Toward a National Eviction Data Collection Strategy Using Natural Language Processing

Tim Thomas
Alex Ramiller
University of California, Berkeley

Cheng Ren
University at Albany, State University of New York

Ott Toomet
University of Washington

Abstract

During the past decade, eviction research has relied heavily on preprepared (structured) data from third parties and state agencies who have taken the effort to create readable and accessible filing data. However, massive data gaps across the country exist because third parties may not provide a complete count of filings and many states do not have a formalized process to digitize, enumerate, analyze, or release information on evictions. In some states, the bulk of eviction filings are buried in court filings.

To address this issue, the Eviction Research Network developed a natural language processing (NLP) approach to mine court record images to enumerate and map eviction filing counts at the neighborhood level and help researchers identify disparities by location, race, and gender. This approach involved downloading eviction court record images from online county court systems, digitizing the text, isolating and geocoding addresses, and estimating demographics based on names and location.

In a case study for the State of Washington, millions of pages in more than 110,000 eviction filings from 2004 to 2017 were processed to demonstrate this approach. The research shows massive racial and gender disparities, where up to one in five African-American/Black female-headed households were named in eviction filings. Eviction rates peak in areas with the lowest rent and in the most diverse neighborhoods when analyzing neighborhood dynamics related to eviction. This research helped pass several tenant protection policies in the state and informed other strategies on how to address housing precarity. A suggested strategy for collecting eviction data across the country concludes the article.

Introduction

For renters, there are absolutely no benefits that come from an eviction. At best, the mark of an eviction impedes access to preferable housing for years to come (Franzese, 2018), even if a case had a favorable resolution for the tenant.[1] At worst, it forces vulnerable households to move within an average of 3 weeks or fewer (Davidson, 2019), often to lower-income neighborhoods with higher crime rates (Desmond and Shollenberger, 2015), and increases the odds of homelessness to 1 in 5—even without accounting for related housing precarity risk factors (Shinn et al., 2013). These consequences induce severe economic, mental, physical, and social harms, including higher debt, arears debts, court fees, and security deposits; declining credit scores (Parker and Smith, 2021); food insecurity; lower school performance (Fowler, Henry, and Marcal, 2015); disrupted social ties (Desmond and Shollenberger, 2015); longer commutes to work; unexpected time off or job loss (Desmond and Gershenson, 2017); depression; greater suicide risk; and exposure to infectious diseases (Benfer et al., 2021; Fowler et al., 2015; Hatch and Yun, 2021). Evictions feature severe racial and gender disparities, where the highest eviction rates fall overwhelmingly on the backs of African-American/Black renters, particularly female-headed households (Hepburn, Louis, and Desmond, 2020). As the United States exits pandemic-era renter assistance and policy protections, 19 million U.S. renters burdened with housing costs (U.S. Census Bureau, 2022) face new challenges such as higher inflation, higher costs for food and basic necessities, and even record-breaking eviction rates in some states (Legal Services Corporation, 2023).

Although the harms of eviction are well documented, data for measuring current coverage and trends remain largely fragmented and incomplete (Pan, Zainulbhai, and Robustelli, 2021). These information gaps leave many policymakers woefully uneducated about the trends and extent of the problem in their jurisdiction and, consequently, ill-prepared to act swiftly or advocate for better housing policies. The reason for these gaps is that eviction records are generally processed in jurisdictional court systems that vary widely in recording and storage protocols. Several scholars and institutions have been able to collect structured data from some of the better-organized court systems, which include names, addresses, judgment amounts, and resolutions. Some states only provide county-level counts, which conceal very important details about neighborhood-level processes, such as housing market effects and concentrations in marginalized communities. For the rest of the country, eviction details are buried deep within court filing texts that are either in PDF images or in physical paper form—an almost impossible data source to mine until now.

The focus of this article is to (1) describe how data science tools can be used to extract records from jurisdictions with opaque eviction recordkeeping and (2) illustrate how these tools can supplement existing data collection practices to build a comprehensive national dataset. This type of dataset would allow scholars, policymakers, and the public to analyze in and between jurisdictional trends, measure the severity of the eviction problem, and identify solutions. This supplemental approach uses a natural language processing (NLP) technique to mine court records and fill gaps in missing data for underrepresented and underresourced states and counties. This article applies the NLP method to a case study in Washington State and demonstrates its practical application, hurdles, best practices, and findings. It also describes the political impact

[1] *Smith v. Wasatch Property Management, Inc., et al.*

of enumerating unknown populations, which motivated the adoption of several tenant protection policies across the state, including just cause and extending the state's pay-or-vacate notice period from 3 to 14 days. The article concludes with a recommended outline of steps and tools to apply this approach on a broader scale.

Background

Evictions happen to the most vulnerable citizens in the country when the primary causes at the household level are inadequate minimum wages and insufficient welfare support competing with rising rents (Desmond, 2012). At the neighborhood and county level, rental markets and race are two dominate themes. More specifically, regions that have the lowest median rent, volatile[2] (or gentrifying) housing markets, and higher proportion of African-American/Black tracts in Whiter counties have higher eviction rates (National Academies of Sciences, Engineering, and Medicine, 2023). These analyses draw from a subset of eviction data in the country and require a larger dataset to confirm generalizability and to further explore the nuances among different regions. The primary goal of this study is to collect and structure the most difficult to obtain data to improve coverage and further scholarship; however, before discussing this approach, a general understanding of the eviction process and how these data are structured is required.

The Eviction Process and Data Points

An eviction is not a single event, but rather a process that plays out over time and is documented in varying ways with varying outcomes at each stage. There are five primary stages: (1) prenotice, (2) notice, (3) court filing, (4) writ of restitution, and (5) physical removal. This process can play out in less than 5 days or up to 53 days, depending on varying state rules about how each event is executed (Davidson, 2019).[3] The prenotice—also referred to as an informal or illegal eviction—is not documented and therefore impossible to analyze. Notices are rarely collected, despite being the starting point of the legal process. The most commonly available and studied data are eviction filings, followed by sheriff lockout data. An important caveat about eviction research is that the actual number of renters who face eviction is likely severely undercounted because the sum of notices and informal evictions that precede a court filing is unknown. Estimates suggest that there may be anywhere between two and five-and one-half informal evictions for every formal eviction recorded by the courts (Desmond, Gershenson, and Kiviat, 2015). In addition, outcomes and mobility patterns are difficult to determine because tenants may either move at some point within the five stages or strike a deal with the landlord to stay. On rare occasions, filing data will include judgment resolutions, which allows researchers to make a few assumptions about the outcome (e.g., default judgment means the tenant did not show up to court and the landlord's demands were likely favored in the case).

