



---

## **Methodology Report**

**v.1.1.0 (5/7/18)**

# Table of Contents

Table of Contents	1
I. Introduction	2
I.1. How to Cite the Data and Methodology Report	3
II. Data Collection	4
II.1. Court Records	4
II.2. Data Acquisition	4
III. Data Cleaning	9
III.1. Reformatting the Data	10
III.2. Dating an Eviction Case	11
III.3. Commercial Cases	12
III.4. Name Standardization	13
III.5. Preparing Addresses for Geocoding	16
III.6. Separating Street Address from Apartment Designation	18
III.7. City, State, and Zip Code Cleaning	18
III.8. Geocoding and Linking to Census Geographies	21
III.9. Address Standardization	24
III.10. Duplicate Records	26
III.11. Serial Cases	26
IV. Estimating the Prevalence of Eviction	28
IV.1. Classifying Eviction	28
IV.2. Estimating Case Volume Fluctuation	28
IV.3. Imputation	32
IV.4. Aggregation	33
IV.5. External Data Sources	34
IV.6. Denominator	35
V. Validation of Estimates	36
V.1. State Data Merges	36
V.2. State-Reported County Statistics	37
V.3. Low and High Rates	39
VI. Future Research	43

# I. Introduction

In recent years, renters' housing costs have far outpaced their incomes, driving a nationwide affordability crisis. Current data from the American Housing Survey show that most poor renting families spend at least 50 percent of their income on housing costs.<sup>1</sup> Under these conditions, millions of Americans today are at risk of losing their homes through eviction.

An eviction occurs when a landlord forcibly expels a tenant from a residence. While the majority of evictions are attributed to nonpayment of rent, landlords may evict tenants for a variety of other reasons, including property damage, nuisance complaints, or lease violations. A formal eviction occurs when a landlord carries out an eviction through the court system. Conversely, an informal eviction occurs when a landlord executes an eviction without initiating a legal process. For example, a landlord may offer a buyout or perform an illegal lock-out. Until recently, little was known about the prevalence, causes, and consequences of eviction.

The Eviction Lab at Princeton University has collected, cleaned, geocoded, aggregated, and publicized all recorded court-ordered evictions that occurred between 2000 and 2016 in the United States. This data set consists of 82,935,981 million court records related to eviction cases in the United States between 2000 and 2016, gleaned from multiple sources. It is the most comprehensive data set of evictions in America to date.

These data allow us to estimate the national prevalence of court-ordered eviction, and to compare eviction rates among states, counties, cities, and neighborhoods. We can observe eviction trends over time and across geography, and researchers can link these data to other sources of information.

The Eviction Lab at Princeton University is directed by Matthew Desmond and designed by Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton, with research assistance from Henry Gomory, Samarth Gupta, Chase Hommeyer, Hyojin Lee, Sejin Park, Gilian Slee, and Haley Zeng. The Eviction Lab is funded by the JPB, Gates, and Ford Foundations as well as the Chan Zuckerberg Initiative. More information is found at [www.evictionlab.org](http://www.evictionlab.org).

The following report details the Eviction Lab's methodology. For further information, please email [research@evictionlab.com](mailto:research@evictionlab.com).

---

<sup>1</sup> Matthew Desmond, "Unaffordable America: Poverty, Housing, and Eviction," *Fast Focus: Institute for Research on Poverty* 22 (2015) 1-6.

## I.1. How to Cite the Data and Methodology Report

Citations for academic publications and research reports:

*Data:* Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. *Eviction Lab National Database: Version 1.0*. Princeton: Princeton University, 2018, [www.evictionlab.org](http://www.evictionlab.org).

*Methodology Report:* Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. *Eviction Lab Methodology Report: Version 1.0*. Princeton: Princeton University, 2018, [www.evictionlab.org/methods](http://www.evictionlab.org/methods).

Citations for media outlets, policy briefs, and online resources:

Eviction Lab, Princeton University, [www.evictionlab.org](http://www.evictionlab.org).

## II. Data Collection

### II.1. Court Records

In the United States, eviction cases are typically heard in civil courts at the county-level. Although information contained in court records varies between jurisdictions, most records include the date the eviction case was opened, the names and addresses of plaintiffs (landlords) and defendants (tenants), the resolution of the case, and the money judgment (if any). Depending on the jurisdiction, an eviction may be classified as a “forcible detainer,” “forcible entry and detainer,” or “unlawful detainer.”

Courts resolve most eviction cases in one of three ways:

- (1) an *eviction judgment* is rendered, ordering the defendants to vacate a premise by a specific date;
- (2) the case is *dismissed* or ruled in favor of the defendant, allowing the tenant to remain in the residence;
- (3) a *mediated agreement*, often called a “settlement” or “stipulation,” is established between a landlord and a tenant. In a mediated agreement, the tenant must meet certain terms or adhere to a negotiated payment schedule. If the terms are met, the eviction is dismissed; if they are not, an eviction judgment can be rendered.

Table 1 provides an example of the information included in a typical eviction record.

**Table 1.** Sample Eviction Record

Case Number	Defendant First Name	Defendant Last Name	Plaintiff First Name	Plaintiff Last Name	Street	City	State	Zip	Filing Date	Judgment Date	Outcome	Judgment Awarded
2016NYBRO-01	Jane	Doe	John	Adams	12 Main Street	New York	NY	12345	1/1/2016	4/1/2016	Find for Plaintiff	\$700

### II.2. Data Acquisition

The Eviction Lab created a centralized repository of U.S. eviction records by acquiring data from states and counties and purchasing records from two independent data acquisition companies.

We first contacted court clerks in all 50 states, requesting details about eviction records in their jurisdictions.<sup>2</sup> Depending on the staffing, the use of electronic data storage, and the responsiveness of individual courts, the time from initial contact to completion of our request was between one month (North Carolina) and ten months (North Dakota).

We collected eviction records en masse where they were available. As a result, we made requests of court systems for bulk reports. Bulk reports provided a corpus of a court's available eviction case information, viz., names, addresses, dates of eviction filings, and outcomes. Hundreds of thousands of records can be retrieved through a single bulk record request, streamlining data collection. The Freedom of Information Act (FOIA) extends the right to access eviction records; however, local policy barriers and resource limitations pose significant challenges to data acquisition. Some states, such as [Illinois](#), enforce policies preventing bulk data retrieval.

Through bulk requests, we obtained 12,870,070 individual-level, court-ordered eviction records from 13 states. Table 2 represents the number of records we received from each state and the range of years for which we have data.

The Eviction Lab also purchased data sets of public eviction records from two companies: LexisNexis Risk Solutions (LexisNexis)<sup>3</sup> and American Information Research Services Inc. (AIRS)<sup>4</sup>. We obtained 66,512,547 eviction records from LexisNexis, which included all states except North and South Dakota. LexisNexis' cases occurred between February 1970 and October 2017. AIRS provided 10,606,467 eviction records from six states: Arizona, California, New Jersey, Nevada, Ohio, and Oregon. AIRS's cases occurred between January 1988 and March 2018.

**Table 2.** Records Collected from State Courts

State	Years	Number of Observations
-------	-------	------------------------

<sup>2</sup> Not all records are collectible. Some eviction records may be sealed. For instance, California established policies allowing tenants to block public access to prior eviction court records. Wisconsin seals cases ending in a dismissal after two years, pulling them out of their database and destroying the case information.

<sup>3</sup> LexisNexis records were last updated on September 29th, 2017.

<sup>4</sup> AIRS data were last updated on March 9th, 2018.

Alabama	1977-2015	1,148,022
Connecticut	2006-2015	377,095
Hawaii	2006-2016	28,955
Iowa	1995-2015	374,419
Indiana	1983-2016	366,246
Minnesota	2009-2016	414,393
Missouri	2011-2015	534,775
Nebraska	1999-2015	857,687
New Jersey	1995-2015	1,638,451
Oregon	1990-2016	709,467
South Carolina	2006-2015	2,591,938
Pennsylvania	2006-2016	2,769,941
Virginia	2010-2015	1,058,681

---

Note: Pennsylvania data come from a combination of two data sets: Pennsylvania State Court data for all counties between 2006-2016, and Philadelphia County Court data from 1969-2016.

As of April 2018, the Eviction Lab has collected a set of 82,935,981 million court records related to eviction cases in the United States between 2000 and 2016. Not all records and cases were suitable for inclusion in our analysis. Duplicate records and non-meaningful records (e.g., blank records, test records) were removed. Table 3 shows a summary of the number of records we had in our database after dropping unsuitable cases. Numbers are reported between 2000 and 2016, as this is the timeframe we used for analysis and produced on the map.

