In **Salem**, the research team thinks that **Salem Historic Homes** is actually a scattered-site project. The geocoding results are accurate based on the address provided.
- In **West New York**, one development (Horizon Heights) seems to be both in West New York and Union City on adjacent parcels; the research team had identified only the large parcel in West New York.

None of those issues affected the statewide overlap count. The sites were not examples of subsidy overlap, or the overlap was identified despite the inaccuracies.

References

An, Brian, Andrew Jakabovics, Jing Liu, Anthony W. Orlando, Seva Rodnyansky, Richard Voith, Sean Zielenbach, and Raphael W. Bostic. Forthcoming. "Factors Affecting Spillover Impacts of LIHTC Developments: An Analysis of Los Angeles," *Cityscape*.

Cohen, Adam. 2020. "FuzzyWuzzy: Fuzzy String Matching in Python." Python. https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/.

Deng, Lan. 2011. “The External Neighborhood Effects of Low-Income Housing Tax Credit Projects Built by Three Sectors,” *Journal of Urban Affairs* 33 (2): 143–66.

Esri. n.d. “Geocoding and Geosearch.” ArcGIS Online Help. https://doc.arcgis.com/en/arcgis-online/reference/geocode.htm.

Google Maps Platform. n.d. “Geocoding Request and Response | Geocoding API.” Google for Developers. https://developers.google.com/maps/documentation/geocoding/requests-geocoding.

Harrison, Austin, Dan Immergluck, Jeff Ernsthausen, and Stephanie Earl. 2021. “Housing Stability, Evictions, and Subsidized Rental Properties: Evidence from Metro Atlanta, Georgia,” *Housing Policy Debate* 31 (3–5): 411–24.

Lens, Michael C., Kyle Nelson, Ashley Gromis, and Yiwen Kuai. 2020. “The Neighborhood Context of Eviction in Southern California,” *City & Community* 19 (4): 912–32.

New Jersey Housing Mortgage Finance Agency (NJHMFA). 2022. Low Income Tax Credit Properties. Data file.

New Jersey Office of GIS. 2022. “Parcels and MOD-IV Composite of NJ (download).” Shapefile. New Jersey Geographic Information Network (NJGIN) Open Data. https://njogis-newjersey.opendata.arcgis.com/documents/406cf6860390467d9f328ed19daa359d/about.

Prener, Christopher G., and Branson Fox. 2021. “Creating Open Source Composite Geocoders: Pitfalls and Opportunities,” *Transactions in GIS* 25 (4): 1868–87. https://doi.org/10.1111/tgis.12741.

Reina, Vincent, and Michael Williams. 2012. “The Importance of Using Layered Data to Analyze Housing: The Case of the Subsidized Housing Information Project,” *Cityscape* 14 (1): 215–22.

Rowan University School of Earth and Environment. 2023. “NJ Map.” Parcel Explorer. https://www.njmap2.com/parcels/.

State of New Jersey Department of Community Affairs. 2022 (NJ DCA). List of Affordable Developments by County. Data file. https://www.nj.gov/dca/codes/publications/developments.shtml.

Taghavi, Lydia B. 2008. “HUD-Assisted Housing 101: Using ‘A Picture of Subsidized Households: 2000’,” *Cityscape* 10 (1): 211–20.

Teske, Daniel. 2014. “Geocoder Accuracy Ranking.” In *Process Design for Natural Scientists: An Agile Model-Driven Approach*, edited by Anna-Lena Lamprecht and Tiziana Margaria, 161–74. Communications in Computer and Information Science (CCIS), vol. 500. Berlin, Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-662-45006-2_13.

United States Department of Housing and Urban Development (HUD). 2022a. Low-Income Housing Tax Credit (LIHTC): Property Level Data. Data file. https://www.huduser.gov/portal/datasets/lihtc/property.html#data.

———. 2022b. Multifamily Assistance and Section 8 Database. Data file. https://www.hud.gov/program_offices/housing/mfh/exp/mfhdiscl.

———. 2022c. Public Housing Buildings. Data file. https://hudgis-hud.opendata.arcgis.com/datasets/52a6a3a2ef1e4489837f97dcedaf8e27_0/explore?location=35.877966%2C-115.070600%2C4.47.

Wilson, Nicole E., Michael Hankinson, Asya Magazinnik, and Melissa Sands. 2023. "Inaccuracies in Low Income Housing Geocodes: When and Why They Matter," *Urban Affairs Review*, March. https://doi.org/10.1177/10780874231165767.