[2] Volatility is measured as the median rent gap in a county: The degree to which tract median rents are lower or higher than nearby tract rents. In counties where nearly all tracts share similar median rents, little volatility appears in the rental market, and vice versa.

[3] Georgia, Indiana, Maryland, Minnesota, Missouri, New Jersey, North Carolina, Oklahoma, Pennsylvania, and West Virginia combine the notice and court filing period into one event.

Filings may be available in one of three different primary forms: structured data, data within digital images of court records, and data only available on paper records. Structured data generally refer to easily accessible tabular formats that record the attributes of each eviction record, facilitating easy analysis. However, these types of data are only available in certain jurisdictions where court systems or sheriff's offices have decided to record data in that manner. Other jurisdictions may provide aggregate counts of evictions within specific timeframes. More commonly, courts will only have scanned images of civil court records that require computational language tools to mine the text or paper copies, which require the extra step of being scanned.

One of the most important, but often missing, data points is the evictee's address. Addresses allow researchers to get a closer look at eviction trends otherwise concealed by aggregated county counts. Addresses allow for examining neighborhood dynamics and where rates may concentrate within a city and provide a deeper understanding about demographic stratification of eviction through racial and gender estimation. Structured data occasionally provide addresses, but language models are necessary to extract them from court texts.

Current State of National Eviction Data Collection

The current landscape of eviction data collection in the United States consists of several independent organizations and research institutes that each contribute to a more complete national picture of evictions. The data collection strategies of these organizations fall into several main categories. First are organizations such as the Eviction Lab[4] and Legal Services Corporation (LSC)[5] that aim for a truly national data collection strategy, gathering structured data from as many sources as possible, even if those data are only available at higher geographic scales. Second, several local organizations have opted for a more targeted approach, collecting comprehensive data in a single county, region, or state and supplementing this analysis with local knowledge about the legal and technical specificities of the eviction process.[6] A final emergent approach, exemplified by groups such as the Anti-Eviction Mapping Project (AEMP)[7] and the Eviction Research Network[8] involves a combination of these approaches, combining local knowledge with sustained data collection efforts across multiple different jurisdictions.

The most comprehensive data currently available on evictions in the United States are currently held by the Eviction Lab, which uses a combination of data collection techniques to construct an eviction dataset with a national scope. For structured case-level datasets, Eviction Lab makes bulk data requests to state court systems throughout the United States, which has yielded records from 16 states and the District of Columbia. For county-level eviction data, Eviction Lab submits annual requests to state and county court systems, receiving data from 2,204 counties across 46 states.

[4] See Eviction Lab at https://evictionlab.org.

[5] See Legal Services Corporation's Eviction Tracker at https://civilcourtdata.lsc.gov.

[6] See the Atlanta Regional Eviction Tracker at https://metroatlhousing.org/atlanta-region-eviction-tracker/; Root Case Research at https://www.rootcauseresearch.org/lel; The University of Michigan's Eviction page at https://poverty.umich.edu/research-funding-opportunities/data-tools/michigan-evictions/#:~:text=Key%20Findings,6%20rental%20units%20(17%25); and the Richmond Virginia Eviction Lab at https://rampages.us/rvaevictionlab/.

[7] See the Anti-Eviction Mapping Project at https://antievictionmap.com.

[8] See The Eviction Research Network at https://evictionresearch.net/.

Finally, in an effort to fill in the gaps left by these methods, Eviction Lab purchases proprietary data from groups like LexisNexis Risk Solutions that document some cases from the local level but also include manually written, paper-based case management systems (Gromis et al., 2022). Like Eviction Lab, LSC aims for comprehensiveness. Eviction Lab either receives datasets from collaborators or purchases them from third-party data vendors, whereas LSC scrapes digital data from county court systems to create its own structured datasets. LSC has collected data from 32 states, including 18 with complete state coverage. Data from 13 of those 32 states contain name and address information, including 5 with complete state coverage.

As an alternative to this comprehensive approach, which is wide-ranging but less able to capture hard-to-reach records or compare legal differences between jurisdictions, hyper-local approaches to eviction data collection involve collecting data in a single city or region and leveraging those data to explore specific local research questions and challenges. Researchers at Georgia Tech (Raymond et al., 2020), for example, have leveraged the collection of eviction data in the Atlanta region from local county court systems to explore the relationship between evictions and the growing trend toward investor purchases of single-family rental properties, which is a particularly acute problem in Atlanta and other Sun Belt cities. The Evicted in Oregon team at Portland State University (Bates, 2023) has similarly focused on evictions, primarily within the state of Oregon, and gathered knowledge about the local legal landscape through a mixed-methods approach, including analyzing data, conducting interviews, and observing courtrooms. Some organizations have branched out from this local approach by collecting data from multiple different jurisdictions across the country, combining local specificity with an increasingly nationwide scope. This work includes the AEMP, which began by collecting eviction data in the San Francisco area in partnership with local tenant-advocacy organizations and has since expanded its data collection to Los Angeles and New York. Finally, the Eviction Research Network applies data science tools to process, analyze, and visualize collaborator data and to collect data from more difficult areas, such as Washington, California, Chicago, and Baltimore.

The current approaches to eviction data collection reveal an important gap in the landscape of eviction research. Although national-scale efforts by Eviction Lab and LSC have been successful in collecting high-level data for many jurisdictions and detailed records for a smaller number, these organizations are not able to capture less accessible records or data that only appear in unstructured formats. These forms of data may be usable to researchers working on data collection for a single city or region, but a significant amount of resources would be required to collect, clean, and process a larger set of unstructured eviction data, such as digitized PDFs of court records and physical files. This challenge calls for data science tools like NLP that can dramatically streamline the collection and cleaning process for unstructured data.

Data

To demonstrate the NLP approach, court records from Washington State were examined, where counts are only available at the county level. The goal in this project was to fulfill a request by the Washington State legislature to estimate racial and gender disparities, which required both names and addresses of eviction defendants. The data collection process began with a public records request to the state, in collaboration with the Washington Office of Civil Legal Aid, for case numbers that included names, resolution, and judgment amounts for every county in the entire state from 2004 to 2017. Both resolution and amounts were incomplete, and addresses were not provided, requiring access to the court records to mine these elements.

Each state court system operates and holds records differently. For Washington, these records were stored on three different web portals, where access was granted at the discretion of the elected county court clerk. Permission was requested from most of the 39 court clerks in the state, targeting the more populated counties around Puget Sound, near the Oregon border, and Spokane County in eastern Washington. In addition to county clerks, local legal aid providers who operated in these counties were contacted to see if they would advocate to the clerk for access.