**Table 3.** Relevant Records by Data Source (2000-2016)

<b>Data Source</b>	<b>Number of Records</b>
LexisNexis	63,677,688
State Court Data	10,895,619
AIRS	8,362,674

To retrieve these records, the companies utilized collection methods similar to our own (e.g., automated record collection, bulk record collection). In addition, individual records were collected in person from courts where bulk and online collection were not available. Examining paper court records granted access to name and address information available on both the petition and the final judgment, as well as allowed them to determine the outcome of the case. This included any money judgments related to landlord/tenant cases, dismissals, and instances where a tenant was ordered to vacate the premises. Record information was then consolidated through data entry. Not all available eviction-related records can be collected in person, given the time and resource-intensive nature of the practice. However, these data represent the most comprehensive attempts to collect court-ordered eviction records nationwide.

Our data include all landlord/tenant-related cases. Some landlord/tenant cases do not involve the return of a property to the landlord. For instance, some cases involve a disputed security deposit. To determine which cases resulted in eviction, the companies drew on court procedures, developing criteria for determining whether an eviction took place. Because the data come from independent sources, they contain additional information apart from the standard collected from states. Some states indicate whether the case involved possession of property. If the possession indicator was present, the case was classified as an eviction. Other states had more parsimonious case categorizations, indicating whether a money judgment, eviction, or dismissal took place.

Our team also gathered aggregate state-reported county-level data from 27 states, New York City, and the District of Columbia.<sup>5</sup> We attempted to collect reports of the number of

---

<sup>5</sup> These states include: Alaska, Arkansas, California, Colorado, Delaware, Florida, Georgia, Hawaii, Maine, Maryland, Massachusetts, Michigan, Missouri, Nebraska, Nevada, New Jersey, New Mexico, North Carolina, North Dakota, Ohio, Pennsylvania, South Dakota, Texas, Utah, Vermont, Virginia, and Wyoming.



landlord/tenant cases filed in counties within a state each year.<sup>6</sup> If the reports were not available, then our team made individual requests to each state. Many states did not maintain reports or retain landlord/tenant case information housed at the state-level.

We converted any PDFs or workbooks into CSV files, then wrote a Python program to rearrange and extract relevant eviction count information. We accomplished this by searching for any information in a report pertaining to an eviction-related case type and outputting it into a separate spreadsheet. In Ohio, Massachusetts, and Alaska, filings were reported by courts rather than counties. Our team classified all reported filings into counties (based on courthouse location) to maintain standard geographic units for comparison. Using a list of courts provided by LexisNexis, we were able to identify which county a court belonged to based on indicators in this list. If we could not determine this information from the list, we classified the court into a county based on the address of the courthouse.<sup>7</sup>

Only a handful of states maintained data on how landlord/tenant cases ended, but all had information on number of cases opened and/or filed in a given year.<sup>8</sup> All data from these reports were aggregated together to create a uniform output, which can be found in the data download section of our website. The file lists the year reported, the state and county, and the number of cases reported each year in the report.<sup>9</sup> Table 4 serves as an example output.

**Table 4. Example Aggregate Output**

<b>Year</b>	<b>State</b>	<b>County</b>	<b>State-Reported Cases</b>
2004	WI	Adams	24
2004	WI	Ashland	10
2004	WI	Barron	500
2005	WI	Adams	35

<sup>6</sup> Some states did not report case volume as “landlord/tenant.” The name for these types of cases varied by jurisdiction. We sought to collect all case types that fell under the relevant umbrella but did not include foreclosures.

<sup>7</sup> In some smaller areas, court jurisdictions may cross county lines. Our team was unable to determine the precise coverage area of each individual court, so we relied on the list of courts, as well as the seat of the county courthouse, to determine in which county to report the filings.

<sup>8</sup> Some states reported case volume by the fiscal year. Because we had no way to accurately determine the distribution of cases over the year, we report the year as the first calendar year in the reported fiscal year.

<sup>9</sup> For North and South Dakota, we received lists of case numbers for landlord tenant cases, which county the case pertained to, and filing dates. In these states we summed the number of cases in each year and reported the information in the same format.

2005 WI Ashland 7

2005 WI Barron 425

---

The Eviction Lab collected records from a variety of sources, with many overlaps in jurisdiction. Each source differed in its ability to collect and report eviction cases. For instance, had we reported only state-reported case filings we would have been unable to determine the eviction rate, or where evictions occurred within a county. Our primary objectives with our data and map were to promote comparability between areas over time, and to achieve geographic specificity when reporting. To this end, we used the most nationally comprehensive data source available, which is LexisNexis.

There are a handful of exceptions. Between 2007 and 2016 in Alaska, Arkansas, North Dakota, South Dakota, and Pennsylvania, we lacked sufficient individual-level eviction records to create estimates of eviction filings or evictions. Given this circumstance, we chose to supplement our statewide gaps with data from external sources. For Alaska, Arkansas, North Dakota, and South Dakota, we derived estimates from state-reported volume of landlord/tenant case filings. In Pennsylvania between 2007 and 2016, we included the individual-level state court data.

In the following sections, the counts found in all tables are based on individual-level data that we used to produce the estimates found on the website. Our data includes nationwide LexisNexis records between 2000-2016 – excluding the four states listed above – and Pennsylvania state court data between 2007-2016.

### III. Data Cleaning

The eviction court records in our data set were created for administrative purposes rather than research analysis. Accordingly, the quality of information and level of detail included in these records varied substantially across municipalities. The Eviction Lab’s data represent nearly every jurisdiction in the country, so it was essential to standardize each record to create accurate measures for comparison. It was similarly important to correct errors that occurred during data entry or collection. For example, we eliminated instances where court clerks created a duplicate record of the same eviction case and corrected clear instances of typing errors in addresses. This required thorough investigation and analysis of the data.

We used *regular expressions* in many of our data cleaning procedures. Regular expressions are sets of characters that can be used to search for a specified pattern within text. They rely on wildcard characters and other conventions that specify a character’s location within a string of

text and/or how often a character appears in a string of text. Using regular expressions allowed us to detect irregularities in text variables and to collect and rearrange pieces of information.

We implemented regular expressions in Stata, predominantly using the standard set of regular expressions commands:

```

    regexm() to find a specific pattern of text
    regexr() to find a specific pattern of text and replace it with another chunk of text
    regexs() to save a specific chunk of text identified using regexm()

```

To allow an expanded set of special and wildcard characters, we also used their unicode string equivalents: `ustrregexm()`, `ustrregexra()`, and `ustrregexs()`.

### III.1. Reformatting the Data

Renting households are our main unit of our analysis. We focus on households, rather than individuals, because not all household members are listed as defendants when eviction cases are filed in court. Court records do not reveal how many individuals, particularly children or other household members not listed on a rental lease, are affected by an eviction. In our data, each entry records an action between a defendant and a plaintiff associated with an eviction case. We received some data in full case form and some as individual actions within cases. Depending on when records were collected and our data source, we could see an eviction *filing* and an eviction *judgment* for the same defendant as two separate entries sharing the same case number and court identifier. Multiple entries could therefore be recorded under the same court identifier and case number, when there were multiple actions between the same plaintiff(s) and defendant(s), multiple defendants listed on the case, or a combination of both. We reformatted the data so that each record contained all information on defendants and action types pertaining to the case. In this format, each case represented one household. Table 5a below presents an example of data format upon receipt, and Table 5b presents our data format.

**Table 5a.** Example Format of Original Data

Case Number	Court Number	Action Type	Defendant First Name	Defendant Last Name	Plaintiff First Name	Plaintiff Last Name	Street	City	State	Zip
123	456	Filing	Jane	Doe	John	Adams	12 Main Street	Albany	NY	12345
123	456	Filing	John	Doe	John	Adams	12 Main Street	Albany	NY	12345
123	456	Judgment	Jane	Doe	John	Adams	12 Main Street	Albany	NY	12345

123	456	Judgment	John	Doe	John	Adams	12 Main Street	Albany	NY	12345
-----	-----	----------	------	-----	------	-------	----------------	--------	----	-------

**Table 5b.** Format of Reshaped Data

Case Number	Court Number	Number of Filings	Number of Judgments	Number of Defendants	Number of Plaintiffs	Street	City	State	Zip
123	456	2	2	2	1	12 Main Street	Albany	NY	12345

Similarly, we reformatted state court data so that each entry corresponded to a single case, each uniquely identified by a court identifier and case number.<sup>10</sup> We included 38,564,127 total unique cases in our analysis between 2000 and 2016.

