Connecting Address and Property Data To Evaluate Housing-Related Policy

Alyssa J. Sylvaria
The Providence Plan

Jessica Cigna
HousingWorks RI

Rebecca Lee
The Providence Plan

Abstract

Housing conditions can vary greatly from one property to the next, but housing characteristics often are measured at different geographic units because of data limitations. This article discusses the process of connecting address-level datasets to create meaningful analyses at the property level in the absence of a comprehensive address-to-parcel crosswalk. To demonstrate this process, the authors describe linking child lead screening, lead property compliance, foreclosure, and tax assessors' property records for a U.S. Department of Housing and Urban Development-funded Lead Technical Study in four Rhode Island core cities. Using the linked data analysis, robust property-level findings can lead to an effective evaluation of policies that affect properties, particularly for urban communities with high proportions of multifamily housing.

Introduction

Connecting existing datasets to conduct policy evaluation is a smart way to make the best use of available resources. Administrative datasets across multiple domains contain addresses and can be linked to gain insight regarding housing conditions and policy. In some situations, however, researchers prefer data about entire properties to address-level data when describing housing issues. Many multiunit residential properties have more than one address and, when researchers

try to collect information about all residential units within properties, address listings are often insufficient. This concern is particularly evident for analysis in urban communities, where a high proportion of the housing stock contains more than one unit.

Robust statewide data systems ideally would exist and would enable researchers and city administrators to easily link address-specific data to property-level data. In Rhode Island, as we suspect in many other states, that ideal is not yet the reality. Therefore, extensive preparatory work was completed to conduct a property-level analysis of childhood lead exposure, lead compliance certificates, and foreclosures in four Rhode Island cities. In this article, we discuss the process of connecting a variety of separate address-level datasets with unique variables and coding systems. We provide background information that defines the study's purpose and describe how we created a master lookup table, matched our datasets to it, and analyzed the data. We also share the lessons we learned from this effort.

Context

The 2005 Rhode Island Lead Hazard Mitigation Act requires owners of nonowner-occupied properties built before 1978 (when residential lead-based paint was banned in the country) to comply with a series of actions aimed at reducing lead exposure. These requirements include attending a lead hazard awareness class, inspecting rental properties, providing tenants information about lead hazards and a copy of the inspection report, responding to tenants' concerns about any lead hazards, fixing lead hazards, and using lead-safe work practices when performing any maintenance. After the owners comply with the requirements, they receive a Certificate of Conformance, which needs to be kept current.[1]

For this U.S. Department of Housing and Urban Development-funded study, we sought to evaluate outcomes associated with the Rhode Island law. We identified the number of residential properties that were in compliance with the law, whether lead-exposed children were more likely to reside in noncompliant homes, and whether foreclosure had an impact on lead exposure and compliance. The analysis centers on lead exposure and other risks that are likely to pervade entire structures, and the unit of analysis was at the property level rather than address level. In addition, the law has implications for property owners, which bolsters the rationale for a property-based approach. We studied four cities in Rhode Island that have high risks of substandard housing concerns and lead-exposed children: Central Falls, Pawtucket, Providence, and Woonsocket (Healthy Housing Collaborative, 2012). The first results of our analysis, in which we compared blood lead levels of children with a property owner's compliance with and exemption from the Lead Hazard Mitigation Act and which included a summary of the methods that we used, were recently published and received notable press coverage in local media (Rogers et al., 2014).

Preparing the Data

In the four cities we studied, most residential properties have two or more units (U.S. Census Bureau, 2014), and many of those properties have more than one street address. Thus, to obtain

[1] State of Rhode Island General Assembly. 2003. Chapter 23-24.6 Lead Poisoning Prevention Act.

accurate property-level counts, we first determined which distinct addresses were part of the same property and then, based on the knowledge of all addresses for each property, aggregated all the address data from various sources to the property level. For example, if three children were exposed to lead at one address and two at another, but those addresses were both part of the same multifamily property, that property housed five lead-exposed children.