The reception for requesting online access was mixed. One county clerk enthusiastically granted permission, and two others willingly provided access through local legal aid contacts; the fourth county was more difficult because the clerk was unwilling to waive the 25 cents per downloaded or printed page charge—an estimated $350,000 cost for all the records from 2004 to 2017. Luckily, a subset of records was provided by a local legal aid provider. The most common reasons for unsuccessful attempts were in part due to clerk unresponsiveness and a lack of support resources. One county clerk said that the state of Washington had the second least funded court system in the country, which might explain the unwillingness among most counties to extend any resources toward a project not mandated by state law.

In the end, full access to Pierce, Snohomish, and Whatcom's online portal was provided for the study period of 2004 to 2017 and limited access to King County's portal, where records were pulled from 2006, 2010, 2011, 2013, and 2017. Using the case numbers provided by the state, an HTML scraper script was built that input the case number in the respective portal and pulled the record at a delayed interval in order to avoid overloading the servers. The scraper downloaded 111,740 PDF or TIFF court record images. Each record held several parts of the eviction proceeding, the most important of which was the eviction summons,[9] which typically listed both the defendant's name and address of residence.

Methods

Extracting Geographic Data from Court Documents

Text extraction proceeded broadly in the following fashion: First, the original document image files were converted to text; the defendant's address was then extracted from the converted text; and finally, the address was geocoded in a census-tract and estimated demographics. The first step of converting images to text was completed using *Tesseract* optical character recognition (OCR) software (Tesseract-OCR, 2023). Because Tesseract cannot handle PDF and TIFF images, a preprocessing step converted these files to PNG raster images at 200 dots per inch (dpi) resolution using the *ImageMagick*® software suite ("ImageMagick," 2023).

[9] An eviction summons is an official notice issued by a court to renters stating that an eviction process has been initiated against them. It typically lists name(s) of the lessees and the landlord(s), address of the premises, and attorneys name(s) if any are involved. The notice also gives lessees a response deadline.

The extracted text quality relied heavily on image quality. Poor quality images can lead to a host of problems, such as causing errors in the text interpretation. For example, the numeral “1” or capital letter “I” turning into the pipe “|” symbol, superscript letters like $^{\text{“th”}}$ used in street names transforming into a quote mark “"”, or in other cases, misspellings due to misidentification of similar-looking characters (e.g., “Seattle” may turn into “Seatile” or “Bellingham” to “Bellmgham”). Another error comes from how court documents are formatted with numerals on the left side of the page, which leads to misclassification of columns (e.g., an address spanning row number 17 and 18 incorporates “17” and “18” in the middle of an address, such as “123 4th street S 17 18 Seattle, WA,” where “17” and “18” in the middle of the address are the row numbers). An additional challenge comes from handwritten documents, which Tesseract cannot recognize. Finally, letters and numbers may be concatenated if a separator, such as a space, is missing (e.g., “123 N Jackson Street” may turn into “123N Jackson Street”). Despite these issues, most addresses were extracted correctly.

In the second step, text is taken from the first step where all addresses are extracted using two approaches: *rule-based recognition* using regular expressions and *named entity recognition* (NER) using the neural network–based *spaCy* library (“spaCy”, 2023). The rule-based address recognition approach uses a complex set of rules to recognize addresses in the text by looking for house numbers, street type patterns (e.g., *“street,” “st,” “str,” “ave,”* etc.), ZIP Code patterns, and so on, in a given order. Although most addresses follow a clear distinct pattern, several less-common cases are harder to handle, including addresses ending in a county name; missing ZIP Codes or states; names that look like something else, such as “Federal Way” being a city, not a road, and “street court” being a street type, not a separate street and court; long multiword names like “Martin Luther King Jr,” which is a single four-word street name; and false positives, such as “Superior Court of the State of Washington,” which tends to be misclassified as an address in Washington. The algorithm starts by looking for the address on the first page of the summons file, where they are normally located. If it cannot find an address on that page, the algorithm looks at all other documents in the respective case and extracts all the addresses it can find. Any additional addresses are parsed in a string of components, such as house number, direction, street name, and street type. This information is used afterward to convert the written addresses to a normalized form in which all acronyms are replaced with expanded lowercase words.

As an alternative to the rule-based method, a NER approach using neural networks trains the *spaCy* library to recognize addresses in documents. The training data are generated through multiple steps that include fake addresses; a negative sample of other kinds of similar objects, such as phone numbers and dates; and a set of actual documents with addresses labeled correctly. Based on these training data, the model outputs a large number of addresses (and a few false positive entities). Next, these addresses are parsed into components like in the rule-based method using the *spaCy* neural networks, and this trains the algorithm to distinguish components using 600 manually parsed addresses (including negative examples). Both steps work rather well, but the parser is better at skipping spurious words and symbols related to the problem of numbers in the first column.

The main issues include missing separators between address components (e.g., missing spaces between a street number and prefix). Sometimes, the address components are misidentified or combined in a wrong order. For instance, “123 N Alder street” may be read as “123 n.alder street,” and the resulting street name will be “n.alder.” Other issues include mistaken conversions between letters and numbers (e.g., prefixes like “s” can be mistaken as a house number like “5”). In this case, the house number will either be wrong or completely missing in the result. Finally, multiword street or city names may confuse the parser, causing it to skip part of the name, mistakenly considering it being a spurious word from another column.

To distinguish defendant addresses from other addresses (e.g., attorney’s office), a simple Naive Bayes algorithm is used to look at words surrounding the address and computes the probability that this result is the address of interest. Thereafter, the algorithm picks the address with the largest such probability.

Geocoding

Next, structured addresses are geocoded for spatial analyses and demographic estimation. Several geocoding platforms were tested, each using slightly different databases and methods for geocoding, including Open Street Map (OSM); Census Bureau, Google, Azure, and Esri Business Analyst (BA) data; and Master Address File (MAF) for King County. Free Open Street Map and Census Bureau geocoding platforms struggle with interpreting ambiguous address strings. Pay-as-you-go services Google and Azure operate with a much greater degree of effectiveness, each of which using proprietary fuzzy string-matching algorithms to find matches—even for more ambiguous address strings. BA geocoding data operated through Esri’s ArcMap program are purchased up front for 1 year. Finally, the King County MAF data are free to the public and provide a record of every single address location within a local jurisdiction for administrative recordkeeping and emergency services.

Each geocoding platform offers different strengths. OSM and Census Bureau geocoders are optimal in terms of accessibility, although both Google and BA offer various levels of address accuracy from the administrative unit to the rooftop. Crucially, Census Bureau and BA approaches share the advantage of reproducibility because they are based on static benchmarks that can be referenced at any point in the future (although the Census Bureau’s dataset is regularly updated, it saves benchmarks at each point in time). Other data sources, such as OSM, Google, and Azure, meanwhile, continually update without retaining previous records, which makes it impossible to reproduce the exact same results at a later point.