Not all unique cases were included in the analysis. Cases were excluded for three reasons. First, we excluded cases that duplicated information contained within another case already present in the data. Second, we excluded cases with commercial defendants. Commercial defendants indicate that the eviction is being carried out against a business rather than renting household. Identification of commercial defendants is discussed in more detail below. Finally, we excluded cases that occurred before 2000 or after 2016. Case collection was more consistent after 2000, and 2016 is the most recent year for which we have complete data.

## III.2. Dating an Eviction Case

Nearly every action on an eviction case was associated with a date in the data provided by LexisNexis.<sup>11</sup> When the action was a new case filing, this date reflected the filing date. When the action was the dismissal of the case, this date reflected the dismissal date. When the action was the result of a judgment, such as awarding rent owed to a landlord or issuing a writ of restitution for the property, the judgment date was recorded. The Pennsylvania court data only included

<sup>10</sup> The state court data we received from Pennsylvania (this excludes data from Philadelphia County) were missing several key variables, including defendant names, plaintiff names, case numbers, and filing years. As a result, we had to determine which observations corresponded with which unique cases. Many of the cases in these data included a unique judgment number that associates with a unique case. For cases without a judgment number, we generated a unique case identifier using a combination of relevant variables: county name, court office code, judgment disposition date, judgment disposition, and claim amount. We then assigned this identifier to cases as a proxy for case numbers.

<sup>11</sup> There were 68 cases in the data that were missing action dates. These cases were excluded from the analysis.

judgment dates for cases.<sup>12</sup> There were no records of when the case was initially filed. The Philadelphia court data included both filing and judgment dates.

Of the 38,564,127 cases in the analytic sample, 35.1% of cases did not include a record of the initial filing date, while 38.1% of cases did not include information about what occurred between the landlord and tenant after the case was filed (i.e., no judgment or dismissal information was present). To calculate yearly filing and eviction rates, each case was assigned a fixed *date of record* that was used as a timestamp. We assigned the date of record using the *earliest action on a case*. In approximately 65% of cases this was the filing date. In the remaining cases, it was the date of the first judgment or dismissal. The date of record was used to aggregate cases over time. For consistency, cases are assigned to the year of their date of record. For example, if a case showed a filing date of November 30, 2010, and a first judgment date of January 5, 2011, that case is recorded as having occurred in 2010.

LexisNexis data also contained release dates that were populated if the entry recorded an action involving a “judgment release.” A judgment release is the official notice from the court that a judgment has been fulfilled. If a tenant is ordered to leave the property at the conclusion of an eviction case, a landlord may return to court to file a release confirming that the tenant vacated the property. The action dates associated with release entries reflect the date of the judgment the release satisfies. For consistency, when the only action present on a case was a judgment release, the action date, rather than the release date, was used to assign the date of record.

Finally, a case may also include a “vacated judgment.” A vacated judgment indicates that a previous judgment in the case was overturned. For these records, the action date corresponds to the judgment being overturned, and the release date indicates when that judgment was vacated. If an action date on a judgment corresponds to an action date of a vacated judgment that occurred later in the same case, the original judgment was marked as vacated. Vacated eviction judgments were not counted as evictions.

### III.3. Commercial Cases

Our data include evictions filed against households (residential cases) and businesses (commercial cases). A case is marked “commercial” if at least one defendant was identified as a commercial entity. Because our research examines eviction at the household-level, we excluded commercial cases from our estimates. We did not distinguish between individual and commercial plaintiffs, since both individual and commercial landlords evict residential households.

---

<sup>12</sup> Fewer than 1% of total cases (N = 822 cases) had multiple judgments dates associated with the same case. When this occurred, we selected the latest judgment date, it being most affiliated with the final outcome of the case.

We identified commercial cases in our data using two methods. First, where available, we used variables provided in the raw data that identified commercial cases. This indicator was available in 97.3% of cases in our data. We developed a master list of keywords to identify commercial cases in data without an existing indicator. This list includes common words clearly and consistently associated with businesses:

- **Legal terms for businesses:** e.g., Inc., Corporation, LLC
- **Types of organizations:** e.g., Church, School, College, Trust, Foundation
- **Types of services:** e.g., Laundry, Barbershop, Dentist, Financial
- **Types of products:** e.g., Pizza, Toys, Tires, Books, Goods

We added keywords to the master list as we encountered new ones. We then used regular expressions to determine whether defendant names included any of the keywords from the master list. We checked all data against the latest version of the master list each time it was updated.

We validated this second method using regular expressions by testing it on records that already had a commercial indicator. We found that the regular expressions procedure detected approximately 75% of the cases marked by the commercial indicator (where comparison was possible). This discrepancy was often due to situations in which the record appeared to list only the owner name without an accompanying business name. We ultimately relied on the commercial indicator and supplemented identification of commercial cases via the regular expressions method where the indicator was not available (0.58% of cases in our data).<sup>13</sup> In total, we identified and excluded 597,327 commercial cases in our data set; this amounted to 1.55% of all eviction cases.

### III.4. Name Standardization

Some defendant and plaintiff names appeared multiple times within eviction court records. However, the same name may not have always appeared in the same format. This can occur for several reasons. Sometimes, only first and last names appeared on a record; other times, middle names and initials were included. Additionally, entry errors are always present in administrative data. Identifying name variations was essential for determining the number of defendants and plaintiffs associated with each case and for identifying duplicate cases. To accomplish this, we developed a protocol for creating a standardized version of names.

#### **Plaintiff Names**

---

<sup>13</sup> We were unable to perform this procedure for Pennsylvania Court Data, which did not contain defendant names.

Plaintiffs listed as businesses can have non-uniform name entries. For example, a plaintiff may appear in the data as both “SUNNYSIDE APARTMENTS” and “SUNNYSIDE APARTMENTS LLC.” Accordingly, we removed common abbreviations for commercial entities, e.g., “LLC,” “INC,” and “CORP,” using regular expressions. We also removed common symbols and words that may have otherwise generated multiple entries for the same plaintiffs (e.g., “&”, “AND”).

We also standardized spellings and abbreviations for other words that were common to many plaintiff names. For example, the word “MANAGEMENT” may appear as “MGMT”, “MANAGMNT”, or “MANGAMENT.” For these cases, we created flexible search strings using regular expressions to find such variations and replace them with “MANAGEMENT.” The following is an example:

```
regexr(plaintiff_name, "(^| ) (MANAGE[MENT ]+|M[GMN] [GMN]?T) (| $) ", "MANAGEMENT").
```

Plaintiff names were then segmented into separate pieces using spaces (“ ”) as the delimiter. All name pieces were compared within the same street address and then within the same case number. If all name segments of a plaintiff name were contained within another plaintiff name appearing at the same street address or within the same case, these entries were assigned the same standardized name. Public Housing Authorities and University Regents were particularly likely to enter the data in many versions in areas with concentrations of subsidized or university housing, respectively. Local Housing Authorities were identified by searching for the occurrence of both “HOUS” and “AUTH” in plaintiff names; University Regents were identified using “REGENTS.” Both of these entities were then standardized within the same cases and addresses.

Finally, plaintiff names were compared within cases and addresses for minor variations in spelling. We calculated the Levenshtein distance between each pairwise combination of plaintiff names. Levenshtein distance represents the number of edits necessary to make one string match a second string. Edits can be deletions, insertions, or substitutions of any character that appears in the string. We used a threshold of two or fewer edits to determine that two versions of a name matched and should be assigned the same standardized name.<sup>14</sup>

## Defendant Names

We also searched for common spelling and format variations to check for duplicates and ensure consistency across defendant names. Initially our data had defendant names contained in two variables, one for the full defendant name and another for the name suffix (e.g., Jr, Sr, III). Defendant names were formatted as LASTNAME, FIRSTNAME MI (if present). We separated

---

<sup>14</sup> We implemented the additional requirement that names be at least seven characters long to be determined a match. This requirement was set to ensure that very short plaintiff names were not erroneously linked.

the names into three components: first name, middle name or initial (if present), and last name. We kept the name suffix separate.<sup>15</sup>

Like plaintiff names, we compared each pairwise combination of defendant names within the same case numbers and street addresses. The following rules were used to determine if two variations of name represented the same defendant with the same case number:

1. Exact match of first name, middle name, and name suffix, Levenshtein edit distance of two or fewer characters in the last name, when the last name was longer than two characters. This rule was used to catch minor spelling differences in last name.
2. Exact match of middle name, last name, and name suffix, Levenshtein edit distance of one character or less in the first name, when the first name was longer than one character. This rule was used to catch minor spelling differences in first names.
3. Exact match of middle name, last name, and name suffix, one version of the first name is contained within the other. This rule was used to catch nicknames or initials.
4. Exact match of first name and name suffix, one or less character Levenshtein edit distance in the last name and the two middle name fields start with the same letter. This identifies cases in which the full middle name was used instead of just the middle initial.
5. Exact match of first name and name suffix, the middle and last names are combined in one instance and separated in the other. This rule catches differences that occur when a second last name is entered in court records.
6. Exact match of first name and name suffix, one character or less Levenshtein edit distance in the last names and Levenshtein edit distance of one character in the middle name, when the middle name is longer than one character. This catches minor spelling variations in middle name.
7. Exact match of first name and name suffix, one or less character Levenshtein edit distance in the last name and one version missing the middle name. Sometimes the middle name is entered for a defendant; other times it is not.