Creating a Master Lookup Table

To overcome the obstacles associated with having multiple addresses per property, we created a crosswalk tool called the master lookup table, or MLT, which links each address to its property identifier code as well as other basic descriptive data about the property. Our method for creating the MLT differed between Providence and the other three cities. For Providence, the largest city in the state, the MLT was developed to be a more robust resource (as described further in the Providence MLT Online Tool section that follows). Two key pieces that enabled the work for Providence were (1) the availability of an up-to-date parcel shapefile, which identifies the plat and lot numbers for properties and the size and shape of the parcel of land, and (2) a cooperative relationship with the city tax assessor's office. Having staff members with Geographic Information System (GIS) experience and interns to assist in the time-consuming portions of creating a reliable MLT were also integral to the process.

Another indispensable resource for this work was a dataset of all addresses for occupied and unoccupied structures in Rhode Island, which was initially developed for the emergency 911 telephone and response system. These addresses are available through the state's GIS data website as a point shapefile (Rhode Island Geographic Information System, 2014). In ArcGIS software, we were able to join the Providence parcel data shapefile with the emergency 911 addresses to create a citywide map of all properties. To increase accuracy, we consulted paper maps from the tax assessor's office and, on occasion, staff members physically visited properties to verify the address-to-property crosswalk. Today, where available, Internet-based streetview maps can serve to validate addresses or other property information by zooming in on the address in question. Parcels with more than one street address can be easily identified through this combination of data—any address that matches to a given parcel identifier is then linked to that parcel.

For Central Falls, Pawtucket, and Woonsocket, addresses from the emergency 911 point shapefile were standardized and matched to parcel records from each city's taxroll records. This process resulted in separate files specific to each municipality. The geocoded shapefile associated with the emergency 911 addresses served as a good starting point for identifying all the properties in each municipality. Inaccuracies and differences in completeness in these data continue across municipalities, however, and so we would not suggest relying solely on this source. For example, some cities' records include parcel identifiers in the file, but many do not. Furthermore, because the shapefiles are point data, rather than polygons, they provide no sense of the size and shape of property lines in relation to one another in the city.

The next step in our data preparation entailed actively looking for multiunit properties that were likely to have multiple addresses. Within each city, we selected the tax class that corresponds with

two- to five-family properties and, for each address, added 2 and subtracted 2 from the street number to create fields representing what would be the next address to the left or right of the property. If those addresses matched an existing address in the data, we discarded them, but, if they did not, we kept them as potential matches to expand the coverage of the MLT. If address numbers were recorded as ranges, we split them up to create multiple address records from the original.

Adding the municipality-specific datasets into one master file was the final step. To avoid any loss of data that could arise if plat and lot numbers overlapped in different cities, we created a city-specific parcel identifier field called "CKEY." The identifier field concatenated a two-letter abbreviation of the city name—CF (Central Falls), PA (Pawtucket), PR (Providence), and WN (Woonsocket)—and the parcel identifier code: the plat and lot numbers. For example, if the plat and lot numbers for a property in Providence were 1 and 1, the CKEY would have been PR 1-1.

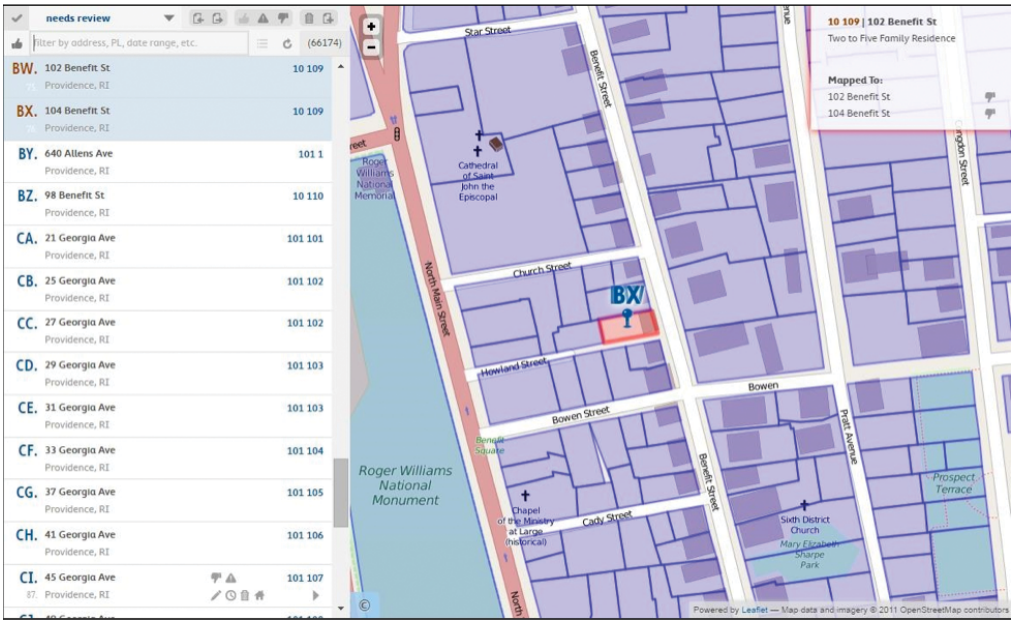
The most basic version of the MLT file contains all residential addresses, the municipality, and the corresponding CKEY, with all text in uppercase lettering to ensure standardization. The basic structure of the file is provided in the appendix. Variables that can be included as needed in the MLT include address number, tax class category, property type, and year built—essentially, any descriptive attribute that is unlikely to change regularly. Because the MLT is focused on linking addresses to parcels, each address within a city should be listed only once and addresses should not match to more than one plat and lot. Parcel identifiers can have duplicates, however, because when they match to more than one address, they are listed in separate rows in the table. The MLT created for the Lead Technical Study currently contains more than 100,000 address-to-parcel linkage records, with about 65 percent of those in Providence.