Each geocoder was tested on approximately 5,000 manually extracted 2013 King County addresses. These addresses had a success rate of 85.1 percent for OSM, 90.6 percent for the Census Bureau geocoder, 92.1 percent for the King County MAF, 97.8 percent for Esri BA data, and 99.9 percent for Azure. For the NLP addresses, success rates were consistently lower: 75.6 percent for OSM, 77.2 percent for the Census Bureau geocoder, 84.5 percent for the King County MAF, 90 percent for Esri BA data, and 99.9 percent for Azure. The high return rate for Azure includes geocodes at the center of some places: 81.6 percent for “address point,” 12.3 percent for “address range,” 5.5

percent for "street," 0.3 percent for "geography," and 0.3 percent for "cross street."[10] Azure was the most successful in geocoding the provided addresses; however, in terms of geocoding tools with a high degree of replicability, Esri BA data had the highest rate of success. Azure's spatial accuracy will range from rooftop coordinate to the centroid of a ZIP Code or county. The final dataset uses a subset of addresses with rooftop to block accuracy, which is about 93 percent of the NLP addresses. Addresses with accuracies at the "street" centroid or "geography" (e.g., county centroid) were omitted because they likely fell within tracts far from their true location.

Validation

Address Validation. Did the correct defendant address get extracted? A Naive Bayes probability approach alone determined that the correct address was obtained with a confidence of 98 percent. Repeated addresses from a large number of cases were also manually analyzed. In most cases, these addresses were apartment complexes, although at least one incorrect place slipped through (e.g., Pierce County courthouse). However, the Naive Bayes probability was relatively low in cases with the 90th percentile being less than 90 percent.

NLP address components were also compared to the manually extracted addresses from 2013, using the average ratio of the Levenshtein distance as a metric of accuracy—a method to measure the differences between two strings. The ratio is calculated by the python-Levenshtein package (version:0.12.2) with the formula:

$$ratio\ (a, b) = \frac{(lensum - ldist)}{lensum} = \frac{((len(a) = len(b)) - dist\ (a, b)}{(len(a) + len(b))}$$

For example, the distance between "apple" and "appl" is 1, and the ratio is 0.89. After comparing the manually extracted addresses and algorithmically extracted addresses, the average ratio was 0.82. In this situation, if the ratio was more than 0.7, the two pairs were usually on the same street or a close street, although when the ratio was lower, the two pairs may be in the different cities. A common error is from algorithmically extracted addresses; the method extracts the address of the court or attorney's office, rather than the address of the defendant. The final result for this comparison yielded an accuracy of 93 percent.

Geographic validation. Address accuracy was evaluated by comparing the distance between the coordinates of the 5,000 manually extracted addresses and NLP coordinates and determining whether the two addresses landed in the same census tract. When using Azure's geocoder and isolating locations within the county of the case, the average distance was 3.2 miles. When joining these locations to census tracts, 83.1 percent of the pairs shared the same census tracts on Azure. For OSM, the accuracy was 63.3 percent, which is much lower because OSM was unable to geocode as many addresses as Azure.

[10] Geocoders provide various accuracies for a given address. "Address point" is the highest level of accuracy where the latitude and longitude mark the exact point where the address is located. "Address range" means the latitude and longitude falls within a range of addresses between cross streets, but not at the exact address location. "Cross street" means the latitude and longitude marks the nearest cross street. "Street" means the address number was difficult to find, so the latitude and longitude marks the centroid of the street. This can be problematic when the street is long and crosses several geographies such as census tracts. Finally, "geography" marks the centroid of a much larger geography such as a city, county, or state.

Race and Gender Estimation

With addresses and names, the race of the defendant is estimated by running a first batch of surnames through the R packages *wru* (Khanna et al., 2023) and then *rethnicity* (Xie, 2022) for the few hundred names that could not be estimated in the first round. The ecological inference package *wru* uses the Bayes' Rule to examine the racial probability of a surname compared to the racial composition for each neighborhood (census tracts) where evicted defendants lived.[11,12] This estimates a racial category of White, Black, Latinx, Asian, or other for everyone with geocoded addresses. The few names that were not estimated by the *wru* package are passed through the *rethnicity* package—a Bidirectional Long Short-Term Memory (Bi-LSTM) model based on a recurrent neural network architecture commonly used for natural language processing. This package also uses voter registry records for training data. Although these methods have been widely used in various studies, including eviction research (Hepburn, Louis, and Desmond, 2020; Hepburn, Rutan, and Desmond, 2022; Thomas, 2017; Thomas et al., 2020), they struggle with correctly assigning race and ethnicity in states outside of its training data, which were voter files based in Southern states. Sampled observations of the outputs showed that *wru* did well in estimating Asian, Black, Latinx, and White surnames, whereas *rethnicity* did well with Black and Latinx names.[13] Gender is inferred using the R package *gender* (Mullen, Blevins, and Schmidt, 2023) which cross-validates the first name of an individual with the Social Security Administration (SSA) Name Registry from 1932 to 2012 and the U.S. Census Bureau Integrated Public Use Microdata Series (IPUMS). To see if estimates were close to real-world conditions, demographic estimates of the manually extracted addresses from 2013 were compared to aggregated intake data from a King County tenant attorney's office for that year, and it was found that the compositions of each group between the two datasets were within several percentage points. Although this comparison improves confidence in the demographic estimates, there should still be some caution regarding the reliability and interpretation of these values.

The final dataset for King, Pierce, Snohomish, and Whatcom Counties held 111,740 cases, 91.7 percent of which were geocoded and had racial estimation, 95.7 percent had gender estimation, and 87.7 percent had both race and gender[14] estimation. The data was then subset to 2013–17 to compare the 2017 5-year PUMS data for renter households by race and gender. Racial demographic counts for King County's missing years of data between 2014–16 were interpolated using a linear regression controlling for the data's county, counts, and year.

[11] The most recent version of Khanna et al.'s fully Bayesian Improved Surname Geocoding (fBISG) method is used for the estimates.

[12] For example, a person with the last name Jackson, a common Black surname, living in a high-Black neighborhood would have a greater likelihood of being Black. Whereas the same name found in a high-White neighborhood would have a lesser probability of being Black.

[13] Asian and White estimates in *rethnicity* included more Middle Eastern and non-European names, which *wru* includes in the "Other" category. The final racial estimate consisted primarily of *wru's* estimates where *wru's* "Other" identification was replaced for any names that received a Black or Latinx designation from *rethnicity*. This helped improve a few hundred estimates from "Other" to Black or Latinx.

[14] Race and gender estimates were calculated by multiplying the race and gender probabilities together.