These same rules were used to compare defendant names within the same street address, with the exception that Levenshtein edit distance between last names could only be one character or less,

---

<sup>15</sup> There were very few instances in which part of a defendant's first name was erroneously split into the middle name field due to a presence of a space. These were corrected by comparing concatenated first and middle names within cases. If the concatenated first and middle names matched another first name on the same case (and the last names and name suffixes matched), the first name and middle name fields were rejoined.



rather than two characters or fewer. This change accounted for the increased likelihood of two distinct defendants with similar names located at the same address over time as compared to being contained in the same case.

Each defendant name within the same case number or street address was compared twice. Comparing names twice allowed for standardized versions of names to be updated in sequence, and was more precise than creating less restrictive standards for matching two versions of a name. For example, the following three names may appear at “123 MAIN STREET, ANYWHERE, US”: “BUDDY HOLLY”, “BUDDY HOLY”, and “BUDDDY HOLY.” In the first iteration, “BUDDY HOLY” will be standardized to “BUDDY HOLLY” and “BUDDDY HOLY” will be standardized to “BUDDY HOLY”. On the second pass through, the standardized name for “BUDDDY HOLY” will be updated once more from “BUDDY HOLY” to “BUDDY HOLLY,” resulting in standardization across names within addresses. In total, 2.7% of names in our data were updated during these standardization procedures.

### III.5. Preparing Addresses for Geocoding

To pinpoint the location of an eviction, we geocoded the defendant address associated with each court case in our data. Cleaning addresses was essential to the success rate of our geocoding process. Geocoding involves matching street addresses against a standardized data set of street addresses and locations then assigning latitudes and longitudes to each address. Extraneous information (e.g., name of apartment complex, unit number), inconsistent formatting, and spelling errors can impede geocoding.

Some cases listed multiple addresses for defendants. In some cases, different addresses were associated with different defendants on the case. In other cases, multiple addresses were listed for the same defendant. This may occur if new addresses were entered for defendants after they left the landlord’s property but the case was still ongoing (e.g., the defendant may still be ordered to pay the landlord back rent). In 1.8% of cases from LexisNexis, a single eviction case contained multiple defendant addresses.

For this reason, an address associated with a new case filing, forcible detainer judgment (judgment for restitution of premises to the landlord), or case dismissal is most likely to correspond to the property from which the landlord sought to evict the tenant. In the LexisNexis data, 67.7% of case addresses were associated with filing records, 14.6% were associated with forcible detainer judgments, and 3.4% were associated with case dismissals. Only 14.3% of addresses were taken from entries associated with monetary judgments. In 0.4% of cases, no address was present for the defendant(s).

If multiple addresses were listed, we used the following criteria in the order presented to select the representative address for a case:

1. We selected the address associated with the original eviction filing, if present.
2. We selected the address associated with the forcible detainer judgment, if present.
3. We selected the address associated with the earliest case action, with preference given to non-missing, in-state addresses. If multiple addresses were associated with the earliest case action (and were non-missing and in-state), we selected the most common address. If further delineation was necessary, we gave priority to the addresses that were able to be geocoded.

We used this protocol to select the address with the highest likelihood of representing the location where the eviction occurred.

Due to anonymized records and lack of filing dates, we performed slightly different procedures in the Pennsylvania court data (except in Philadelphia County). Within a case with multiple addresses, we prioritized in-state non-missing addresses and randomly selected an address from these candidates. This procedure affected 1.7% of the total cases. In the Philadelphia state court data, only one address is associated with each case, so a procedure for selecting addresses was not needed.

A primary goal of the cleaning process was to ensure addresses were accurately split into five distinct fields: street, apartment, city, state, and 5-digit zip code. Some data arrived already split into these fields; other data were often incorrectly or incompletely split. We standardized all address data by capitalizing alphabetic characters, removing trailing and leading spaces, and reducing spaces within a string to one space each (e.g., “123 Main St” became “123 MAIN ST”).

We employed regular expressions to search for two capital letters followed by a 5-digit numeric sequence to detect cases where city/state/zip data was housed in the *street* or *unit* fields. We then employed additional regular expressions to place text within the appropriate field. Additionally, we used regular expressions to search for the presence of numbers to check the integrity of the *city* field. Numbers indicated either a nonsensical value or the presence of street, unit, or zip code information within the city field. We also removed punctuation, with the exception of apostrophes. We ensured that all *state* values consisted solely of two letters (e.g., AK, AL) and that all letter combinations matched a legitimate U.S. state abbreviation. Cases that did not, but where the true state could be identified (e.g., a transposition of letters), were spot corrected. *Zip code* cleaning consisted of verifying that the zip code field contained either 5 digits in a row or 5

digits followed by a dash and 4 additional digits. Non-numeric information was removed from this field. We left partial zip codes unchanged and shortened 9-digit zip codes to 5 digits.

### III.6. Separating Street Address from Apartment Designation

We employed regular expressions to remove unit information from the street field and place it in the unit field. We searched for unit designators, such as “unit,” “apartment,” “room,” “suite.” Below is an example of a regular expression used for this purpose:

```
regexpm(street, "(APA?R?TM?E?N?T?|GARAGE|SPACE|SU?I?TES?|RO?O  
?M|NO\\.?.|TRAILER) (# ?|NO\\.?. ?|NUMBER ?)?([0-9A-Z\\-]+)").
```

We also corrected cases where the data populating the street and unit fields were flipped, i.e., valid street address information ended up in the unit field or vice versa. We again used regular expressions to search for cases where the unit field began with a set of numbers followed by alphabetic characters and the street field did not begin with a numeric value. This pattern usually indicated the need for a street-unit swap. We performed spot checks and additional customization to refine the accuracy of the code.

Next, we used regular expressions to search for Post Office Box (PO Box) information within any address field. We removed PO Box numbers and placed them in a separate field to prevent erroneous matches to numeric street names during geocoding.

Additionally, we employed regular expressions to standardize street field abbreviations. Specifically, we replaced abbreviated versions of the following street designations with the full word: *avenue, boulevard, circle, country road, court, canyon, drive, expressway, glen, highway, lane, manor, place, parkway, road, rural route, state route, state highway, state route, state road, square, street, terrace, trail, valley, way*. We also expanded numeric suffixes like 1<sup>st</sup> and 2<sup>nd</sup> to read First and Second.

### III.7. City, State, and Zip Code Cleaning

We cleaned city names, state abbreviations, and five-digit zip codes associated with eviction court records via an automated comparison to a standardized listing of city names and state abbreviations associated with U.S. zip codes.<sup>16</sup>

---

<sup>16</sup> We did not apply this procedure on defendant addresses in the Philadelphia Court data. The original data we received had already undergone an address-cleaning procedure, in which addresses were cleaned and normalized

We conducted this standardized comparison listing using two sources: Zip Codes To Go and Zip Code Tabulation Areas. Zip Codes To Go ([www.zipcodestogo.com](http://www.zipcodestogo.com)) maintains an up-to-date listing of all zip codes used by the United States Postal Service, along with the preferred city name and state abbreviation associated with each zip code. The U.S. Census provides Zip Code Tabulation Areas (ZCTAs) that roughly align with USPS zip codes.<sup>17</sup> We combined the listings of city names, state abbreviations, and zip codes from both sources and created a listing of valid city name, state abbreviation, and zip code combinations in the United States.