Providence MLT Online Tool

A secondary outcome of the MLT effort was the creation of an online tool that enables the user to import addresses to tag with parcel identifiers or download lists of addresses and parcels. It also enables registered users to modify records when new information is available or errors are found. See exhibit 1 for a visual example of how the MLT acts as a crosswalk between addresses and parcels. This online version was created using Leaflet, an open-source resource for interactive maps. Although the tool is currently available for Providence only and is a private website that requires a login, the concept is replicable.

Exhibit 1

Providence Master Lookup Table



Note: This screenshot of the master lookup table website provides an example of the interactive nature of the online tool. By clicking on records BW and BX, a pin shows up on the map. By clicking on the parcel, a box in the top right corner of the screen provides details about the property type, property owner (suppressed), and addresses associated with the particular parcel identifier.

Connecting the Data

After the MLT was complete, the next steps involved gathering the various datasets necessary to answer the research questions and then linking them together.

Study Datasets

Using the MLT for the four cities in the analysis, we linked city tax assessors' datasets, two lead compliance certificate datasets (from the Rhode Island Department of Health [HEALTH] and Rhode Island Housing Resources Commission), blood lead screening surveillance data from HEALTH, and foreclosure deeds datasets. Taxroll data provided the details needed to identify if a property was subject to the law, including year of construction and the owner's address. The lead compliance certificate records were integral to the policy evaluation, which aimed to assess whether outcomes differed based on having a certificate. Lead screening records provided the primary outcome variable—blood lead level screening results—and the address at the time of each screening. Rhode Island law mandates that healthcare providers screen all children for lead twice by age 3 and report the results, so the lead screening dataset covers most young children in the state. Foreclosure deeds allowed for compliance and lead exposure comparisons at properties being considered for possible foreclosures.

Each dataset used different formatting of address, unit, and property records. With the exception of the taxroll and foreclosure records, parcel identifier codes often either contained obvious errors or were completely missing in the datasets analyzed. The lead compliance certificate datasets had errors and missing records for the parcel identifier fields. The screening surveillance data records included only patient address fields, which were subject to data entry errors. In some cases, addresses included post office box addresses, which are not useful for this work because they do not indicate the physical address of residence and thus were excluded. Standardizing the addresses was essential.

Preparation for Matching

To prepare the existing address-level datasets, we ensured that any addresses conformed to the same format as the MLT. Geocoding the addresses (we used ArcGIS software) is a good way to begin the process, because it can correct some spelling errors or other inconsistencies automatically. Further editing required reviewing all addresses for errors, stripping any unit designations out of the addresses (apartment numbers, floor numbers, and so on), making sure public housing and similar complexes were in the same format, and standardizing common abbreviations such as "N" instead of "North." To match the format of the MLT, we also ensured that any text was all upper case. We used statistical software to clean up the datasets after geocoding, which systematized the process, because the syntax can be adapted to apply to each dataset and reused when new data need to be processed. We used IBM SPSS Statistics software, but other statistical packages would serve the same purpose. Each type of data we prepared for analysis had one file that contained address and municipality fields.