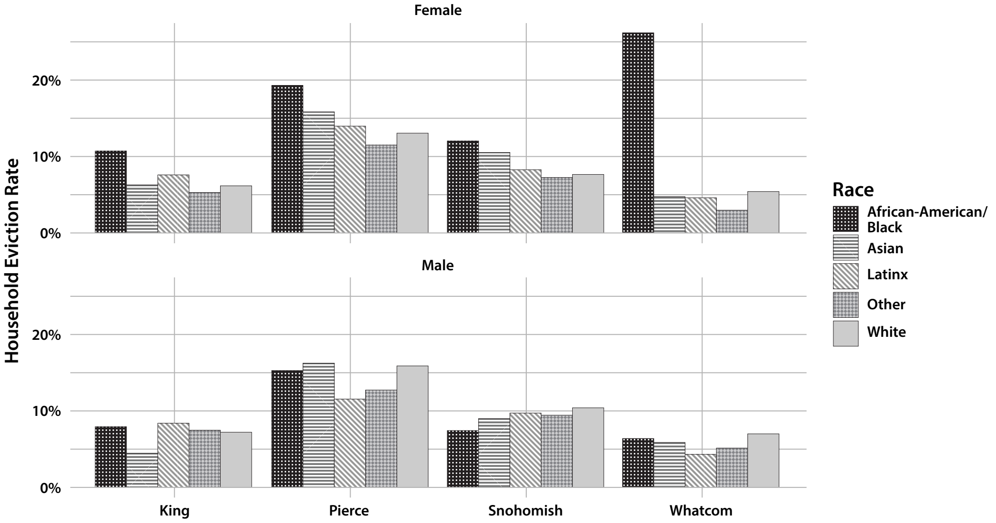
Findings

The NLP address extraction and demographic estimation process revealed sizable racial disparities in the data that would have otherwise been unknown (see exhibit 1). Particularly, one in five (19 percent) of Black female-headed households in Pierce County were named in the process, and one in nine (11 percent) were named in King County. What is important about this finding is that Pierce County was one of the largest recipients of new movers from across the country and from neighboring King County, home of gentrified Seattle (Balk, 2023; 2017). This movement increased demand in the area and applied market pressure on more affordable, and often more racially diverse, communities. King County is unique in that the majority of evictions were located on the south side of the county, which is a common destination for Black families displaced from Seattle (Thomas, 2017). Whatcom County had the smallest number of Black female householder cases (n=15) but the highest rate of 26 percent (one in four households).

King County recorded judgment amounts in less than one-half of all cases, with approximately 41 percent of these cases having estimated race (see exhibit 2, truncated to $10,000). Not much variation in the medians by race exists, but Black households had the lowest median judgment amounts at $2,840.

Exhibit 1

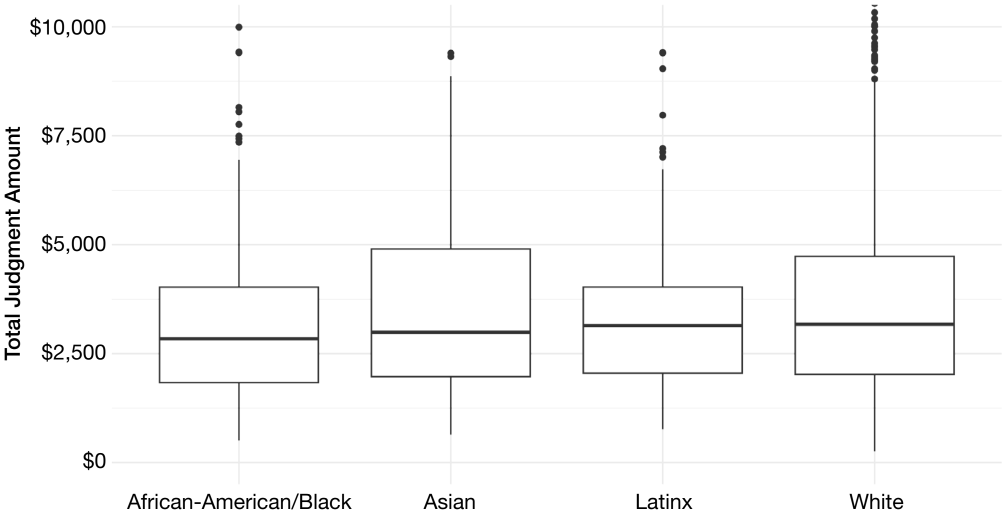
Five Year Eviction Rates by Race and Gender in Puget Sound, 2013–17



Source: Estimated demographics of heads of household with eviction filings over race and sex of renter heads of household (U.S. Census Bureau 2017 5-year American Community Survey)

Exhibit 2

Judgment Amount Totals by Race in King County, 2017



Source: Unlawful detainer judgment amounts from the Washington State Courts and estimated race and ethnicity

Judgment resolution by race shows that default judgments in which the tenant did not show up in court were the most common outcome, with the highest share being among Latinx householders (see exhibit 3). Next was dismissal without trial, which can mean a failure to appear for trial for both parties or that the final documents were not filed. Between 9 percent and 15 percent of cases ended in settlement or agreed judgment, which most likely resulted in some sort of payment plan or agreement to resolve the unlawful detainer.

Exhibit 3

Judgment Resolution by Race in King County, 2017

Resolution	Asian (%)	Black (%)	Latinx (%)	White (%)
Default Judgment	48.3	46.3	**51.6**	48.5
Dismissal Without Trial	**25.4**	21.9	25.0	22.0
Settled by Parties and/or Agreed Judgment	12.7	**14.9**	9.4	12.3
Closed by Court Order After a Hearing	7.6	10.8	**10.9**	10.7
Dismissal by Clerk	2.1	2.3	1.6	**3.4**
Uncontested Resolution	1.7	**1.8**	1.0	1.3
Court Decision After Trial	0.8	0.3	NA	0.2
Summary Judgment	0.8	**1.8**	0.5	1.4
Transferred to Federal Bankruptcy Court	**0.4**	NA	NA	0.2

NA = data not available.

Source: Unlawful detainer judgment resolutions from the Washington State Courts and estimated race and ethnicity

Neighborhood analysis shows that evictions occurred in communities with the most racial diversity and lowest rents in the region. Temporal spatial analysis shows that case frequency was greater in South Seattle's racially diverse neighborhoods in earlier years and then increased outside of the city in South King County, both of which are known displacement destinations for Black, Indigenous, Latinx, and Asian households. In Pierce and Snohomish Counties, evictions concentrated largely in segregated urban spaces.