We then compared each city name, state abbreviation, and zip code in the eviction records to the standardized listing. We used the MatchIt program in Stata to find the best standardized match for each record. MatchIt is a user-written program that allows inexact (or “fuzzy”) matching of two sets of string (or text) variables. We used bigram vectorial decomposition distance to match combined city name, state abbreviation and zip code variables to the standardized listing. Bigram vectorial decomposition creates a similarity measure for two strings by breaking each string into two-character pieces (bigrams) and then calculating how many of those bigrams are shared by both strings. This procedure allowed us to correct data-entry errors in city names, state abbreviations, and zip codes. The bigram vectorial decomposition algorithm created a similarity score for each set of strings in each of the court records and standardized listings.

After running the matching algorithm, we retained the standardized city, state, and zip code listing with the highest similarity score for each eviction record. We retained all standardized candidates if multiple matches tied with the highest similarity score. The highest possible similarity score was 1. A score of 1 indicated that a city name, state abbreviation, and zip code from an address in the eviction records perfectly matched a city, state, and zip listing in the standardized data set. Matches with a similarity score of 1 (95.6% of all defendant addresses) were marked as correct matches.

Because the matches between the eviction record addresses and the standardized listing of cities, states, and zip codes were inexact, some of the suggested standardized matches with similarity scores below 1 were incorrect. We used the following criteria to classify as correct standardized listing matches with similarities scores below 1:

---

using the Google Maps Geocoding API. Nearly all of the cases (99%) were assigned a recognizable address. Because of the high success rate in this address-cleaning process, we did not apply the city, state, and zipcode cleaning procedures in the Philadelphia Court data.

<sup>17</sup> ZCTAs do not align perfectly with USPS postal codes. ZCTAs are used for estimation of demographic statistics, as opposed to location identification for mail delivery and reflect other U.S. Census geographical boundaries (e.g., Census blocks) that are not considered in USPS assignment or realignment. ZCTAs may also include historical zip codes as well as secondary city names associated with a zip code in some areas.

1. A bigram string similarity score greater than 70% with a perfect match for the city name and zip code. In these cases, only the state abbreviation in the eviction record differed.
2. A bigram string similarity score greater than 70% with a perfect match for the state abbreviation and zip code. These matches represented misspellings in city names.<sup>18</sup>
3. A bigram string similarity score greater than 70% with a perfect zip code match. These matches represented minor misspellings in city names and incorrect state abbreviations.
4. A bigram string similarity score greater than 75% with a perfect match for city name, state abbreviation, and the final two digits of the zip code but a mismatch of the first digit of the zip code. Here, the only unmatched character was the first digit of the zip code. Owing to assignment of zip codes across the United States, first digits of zip codes are typically consistent with states.
5. A bigram string similarity score greater than 75% with a perfect match for city name, state abbreviation, and all but the second digit of the zip code.
6. A bigram string similarity score greater than 75% with a perfect match for city name, state abbreviation, and all but the third digit of the zip code. In some places, this may have represented a re-categorization of zip codes in an area over time.
7. A bigram string similarity score greater than 75% with a perfect match for city name, state abbreviation, and all but the fourth digit of the zip code.
8. A bigram string similarity score greater than 75%, with a perfect match for city name, state abbreviation, with the second and third digits of the zip codes flipped.
9. A bigram string similarity score greater than 75%, with a perfect match for city name, state abbreviation, with the third and fourth digits of the zip codes flipped.
10. A bigram string similarity score greater than 75%, with a perfect match for city name, state abbreviation, with the fourth and fifth digits of the zip codes flipped.

If the matches between the address records and standardized listing were marked as correct, we updated the cleaned address fields from the eviction records with the standardized cities, state abbreviations, and zip codes. City names, state abbreviations, and zip codes were updated for

---

<sup>18</sup> We performed initial checks during the matching process to ensure that a 70% similarity score provided a reasonable basis for this assumption. This score is conservative and misses some of the more egregious errors in spelling of city names and abbreviations.

1.75% of defendant addresses using these criteria. Standardized matches that did not meet any of the above criteria were not used to update the eviction address variables (2.70% of addresses).<sup>19</sup>

We developed the above criteria to maximize the number of true errors corrected without introducing changes that would distort the original address. We tested these thresholds on address data from LexisNexis records in three states in different regions of the country – Alaska, California, and Georgia – to ensure that they met our standard of correcting without distorting the data.

### III.8. Geocoding and Linking to Census Geographies

After we completed our address cleaning process, we submitted the files to Environmental Systems Research Institute (ESRI), for geocoding and linking to U.S. Census geographies. Geocoding assigned latitude (Y) and longitude (X) coordinates to addresses. These coordinates were based on spatial reference points in a locator file of U.S. street addresses. When possible, each address was linked to a standardized address in the locator file with the highest possible match score. The higher the match score, the greater the similarity between the text of the original address and that of the official address included in the locator file.

Addresses can be matched to latitude and longitude coordinates at varying levels of geographic specificity. The level of geographic specificity to which an address can be linked depends on how much information is included in the address, how closely it matches an address included in the locator file, and how accurately the locator file represents all possible addresses within a geographic area. We undertook the address cleaning steps described above to maximize the number of usable pieces of information in each address. ESRI used the most up-to-date locator files (published in 2016) to geocode all defendant addresses in our data set (N = 64,700,554).<sup>20</sup> The overall accuracy of locator files tends to increase over time, as U.S. spatial mapping is continuously improved.

Each address was matched to coordinates at the most specific level of geography possible. ESRI used an iterative geocoding process, designed to achieve the most specific match possible for an address. First, match attempts were made at the point address-level. If no standard address in the locator file satisfied the minimum match score, a match was attempted at the street address-level.

---

<sup>19</sup> We also compared city names and states abbreviations in cases where a match could not be found for city, state, and zip code. We used this measure to confirm the existence of a city within a state. Of the addresses that were not matched to a standardized city name, state, and zip code, we verified that 72.2% registered a valid city and state combination.

<sup>20</sup> An alternative strategy would have been to use versions of locator files that were published closer to the date of an eviction record. We choose to use only the 2016 locator files for all addresses, however, after discussing the improvement of accuracy in locator files over time directly with ESRI.

If no street address-level match was possible, a match was attempted at the street name-level, and so on.

Each level of geography has a minimum similarity score that must be met to assign a match. Table 6 lists the levels of geographic specificity, from most to least specific, along with the minimum required match score.

**Table 6.** Levels of Geographic Specificity in ESRI Geocoding

<b>Match Type</b>	<b>Description</b>	<b>Min. Match Score</b>
Point Address	The full street address, including house number, was matched with a standard address appearing in the locator file.	93%
Street Address	The street address was placed within a valid range of house numbers on the street. The range of house numbers was narrow enough to place an address on the correct side of a street within a typical street block.	85%
Street Name	The house number could not be matched to a specific point on a street and coordinates were only be assigned for a central point (centroid) along the street.	85%
Postal	Only the zip code was used to assign coordinates; these coordinates reflect only the central point (centroid) of the zip code.	100%
Administrative Place	Only the place name, usually a city or a municipality name, was used to assign coordinates; these coordinates reflect only the central point (centroid) of the administrative unit.	96%

The level of geographic specificity of a geocoded address is important because it affects our ability to count the number of evictions that occurred within small areal units, such as Census block groups. Geocodes at the point- and street address-levels are assigned coordinates that represent the exact or a very close approximation of the location of a street address, rather than a

centroid point associated with the street name, zip code, or other administrative boundary. At the point- and street-levels of specificity, we can be sure that an assigned Census block group reflects that of the address. As shown in Table 7, of the 64,700,554 records with available addresses, 93.4% were geocoded at either the point address or street address-level.

After an address was assigned latitude and longitude coordinates, its location could be pinpointed on a map and linked to other geographical areal units, such as Census tracts. We assigned each address both Census block group FIPS codes and Census place FIPS codes using Shapefiles provided by the U.S. Census Bureau. We assigned addresses to the Census block group or place that contained their latitudinal and longitudinal coordinates. We chose to use Census block group as the smallest areal unit because it is the lowest geographic level at which the U.S. Census reports income data. Census tract and county FIPS identifiers can then be directly inferred from Census block group FIPS codes.<sup>21</sup> For each geocoded address, then, we were able to potentially identify the Census block group, tract, place, county, and state where the eviction occurred. This geographic information allowed us to link eviction rates to demographic characteristics in a local area.