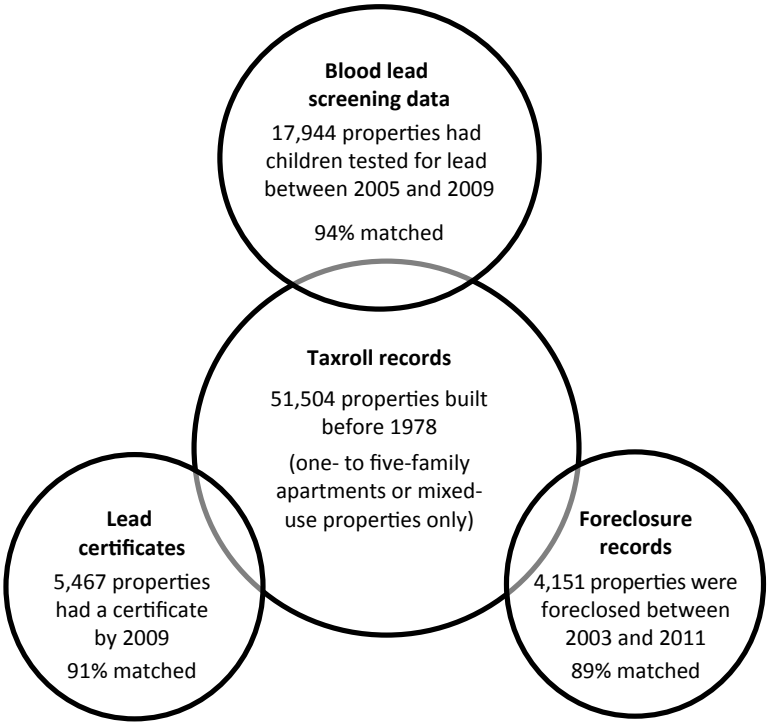
Matching

When the separate data files were fully prepared, we could match the addresses in the datasets for analysis to the addresses in the MLT. To avoid any complications with duplicate addresses across municipalities, the files were matched based on two variables: address and municipality. Matching one file per city at a time to the MLT records for that city could also avoid duplicative addresses. By processing all the address-level data through the MLT, we ensured that the addresses in each of the various datasets correspond to the same properties. Matching to the MLT resulted in each file containing the CKEY parcel identifier field.

After all the files contained a CKEY field, one main property-level analysis file could be created. The taxroll file at the CKEY level served as the base file to match the other property data. For the purpose of the Lead Technical Study, a subset of residential properties built before 1978 was selected—one- to five-family properties, apartments, and mixed-use properties—based on the corresponding codes in the taxroll data. Properties were then classified as either owner- or nonowner-occupied. We aggregated the address-level taxroll, foreclosure, and lead certificate records to the parcel level using the CKEY variable and matched the files to the taxroll. New fields indicated if a property had lead certificate records, any foreclosures, or was exempt from the law. See exhibit 2 for example match rates based on the foreclosure analysis from 2005 through 2009.

Exhibit 2

Properties That Matched Through the Master Lookup Table



Note: The match rates report the percentage of properties in the dataset from which they originate that matched the properties in our pre-1978 taxroll records.

The way we handled the lead screening data differed slightly depending on the unit of analysis. To prepare the lead screening data after matching it with the MLT, we analyzed one record for each child. If a child matched to more than one property, we kept one record per child per property. For the child-level analysis, we matched the property-level file of taxroll, foreclosure, and lead certificates data to the lead screening file to describe the property where the child was living and focused on the first result per child per property. For the property-level analysis, we aggregated the number of children who had screening records at a given address and identified whether one or more children's maximum test results were considered elevated. Thus, conducting analysis with these methods allowed for flexibility at the aggregation level.