Future Strategies

Several alternative approaches stand out for future consideration. The first is document layout analysis, which employs computer vision and NLP to extract contents of interest by their location in the image and would help reduce errors such as misidentifying the defendant address. Document layout analysis has its limitations, such as the high cost compared with the technique introduced in this article.[15]

Another promising technique involves the use of large language models like ChatGPT, where researchers can input unstructured court file content and request outputs like the defendant address. This method was tested on the 2013 training data and highlighted several concerns, such as how current models collect data, which can lead to disclosing sensitive information, and the quality of the returned content can sometimes fall short of expectations.[16] Additional efforts may thus be required to transform this output into a structured dataset. Despite these drawbacks, these two methods should lessen barriers for researchers with limited computational skills in text analysis and even help extract more information, such as the cause of eviction or amounts.

Conclusion and Discussion

Combining data science tools with existing eviction data collection efforts can fill in massive gaps. The NLP approach can extract addresses from court record images, creating the opportunity to analyze demographic disparities in the State of Washington and create a clearer picture of who was being evicted and from where. This work also provided evidence-based research for several statewide policy changes, including extending the pay-or-vacate notice period from 3 to 14 days (Senate Committee on Housing Stability & Affordability, 2019) and just cause.

A recommendation to build a more comprehensive national eviction dataset consists of continuing the practice of collecting structured data where available and then gathering court records for text mining in regions with lesser-known trends. This process will more than likely involve public record requests or direct contact with the respective jurisdiction's datakeepers to gain access, possibly requiring document scanning if records are not already in image formats. With records in hand, apply the following NLP process to these texts: (1) conduct OCR to convert images to text using tools like Tesseract or paid services like Azure or Amazon Web Services (AWS); (2) extract content, such as addresses, using rule-based recognition or NER through tools like spaCy;

[15] During the model customization and training process, researchers must manually label a small diverse sample and then employ a graphics processing unit (GPU) to train the model for application. This process necessitates a careful tradeoff between resource expenditure and data accuracy.

[16] For instance, a defendant's address may be requested, but the address is intermingled with the defendant's name.

(3) geocode addresses using tools from the Census Bureau or Azure; (4) validate the address; and optionally, (5) apply demographic estimations using packages like wru, rethnicity, and gender.

Several factors should be considered before scaling this approach nationwide. First, the sociological, demographic, and technical nature of this research requires an interdisciplinary team of scholars to ensure that ethical and accurate data integrity practices and considerations are implemented throughout the process of collecting and engineering these data. Second, engaging with court officers around collection can be difficult, especially if no motivation or resources are available on their side to facilitate cooperation with collectors. This process may require mandates from legislatures, grant support, or other considerations not discussed in this article. The upside is that a comprehensive database now seems like a realistic goal.

Appendix

Exhibit A.1

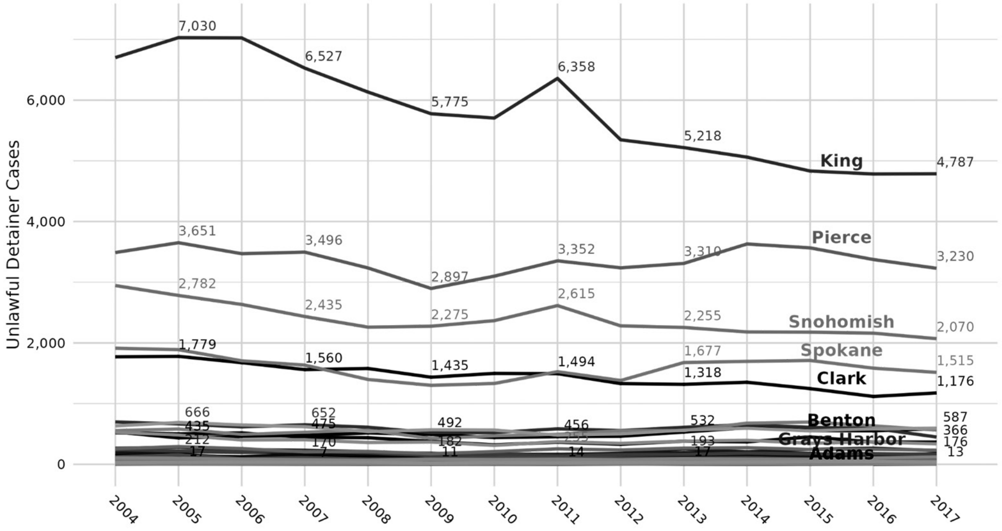
Five Year Eviction Rate by Race and Gender in King, Pierce, Snohomish, and Whatcom Counties, 2013–17

	King	Pierce	Snohomish	Whatcom
Total Cases	24,467	17,042	10,817	1,816
Asian Female Householders	6% (1,233 of 19,663)	16% (460 of 2,905)	10% (337 of 3,215)	5% (27 of 568)
Asian Male Householders	4% (1,420 of 31,607)	16% (453 of 2,777)	9% (390 of 4,321)	6% (29 of 483)
Black Female Householders	11% (1,811 of 16,909)	19% (1,392 of 7,233)	12% (240 of 1,993)	26% (15 of 57)
Black Male Householders	8% (1,330 of 16,727)	15% (1,059 of 6,937)	7% (241 of 3,221)	6% (14 of 215)
Latinx Female Householders	8% (1,315 of 17,431)	14% (754 of 5,389)	8% (547 of 6,618)	5% (69 of 1,480)
Latinx Male Householders	8% (1,654 of 19,654)	12% (751 of 6,502)	10% (749 of 7,675)	4% (62 of 1,411)
Other Race/ Ethnicity Female Householders	5% (698 of 13,040)	12% (674 of 5,847)	7% (260 of 3,563)	3% (42 of 1,401)
Other Race/ Ethnicity Male Householders	8% (774 of 10,274)	13% (653 of 5,109)	10% (309 of 3,247)	5% (45 of 868)
White Female Householders	6% (6,460 of 103,700)	13% (5,011 of 38,386)	8% (3,484 of 45,496)	5% (703 of 12,897)
White Male Householders	7% (7,772 of 106,571)	16% (5,835 of 36,630)	10% (4,259 of 40,777)	7% (812 of 11,518)

Source: Demographic estimates of unlawful detainer heads of household and renter demographics from the U.S. Census Bureau Private Use Microdata Sample