The 93.4% of addresses geocoded at the point- or street-level were assigned Census geography identifiers based on the above procedures. We reassigned the 6.61% of addresses that could not be geocoded at the point- or street address-levels to Census block groups using spatial imputation. Addresses that geocoded to out-of-state locations were also re-assigned to Census block groups within the county where the case was heard.<sup>22</sup> This occurred in 26,403 cases. These addresses were aggregated within counties then randomly assigned to a Census block group within the county using weighted probability proportional to its share of renting households within the county. Weighted probabilities were adjusted to reflect the number of eviction cases already positively pinpointed to households in that block group in the same year. Addresses were assigned iteratively and the weighted probabilities were updated after each assignment.

**Table 7. Geocoding Outcomes**

<b>Geocoding Outcome</b>	<b>Match Types</b>	<b>Number of Records</b>	<b>Percent of Total</b>
Address coordinates	Point, Street	60,413,361	93.7%

<sup>21</sup> Census blocks boundaries are nested within Census tracts, which are nested within counties. Place FIPS codes had to be assigned separately as their geographical boundaries do not perfectly align with block groups, tracts, or counties.

<sup>22</sup> Out-of-state cases were reassigned in this way as the address listed on the court record likely did not represent the location which the eviction was threatened or occurred. This address could have instead represented an alternative residence for the defendant or the new residence a defendant moved to after being evicted.



Centroid coordinates	Street Name, Postal, Administrative Places	4,277,614	6.6%
Unmatched	N/A	9,579	<0.1%
No Address present	N/A	160,043	0.2%

---

### III.9. Address Standardization

Even after separating apartment designations and cleaning city names, state abbreviations, and zip codes, some addresses still appeared in multiple variations in the data. To aid in identifying duplicated records and cases reappearing for the same tenant at the same address, multiple versions of the same address were assigned a standardized form within cases with the same defendant. This occurred when one or more of the following four patterns were observed.

1. The street address, state, and zip code of a defendant's address matched, but one or more entries had a different city name. In many of these cases, otherwise identical address records listed different neighboring city names. For example, "123 MAIN STREET, HENRICO, VA 55555" and "123 MAIN STREET, RICHMOND, VA 55555." When this was observed, all entries were updated to share the same city name. In choosing this standardized city name, preference was given according to the following ordered criteria (e.g., if not a, then b; if not b, then c):
  - a. We selected the city name associated with the address version with the most geographically specific level of geocoding.
  - b. We selected the city name associated with the address version that was able to be geocoded at any level of geography.
  - c. We selected the city name associated with the address version that was verified as a valid city name, state, and zip code combination in the city, state, and zip code cleaning.
  - d. We selected the city name associated with the most common version of the address.
2. The street address, state, and city name of a defendant's address matched, but one or more entries had a different zip code. As with variations in city names, this commonly occurred in bordering zip codes, and all otherwise similar addresses were updated to

share the same zip code. The same ordered criteria was used to assign the consistent zip code.

- a. We selected the zip code associated with the address version with the most geographically specific level of geocoding.
  - b. We selected the zip code associated with the address version that was able to be geocoded at any level of geography.
  - c. We selected the zip code associated with the address version that was verified as a valid city name, state, and zip code combination in the city, state, and zip code cleaning.
  - d. We selected the zip code associated with the most common version of the address.
3. The apartment designation in an address was appended to the street number: e.g., “123 MAIN STREET, APT 2B, ANYWHERE, US 55555” to “1232B MAIN STREET, ANYWHERE, US 55555.” We identified cases these by searching for instances in which both the street and apartment numbers associated with one address appeared within the street number portion of another address within the same defendant name. It was important to identify these cases because addresses with apartment designations were often more difficult to geocode. We updated incorrect versions using those with the correct separation of address components.
4. We synchronized multiple addresses with a shared format if multiple addresses shared the same street number and had street names that had either (1) a Levenshtein edit distance of one character or less or (2) the containment of one street name within the other (e.g., “MAIN” and “MAINE”) within the same defendant name. For the updated version of the address, ordered preference was given via the following criteria:
  - a. We selected the address with the most specific level of geocoding.
  - b. We selected the address able to be geocoded.
  - c. We selected the most common version of the address.

As addresses were updated according to these criteria, they were assigned the geocoded latitude and longitude coordinates and Census geography identifiers consistent with the version of the address used to do the updating.

As a final step, we created a standardized set of address variables – street address, city, state abbreviation, and zip code – for all address records. Those able to be assigned to point- or street-level locations were updated using the matched address provided by the ESRI geocode. For addresses geocoded at centroid points within streets, cities, or zip codes, we used the cleaned and

reconciled address variables to populate these fields. We updated 3.3% of the address records in our data during this standardization process.<sup>23</sup>

### III.10. Duplicate Records

We excluded duplicate records in our estimates to avoid over-counting. First, we dropped perfect duplicates (scenarios when the values of all variables were the same).<sup>24</sup> Then, we excluded duplicate records when all key variables – full defendant name, street address, city, state, zip, action date, action type, and judgment outcome and amount – contained the same value.<sup>25</sup> We considered these cases to reflect data-entry errors in which multiple records referred to the same case.

Some of these issues stemmed from human error, while others were the result of differing systems implemented by county courts throughout the state. To standardize case numbers, we used regular expressions, removing extra white spaces in the numeric entries. Table 8 lists the numbers, as explained above, that were dropped or excluded from our data. These records were neither used nor included in our estimates.

**Table 8.** Duplicates Dropped in All Eviction Records

<b>Duplicates Dropped</b>	<b>Duplicates Marked and Excluded</b>
8,590	697,188

### III.11. Serial Cases

A landlord may file multiple evictions against the same household. These “serial cases” differ from duplicates in that they reflect multiple (real) eviction filings by a landlord on the same

---

<sup>23</sup> We were unable to conduct the first four checks in this section on the Pennsylvania court data due to lack of defendant names. We were able to create standardized address fields using the geocoded matched addresses geocoded at the point- and street-level. The cleaned address fields were used in the standardized variables for those addresses only able to be geocoded at areal centroids.

<sup>24</sup> We were unable to fully perform these steps with Pennsylvania court data, which did not include defendant names, plaintiff names, or case numbers. We were only able to look for perfect (not near perfect) duplicates in these data.

<sup>25</sup> We cannot conduct these procedures for records missing defendant last name, street name, and/or anonymous defendants.

tenant. However, given that evicted tenants are forcibly removed from their homes, multiple repeated filings against within a short period of time suggest the tenant did not in fact vacate the premises. As a result, including all such cases would have overestimated the actual number of evictions. We first observed serial cases in eviction data received from South Carolina. To ensure these cases were not the result of administrative error, we spoke with court clerks and several legal aid representatives from South Carolina Legal Services and South Carolina Appleseed Legal Justice Center, all of whom confirmed the prevalence of serial cases.<sup>26</sup>

To properly identify and categorize serial cases for the purpose of counting evictions, we first standardized names and addresses using the criteria discussed above and sorted observations by defendant name, address, and action date. Next, we grouped cases sharing the same defendant name and address. These groupings of cases were categorized as one set of serial cases. We were unable to identify serial cases if names were missing or anonymized (e.g., Jane Doe, Unknown Tenant), or if an address was not listed for a tenant. The number of cases that were identified as the final case of a series of cases with the same defendant at the same property was 5,461,151.

Next, we referred to the outcome of the most recent case within a set of serial cases to determine whether an eviction appeared to have taken place. If it did, we counted one eviction for the whole set of serial cases. If the last case did not result in an eviction outcome, no eviction was recorded for the set of serial cases. If a set of serial cases included multiple eviction judgments and the last case in the set resulted in such a judgment, we counted that as a single eviction. If a set of serial cases included multiple eviction judgments and the last case in the set did not result in an eviction judgment, then no eviction was recorded for that set. Because serial cases involved multiple (real) filings, each filing was recorded in our final count of active eviction cases in a jurisdiction.

For example, if a landlord took the same tenant at the same address to court three times in three consecutive months – receiving an eviction judgment, case dismissal, then another eviction judgment (in that order) – we would record three filings and one eviction. If a landlord took the same tenant at the same address to court three times in three consecutive months – receiving an eviction judgment, a case dismissal, and another case dismissal (in that order) – we would record three filings and zero evictions.

---

<sup>26</sup> We identified serial cases in all data except Pennsylvania’s court data (excluding Philadelphia County), as the state’s bulk data collection rules prevented us from accessing defendant and plaintiff names.

## IV. Estimating the Prevalence of Eviction

### IV.1. Classifying Eviction

In our data, we classified a case outcome as an eviction if we could see that it contained a record of either a money or forcible detainer judgment for the plaintiff.<sup>27</sup> A money judgment awards the plaintiff some amount of money that is owed by the defendant.<sup>28</sup> A forcible detainer judgment grants possession of the property to the landlord. As discussed in Section III.2, judgments that were later vacated were not counted as evictions. Evictions in New Jersey and Philadelphia were defined differently, as detailed below.