Analyzing the Data

The crosswalk provided by the MLT helped create datasets that could be analyzed with relative ease at various levels of aggregation. For the analysis that led to the *American Journal of Public Health* article (Rogers et al., 2014), we focused mainly on outcomes at the child level and could account for children who lived in different properties or multiple children who lived at one property. We identified whether children who lived in properties that became compliant had declines in blood

lead levels, and we described the burden of lead exposure in exempt properties. For work outside the scope of the journal article, we investigated outcomes at the property level to describe how many properties housed one or more lead-exposed children, comparing lead exposure in compliant and noncompliant properties as well as between exempt and nonexempt properties. These property-based findings were presented to stakeholders and at conferences. Finally, we analyzed children's lead exposure by whether the property where they lived had been foreclosed on within a certain amount of time. When new data become available, these methods can be repeated to keep stakeholders updated on the status of associations of interest.

Discussion

Investing the effort to match data with the methods described in this article has numerous advantages, especially in urban communities with high proportions of multifamily properties. In many cases, administrative datasets that are relevant to housing and health will have address records, but not parcel identifiers. A data crosswalk such as the MLT provides not only a way to link disparate datasets but also rich layers of information to analyze. Without preparing a comprehensive property-level lookup table, a much higher proportion of data would be lost because of nonmatching. In addition, without translating data from address- to property-level status, count, proportion, or density calculations for a given area could mislead readers. The data are particularly misleading at smaller geography levels, such as census blocks and neighborhoods; the differences between property and address information can be meaningful.

The techniques employed to create the MLT and use it as a data-matching system could benefit researchers conducting analysis at the property level in other cities in the absence of established integrated data systems. The tool supported the evaluation of outcomes associated with housing policies—work that would have otherwise been unfeasible given the data landscape. Although the effort associated with having to conduct this work one municipality at a time could make a statewide analysis burdensome, the ability to target at-risk communities and, in some cases, at-risk properties has been valuable to stakeholders in Rhode Island. In summary, ensuring that linkages between property data are accurate and meaningful can lead to meaningful results. Those robust analyses can, in turn, guide policymakers to evaluate housing-related policies more effectively.

Appendix

Exhibit A-1

Fields for Master Lookup Table

Field	Type	Description	Case
SRC	text	Source of address and/or parcel information	
CKEY	text	Parcel level id# with 2-letter municipal code	
LOCATION	text	Original address	Proper
RNG	text	"Y" if multiple address records were created from original address	
ADDR_NUM	text	Address number	
ADDR_X	text	Address number extra (1/2, R, etc.)	Upper
ADDR_NAM	text	Address street name	Upper
ADDR_TYP	text	Address type	Upper
ADDR_SFD	text	Address suffix direction	Upper
PAR_ADDR	text	Full address formatted	Upper
ADDR_NMB	number	Address number	
ADDR_STR	text	Street name and type	Upper
CITY	text	Municipality name	Proper

Acknowledgments

The authors thank Jim Lucht, Kimberly Pierson, and all who were involved with the creation of the original master lookup table. They also thank Michelle Rogers for her input on this piece and her great work on the analysis, as well as Patrick Vivier, Ryan Kelly, Robert Vanderslice, and all those people who assisted with the lead technical study work thus far.

An award from the U.S. Department of Housing and Urban Development provided funding that supported the work that provided the basis for this publication. The substance and findings of the work are dedicated to the public. The authors are solely responsible for the accuracy of the statements and interpretations contained in this publication. Such interpretations do not necessarily reflect the views of the U.S. Department of Housing and Urban Development or the federal government.

Authors

Alyssa J. Sylvaria is a health policy and information specialist at The Providence Plan.

Jessica Cigna is the research and policy director at HousingWorks RI at Roger Williams University.

Rebecca Lee is the director of the Information Group at The Providence Plan.

References

Healthy Housing Collaborative. 2012. "Healthy Housing Data Book." http://www.health.ri.gov/publications/databooks/2012HealthyHousing.pdf.

Rhode Island Geographic Information System. 2014. "Facilities and Structures." http://www.edc.uri.edu/rigis/data/data.aspx?ISO=structure.

Rogers, Michelle L., James A. Lucht, Alyssa J. Sylvaria, Jessica Cigna, Robert Vanderslice, and Patrick M. Vivier. 2014. "Primary Prevention of Lead Poisoning: Protecting Children From Unsafe Housing," *American Journal of Public Health* 104 (8): e119–e124.

U.S. Census Bureau. 2014. 2008–2012 American Community Surveys. http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t.