Exhibit A.2

State Unlawful Detainer Raw Count by County

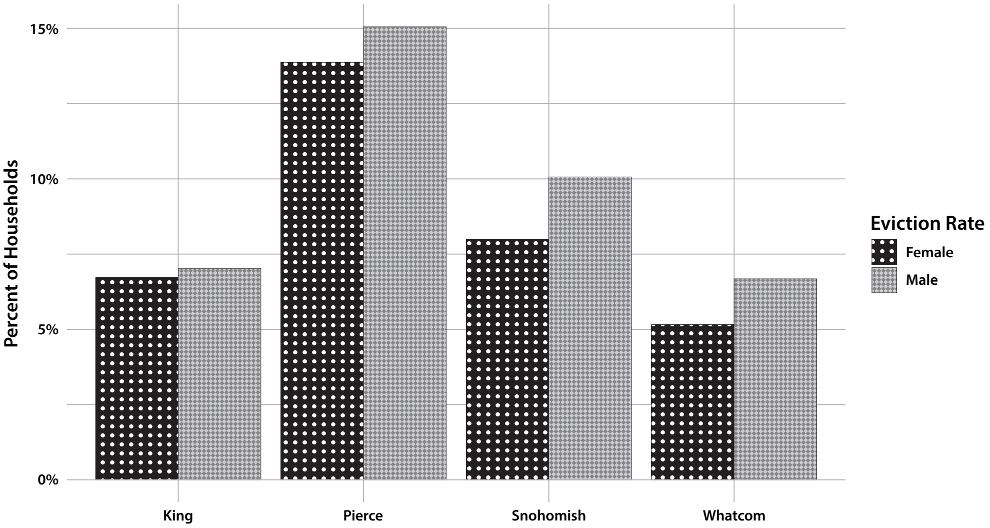


Source: Unlawful detainer filings from the Washington State Courts

County trends show that 5 of Washington's 39 counties had more than 1,000 cases per year, with King County in the lead with less than 5,000 cases in 2017 (see appendix exhibit A.1). Immediately before the Great Recession, state case counts peaked at approximately 22,500 unlawful detainers in 2005 and then declined to approximately 17,500 in 2017. During this time, the Puget Sound region experienced several changes in population growth, housing costs, and demographics. Big technology companies brought in higher-earning workers, spurring gentrification and displacement of Seattle's more vulnerable populations, particularly for Black and brown households. Between 2013 and 2019, King County saw a rapid rent increase from a median of $1,400 to $2,200, which requires a take-home income of approximately $90,000 to avoid rent burden. Black and Latinx median household incomes rested near 50 percent of the area median income of $40,000 and $55,000, which provides little room to afford the cost of living (Thomas et al., 2020).

Exhibit A.3

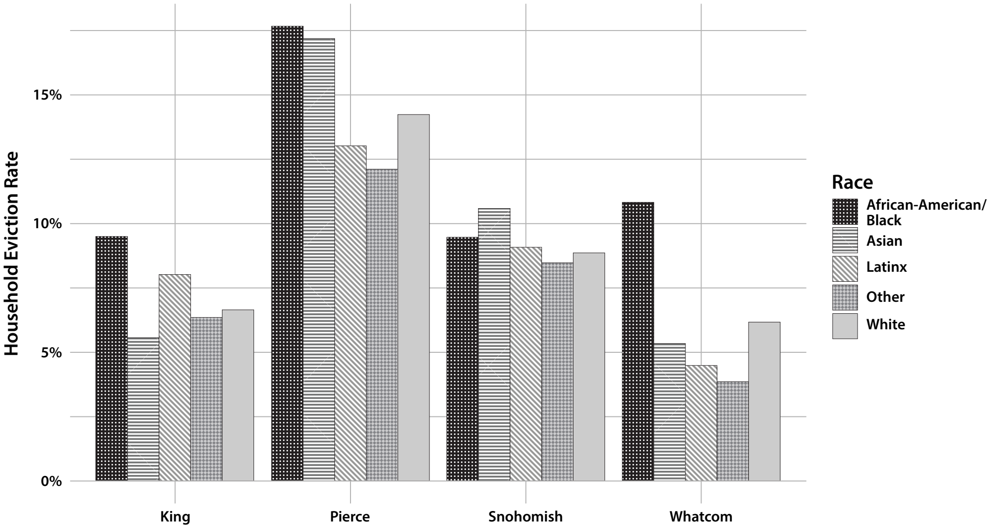
Unlawful Detainer Case Rates by Gender, 2013–17



Source: Estimated gender of heads of household with eviction filings over sex of renter heads of household (U.S. Census Bureau 2017 5-year American Community Survey)

Exhibit A.4

Unlawful Detainer Case Rates by Race, 2013–17



Source: Estimated race and ethnicity of heads of household with eviction filings over race and ethnicity of renter heads of household (U.S. Census Bureau 2017 5-year American Community Survey)

Acknowledgments

This work was funded in part by the Moore/Sloan foundation, Bill and Melinda Gates Foundation, and MacArthur Foundation. The authors would like to also thank the numerous colleagues and organizations that helped make this work possible, including the Snohomish, Pierce, and Whatcom Counties court clerks and tenant attorneys; Edmund Whitter and the King County Bar Association; John Stovall and Washington State Representative Nicole Macri; Jim Bamberger and the Washington Office of Civil Legal Aid; the University of Washington eScience Institute, iSchool, Department of Sociology, and Center for Studies in Demography and Ecology—particularly, Bill Howe, Jose Hernandez, Ian Kennedy, Sarah Stone, and Kyle Crowder; and The University of California Berkeley Institute of Governmental Studies; D-Lab; Berkeley Institute of Data Science; and Karen Chapple and the Urban Displacement Project.

Authors

Tim Thomas, Ph.D., is research director at the University of California, Berkeley (UCB) Urban Displacement Project and director of the Eviction Research Network. Alex Ramiller is a Ph.D. candidate in city and regional planning at UCB. Cheng Ren is a Ph.D. candidate at UCB Department of Social Welfare and lecturer at the University at Albany, State University of New York. Ott Toomet, Ph.D., is an assistant teaching professor at the University of Washington.

References

Balk, Gene. 2023. “When People Move Away from Seattle, Here’s Where They Go.” *The Seattle Times*, April 4. https://www.seattletimes.com/seattle-news/data/when-people-move-away-from-seattle-heres-where-they-go/.

———. 2017. “New Residents Pour in: Pierce, Snohomish Counties See Nation’s Biggest Jump in Movers.” *The Seattle Times*, March 27. https://www.seattletimes.com/seattle-news/data/new-residents-pour-in-pierce-snohomish-counties-top-the-nation/.

Bates, Lisa. 2023. “Evicted in Oregon.” Portland State University. September 2023. https://www.evictedinoregon.com.

Benfer, Emily A., David Vlahov, Marissa Y. Long, Evan Walker-Wells, J. L. Pottenger, Gregg Gonsalves, and Danya E. Keene. 2021. “Eviction, Health Inequity, and the Spread of COVID-19: Housing Policy as a Primary Pandemic Mitigation Strategy,” *Journal of Urban Health* 98 (1): 1–12. https://doi.org/10.1007/s11524-020-00502-1.