We employed additional information on case outcomes in New Jersey using data from AIRS. The AIRS data contained individual-level court records for eviction cases, including judgment information. There was significant overlap between the cases contained in the AIRS data and cases contained in our data. We used court identifier, case number, and defendant name to perform a one-to-one merge linking our data to the AIRS data. We were able to update case outcomes in 3,431,883 records, representing 88.4% of records in New Jersey for the 2000-2016 period. Judgment data from AIRS indicated which party (plaintiff or defendant) a judgment favored and whether restitution of the premises was part of the judgment.

Philadelphia court data came with its own set of outcome codes. In that data, we counted a judgment as an eviction outcome if any of the following took place: (1) the defendant lost the case in a court trial; (2) the defendant lost the case by default; or (3) there was a stipulation and a writ of possession was granted as part of the judgment.

### IV.2. Estimating Case Volume Fluctuation

Sometimes, recorded eviction data are not uniform across space or time. In particular, there were areas where the volume of cases collected was substantially lower than expected, based on demographic variables and state court-reported volume. For this reason, not all of the records

---

<sup>27</sup> We did not separate settlements and stipulations from money or forcible detainer judgements.

<sup>28</sup> There may be situations in which a tenant is the plaintiff and the landlord is the defendant in a landlord/tenant dispute. For example, a tenant may sue their landlord for harassment and be awarded a monetary settlement. We cannot directly observe which party is the tenant and which is the landlord in our data. We checked for any situations in which a plaintiff listed on the initial case filing later became a defendant in the same case by looking for instances in which the parties appeared in the reverse order in any subsequent case actions. We only identified a handful of situations in which party names appeared in opposite fields, and these appeared to be data irregularities rather than a reversal of plaintiffs and defendants (e.g., the same name was listed as both the defendant and the plaintiff on a case).

collected for the 2000-2016 period were included on the Eviction Lab map. Our goal was to provide the most information possible about the spatial locations of eviction in America without presenting statistics that misrepresented the picture of eviction in a given area.

To help determine where coverage of eviction records was most consistent, we compared our data to eviction case volume statistics received directly from local and state courts. We received these data at the county-level for 27 states, New York City, and the District of Columbia, as discussed in Section II.2. We first calculated expected case rates and case volumes from year-to-year in the state-reported court data. We then compared these numbers with our case counts at the county-level. We used the county where the case was filed in this comparison, not the county where the premise address was geocoded.

We assigned each county to a broad *housing market* to better understand expected rates and changes in eviction case volume year-to-year in local areas. We used three different sets of characteristics to categorize housing markets: (1) Urban or Rural; (2) Northern or Southern; and (3) High- or Low-Renter Household Population.

**Urban or Rural:** We distinguished urban and rural counties using the U.S. Census Metropolitan Statistical Area (MSA) delineation files. For years 2007 or before, we used the 2007 MSA delineation; for years after 2007, we used the 2015 MSA delineation.<sup>29</sup> A county was marked “rural” if it was not included in any MSA, or if it was an “Outlying” county in an MSA.

**Northern or Southern:** After recognizing regional effects in housing market behavior (particularly in eviction rates), we assigned a dummy variable to distinguish Northern and Southern states.

**High- or Low-Renter Household Population:** We designated counties having “high-renter populations” if its number of occupied renter households was at or above the nationwide median. “Low-renter populations” were those with renter household populations below the median.

We calculated the annual expected percentage increase and expected percentage decrease in case volume for each housing market using state-reported statistics. We compared these values to the year-to-year percentage change in case volume in our data.

---

<sup>29</sup> U.S. Census Bureau (2007). [Core Based Statistical Areas \(CBSAs\) And Combined Statistical Areas \(CSAs\)](#), July 2007; U.S. Census Bureau (2015). [Core Based Statistical Areas \(CBSAs\), Metropolitan Divisions, And Combined Statistical Areas \(CSAs\)](#), July 2015.

We created four indicators based on these comparisons. The first marked the first year when data were available in a county. The second marked the last year data was available in a county. The third indicator marked years when case volume was substantially lower than the preceding year. The fourth marked years when case volume was substantially higher than the preceding year.

It is difficult to accurately distinguish between expected and unexpected changes when the number of cases is small. Accordingly, we did not include indicators or adjust case volume in counties that (1) were in the first quartile of renting households in the United States or (2) either had an average of fewer than 50 eviction cases per year or never had more than 100 eviction cases in any year.

All remaining counties were evaluated with the four indicators described above to estimate how case volume fluctuated over time.

We marked a year as having an unexpected *increase* in the case volume if it met all the following criteria:

1. The case volume increased by two or more standard deviations from the average observed increase in the same housing market.
2. The case volume increased 100% or more from the previous year.
3. There were at least ten cases recorded in the county in the previous year.

We marked a year as having an unexpected *decrease* in the case volume if it met all the following criteria:

1. The case volume decreased by two or more standard deviations from the average observed decrease in the same housing market.
2. The case volume had a decrease of at least 50% from the previous year.
3. There were at least ten cases recorded in the county in the previous year.

After marking an unexpected decrease, we also noted whether the data ever *recovered*. A recovery occurred when the case volume reached at least 75% of the volume in the year before the unexpected decrease occurred.

### **Marking First Available Year**

All counties began evaluation in 2000. We then sequentially *excluded* years according to the following set of criteria:

1. There was an unexpected increase in case volume the next year.
2. The data was low, relative to the average case volume. This was measured as an average case volume less than 75% of the standard deviation of case volume over time in that county.
3. A year recorded zero cases, and the following year recorded at least 10 cases.
4. A year recorded fewer than 10 cases, and the following year recorded zero cases. This was used to catch instances in our data when a handful of cases was collected in a year, but the long-term trend in the beginning years reflected a lack of overall coverage.
5. There was a 1,000% or greater increase in case volume in a future year. This detected counties that had very low case volume in early years, followed by a later increase in case volume that could only be attributed to a significant lack of complete collection coverage in the earlier years.

### **Marking Last Available Year**

Under a similar strategy, we began a second evaluation in 2016 and worked backwards to mark sequentially *previous* years as unavailable according to the following criteria:

1. There was an unexpected decrease in case volume compared to the previous year, and it was not a result of a temporary spike in case volume.<sup>30</sup>
2. Case volume was less than 75% of the standard deviation of average case volume over time in a given county.
3. There were no cases recorded in that county.
4. We observed fewer than 10 cases with zero cases in the preceding year. Again, this was used to catch lack of overall coverage in later years.

---

<sup>30</sup> A “temporary spike” in case volume was identified as an unexpected increase immediately followed by an unexpected decrease.



5. Case volume decreased by 90% or more in a preceding year. We used this to detect counties with more consistent coverage in early years followed by incomplete coverage in later years.

### IV.3. Imputation

Once the above trends were identified, we marked each year within a county to indicate whether it represented a valid data point. Years before the first available year for data within a county were not included in our data. Likewise, years after the last available year of data within a county were excluded. One exception was made for data in 2016. If a county had available data for 2015 but was missing data in 2016, we pulled the value from 2015 forward to estimate projected case volume in that county ( $N = 93$ ). In these cases, we anticipate updating the 2016 numbers during the next data delivery in 2018. These areas likely reflect counties that have adequate collection of cases but were not fully updated when the data were delivered in September 2017.<sup>31</sup>

After determining first and last available years, we marked years within the available period that had inconsistent case volumes, relative to other years in that county, and thus were unlikely to represent valid data points. We marked unavailable years sequentially until case volume recovered.

Within counties, years that were not marked as valid data points between the first and last years of availability were imputed if there were no more than two consecutive years of missing data. When only one year of data was missing within a county between two years of valid data, the case volume was imputed using the average of the preceding and following years. When two consecutive years of data were missing, we linearly interpolated between the last known and reliable value and the next known and reliable value. A total of 318 counties had at least one year of data imputed at some point in the 2000-2016 period in accordance with these two procedures. Some counties had data imputed in multiple years, resulting in a total of 519 imputed data points. When three or more consecutive years of data were missing, we did not impute the data; thus, these years were not included on the map.