Davidson, Michael Scott. 2019. “Despite Changes, Nevada Eviction Law Still Favors Landlords,” *Las Vegas Review Journal*, June 28. https://www.reviewjournal.com/local/local-nevada/despite-changes-nevada-eviction-law-still-favors-landlords-1697301/.

Desmond, Matthew. 2012. “Eviction and the Reproduction of Urban Poverty,” *American Journal of Sociology* 118 (1): 88–133. https://doi.org/10.1086/666082.

Desmond, Matthew, and Carl Gershenson. 2017. “Who Gets Evicted? Assessing Individual, Neighborhood, and Network Factors,” *Social Science Research* 62 (February): 362–77. https://doi.org/10.1016/j.ssresearch.2016.08.017.

Desmond, Matthew, Carl Gershenson, and Barbara Kiviat. 2015. “Forced Relocation and Residential Instability among Urban Renters,” *Social Service Review* 89 (2): 227–62. https://doi.org/10.1086/681091.

Desmond, Matthew, and Tracey Shollenberger. 2015. “Forced Displacement From Rental Housing: Prevalence and Neighborhood Consequences,” *Demography* 52 (5): 1751–72. https://doi.org/10.1007/s13524-015-0419-9.

Fowler, Katherine A., R. Matthew Gladden, Kevin J. Vagi, Jamar Barnes, and Leroy Frazier. 2015. “Increase in Suicides Associated With Home Eviction and Foreclosure During the US Housing Crisis: Findings From 16 National Violent Death Reporting System States, 2005–2010,” *American Journal of Public Health* 105 (2): 311–16. https://doi.org/10.2105/ajph.2014.301945.

Fowler, Patrick J., David B. Henry, and Katherine E. Marcal. 2015. “Family and Housing Instability: Longitudinal Impact on Adolescent Emotional and Behavioral Well-Being,” *Social Science Research* 53 (September): 364–74. https://doi.org/10.1016/j.ssresearch.2015.06.012.

Franzese, Paula A. 2018. “A Place to Call Home: Tenant Blacklisting and the Denial of Opportunity,” *Fordham Urban Law Journal* 45 (3): 38.

Gromis, Ashley, Ian Fellows, James R. Hendrickson, Lavar Edmonds, Lillian Leung, Adam Porton, and Matthew Desmond. 2022. “Supplementary Information Estimating Eviction Prevalence across the United States,” Eviction Lab. April 25. https://evictionlab.org/docs/Eviction_Lab_Methodology_Report_2022.pdf.

Hatch, Megan E., and Jinhee Yun. 2021. “Losing Your Home Is Bad for Your Health: Short- and Medium-Term Health Effects of Eviction on Young Adults,” *Housing Policy Debate* 31 (3–5): 469–89. https://doi.org/10.1080/10511482.2020.1812690.

Hepburn, Peter, Renee Louis, and Matthew Desmond. 2020. “Racial and Gender Disparities among Evicted Americans,” *Sociological Science* 7: 649–62. https://doi.org/10.15195/v7.a27.

Hepburn, Peter, Devin Q. Rutan, and Matthew Desmond. 2022. “Beyond Urban Displacement: Suburban Poverty and Eviction,” *Urban Affairs Review*, March, 107808742210856. https://doi.org/10.1177/10780874221085676.

“ImageMagick.” 2023. C. https://imagemagick.org/index.php.

Khanna, Kabir, Brandon Bertelsen, Santiago Olivella, Evan Rosenman, and Kosuke Imai. 2023. “Wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation.” R. https://github.com/kosukeimai/wru.

Legal Services Corporation. 2023. “Civil Court Data Initiative.” Data. Eviction Tracker. September. https://civilcourtdata.lsc.gov.

Mullen, Lincoln, Cameron Blevins, and Ben Schmidt. 2023. “Gender: Predict Gender from Names Using Historical Data.” R. https://github.com/lmullen/gender.

National Academies of Sciences, Engineering, and Medicine. 2023. *Addressing the Long-Term Effects of the COVID-19 Pandemic on Children and Families*. Washington, DC: National Academies Press. https://doi.org/10.17226/26809.

Pan, Yuliya, Sabiha Zainulbhai, and Tim Robustelli. 2021. “Why Is Eviction Data so Bad?” Washington, DC: New America. https://d1y8sb8igg2f8e.cloudfront.net/documents/Why_is_Eviction_Data_so_Bad.pdf.

Parker, Brenda, and Janet Lynn Smith. 2021. “Policy Spotlight: Women’s Housing Precarity During and Beyond COVID-19.” SSRN Scholarly Paper 3896504. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3896504.

Raymond, Elora, Sarah Stein, Victor P. Haley, Erik Woodworth, Gordon Zhang, R. Siva, and Subhro Guhathakurta. 2020. “Metro Atlanta Evictions Data Collective Database: Version 1.0.” School of City and Regional Planning: Georgia Institute of Technology. https://metroatlhousing.org/atlanta-region-eviction-tracker/.

Senate Committee on Housing Stability & Affordability. 2019. “Final Bill Report ESSB 5600.” Senate Bill Report C 356 L 19. Olympia, WA: Washington State Senate and House or Representatives. https://app.leg.wa.gov/billsummary?BillNumber=5600&Initiative=false&Year=2019.

Shinn, Marybeth, Andrew L. Greer, Jay Bainbridge, Jonathan Kwon, and Sara Zuiderveen. 2013. “Efficient Targeting of Homelessness Prevention Services for Families,” *American Journal of Public Health* 103 (S2): S324–30. https://doi.org/10.2105/AJPH.2013.301468.

“spaCy.” 2023. Python. https://spacy.io.

Tesseract-OCR. 2023. “Tesseract.” C++. Tesseract-OCR. https://github.com/tesseract-ocr/tesseract#license.

Thomas, Timothy. 2017. “Forced Out: Race, Market, and Neighborhood Dynamics of Evictions.” Thesis, Seattle, WA: University of Washington. https://digital.lib.washington.edu:443/researchworks/handle/1773/40705.

Thomas, Timothy, Ott Toomet, Ian Kennedy, and Alex Ramiller. 2020. “The State of Evictions: Results from the University of Washington Evictions Project.” Seattle, WA: University of Washington. https://evictionresearch.net/washington/.

U.S. Census Bureau. 2022. “More Than 19 Million Renters Burdened by Housing Costs.” December 8. https://www.census.gov/newsroom/press-releases/2022/renters-burdened-by-housing-costs.html.

Xie, Fangzhou. 2022. “Rethnicity: An R Package for Predicting Ethnicity from Names.” *SoftwareX* 17 (January): 100965. https://doi.org/10.1016/j.softx.2021.100965.