Counties had to have at least two consecutive years of valid data to be included. Counties that did not meet this requirement are shown as missing. In total, 171 counties are displayed as missing for all years. The counties and years marked as missing are displayed as "Unavailable" on the map. Four states – Alaska, Arkansas, South Dakota, and North Dakota – did not have consistent data coverage at any point in the 2000-2016 period. For these states, we substituted data on case volume from the state-reported court statistics, as discussed in Section II.2. No counties in these states were included in the imputation procedure.

---

<sup>31</sup> There is often a lag time between when a case is filed within a court and when it can be recorded.

For counties with data adjusted during the imputation, additional cases were assigned Census block group FIPS codes using weighted random assignment. Here, we followed the re-assignment procedure as described in Section III.8. Each Census block group within a county was assigned a weighted probability consistent with its share of renter households. This weighted probability was updated to adjust for the number of cases successfully geocoded in the block group. Additional cases were then assigned randomly using these weights. Cases were assigned to block groups iteratively. After a case was assigned, the block group weights were re-estimated to account for this new assignment and then the next case was assigned. We followed this procedure until all adjusted cases were assigned to block groups within the county.

## IV.4. Aggregation

Eviction cases in our data were aggregated yearly in Census block groups, Census tracts, counties, Census places, states, and nationally across the United States. Once the data were cleaned, each case was marked as representing an active eviction case (filing) and/or an eviction judgment in the year that corresponded with the date of record assigned to the case. Cases contributed to the count of active cases so long as (1) the property at dispute was residential (rather than commercial) and (2) the case did not duplicate a previous case already represented in the data. Cases were additionally counted as resulting in evictions if the case included one of the judgments related to eviction criteria described above, and the judgment was not vacated at a later date. If a case was included in the eviction count, it was also represented in the pool of active eviction cases. There are no situations in which a case is marked as an eviction but not as an active eviction case pending that year. There are many instances, however, in which an active case did not contain a record of eviction.<sup>32</sup> As noted above, judgments on cases that recurred for the same defendant at the same property were not counted as evictions unless they appeared as the final action for a tenant at that property.

Although our data span 17 years, we standardized all geography to reflect 2010 Census boundaries. This allows us to compare the prevalence of eviction cases and judgments within the same area over time. As the Census geography boundaries are redrawn after each decennial Census, particularly at the block group- and tract-levels, we could not compare these areas directly without a standard set of geographical boundaries. For this reason, we had all geocoded eviction cases matched to 2010 Census block group and place FIPS codes.

Aggregation then involved counting the number of active cases and eviction judgments within each level of geography. Cases and judgments were first summed within Census block groups

---

<sup>32</sup> Some of our data sources prioritized cases that ended in a civil judgment. For example, in some areas, dismissals were not recorded, which deflated the total filings numbers. When this happens, filing and eviction rates are similar.

using Census FIPS codes, our smallest areal unit of geography. Counts were then aggregated into Census tracts, as blocks groups are contained within Census tracts. Counts in Census tracts were then summed within counties and counts within counties were aggregated to states. As a final step, all counts within states were collected into an estimate of the number of cases and evictions that occurred nationally each year between 2000 and 2016.

For aggregation we used the county derived from the geocoded address, rather than the county representing the court the case was heard in. We used the geocoded county as this was the best representation of the local area in which the case occurred. As mentioned in Section III.8, cases that geocoded outside of the state the court was located in were reassigned to a Census block group located in the county corresponding to where the case was heard.

Census places were aggregated separately, as their boundaries do not directly correspond to those of Census block groups, Census tracts, or counties. However, Census place boundaries correspond to Census *block* boundaries. Accordingly, after cases were assigned to a Census block group, they were allocated proportionally into Census blocks within that block group using household population as the weighting variable. We obtained allocation factors for each 2010 US Census block group from the [Missouri Census Data Center](#). After cases were allocated to Census blocks, they were aggregated into 2010 Census places.<sup>33</sup> For this reason, the counts of renter households, eviction cases, and evictions for places are often non-whole numbers (e.g. 4.5 evictions). On the map, these numbers are displayed as rounded whole numbers (e.g. 2.3 is displayed as 3 and 4.5 is displayed as 5). All values less than 1 but greater than 0 are displayed as 1. The filing and eviction rates are calculated with the original, non-rounded values. We chose to calculate the rates from the non-rounded values in order to not inflate our estimates. The original, non-rounded values of the eviction case and eviction counts are available in the downloadable files.

## IV.5. External Data Sources

---

<sup>33</sup> Our estimates for Pennsylvania evictions and filings were derived from multiple sources: LexisNexis (2000-2006) and state court data from 2007 onward, when Pennsylvania limited automated, bulk data collection. Because state court data was anonymized and did not include case numbers (as explained above), we could not check for serial nor duplicate cases in these data. Therefore, we used the available LexisNexis data to inform our post-2006 eviction estimates. We calculated the average filing-to-eviction ratio per block group, based on LexisNexis' Pennsylvania records between 2000 and 2006; for block groups that were in the state court data, but not in the LexisNexis data, we used the average filing-to-eviction ratio at the tract-level. We then multiplied this ratio to the state court data's aggregate case count, rounding down the estimates to the nearest integer, to get post-2006 eviction estimates at the block group-level. We applied these procedures specifically for Pennsylvania because we had sufficient data from LexisNexis – verified using published court statistics – to infer eviction counts in the state court data.

As mentioned previously, Alaska, Arkansas, North Dakota, and South Dakota estimates are derived from sources outside of LexisNexis. To address this, our team compared the volume in our individual-level data against the state-reported county statistics to determine an appropriate adjustment to the county statistics volume.<sup>34</sup> We calculated the ratio of county-level case filings found in LexisNexis data to the county-level case filings reported by the state. Then, using the regions used to create the markets described in Section IV.2, we found that in the North, on average, our estimates of eviction case filings were 72% of the volume reported by the state. In the South, this figure was 71%. For estimation purposes, Alaska, North Dakota, and South Dakota were classified as Northern, while Arkansas was categorized as Southern.

As a result, our estimates are the multiplication of the state-reported county case volume by its corresponding market adjustment. For example, if Adams County, North Dakota reported 100 landlord tenant cases in 2016, our estimate of case volume would be 72 eviction filings that year. Estimates found in the data download section and on the Eviction Lab map include these adjustments.

## IV.6. Denominator

To calculate eviction and filing rates, we divided the number of evictions and filings by the number of renter-occupied households in that area. We used estimates of renter-occupied households from the 2000 and 2010 U.S. Censuses and ESRI Business Analyst 2016. We first created renter-household estimates for Census block groups and then aggregated these numbers to create estimates for Census tracts, counties, and states.

The boundaries of 2000 and 2010 Census block groups are not directly comparable. For this reason, we used the Time Series Tables that aggregate 2000 Census demographics into 2010 Census geographies available from the National Historical Geographic Information System ([NHGIS](http://nhgis.org)).<sup>35</sup> We then created yearly estimates by using linear interpolation between the 2000 and 2010 data points. We used this same procedure to create yearly estimates between the 2010 Census and ESRI Business Analyst 2016.<sup>36</sup>

---

<sup>34</sup> While Pennsylvania state court data are used in our estimates for 2007-2016, we did not institute the same penalties. Using court reported statistics for Pennsylvania, we found that the proportion of cases to reported filings was comparable for 2000-2006 using LexisNexis and 2007-2016 using state court data.

<sup>35</sup> Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. *IPUMS National Historical Geographic Information System: Version 12.0* [Database]. Minneapolis: University of Minnesota. 2017. <http://doi.org/10.18128/D050.V12.0>

<sup>36</sup> ESRI Business Analyst 2016 was standardized to 2014 Census geography, which is consistent with the 2010 Census at the block group-level.

To estimate the number of renter-households in Census places, we had to create separate estimates. This time, we started with renter-households estimates from the 2000 Census at the 2010 Census block-level from [NHGIS](#). Then we aggregated 2010 blocks to 2010 Census places. We obtained 2010 Census place renter-household estimates directly from the Census. We allocated the 2016 block group-level estimates from ESRI Business Analyst down to the block group-level, again using files available from [Missouri Census Data Center](#). These estimates were aggregated into 2010 Census places as well. We used linear interpolation to create yearly estimates between 2000 and 2010 and between 2010 and 2016.

## V. Validation of Estimates

We validated our data by gathering additional information on evictions, both at the individual- and aggregate-levels. This included two validation methods at the individual-level and one at the aggregate-level.

### V.1. State Data Merges

We merged and compared LexisNexis data with state court data as a validation measure at the individual-level. These included records from Alabama, Hawaii, Iowa, Minnesota, Missouri, Nebraska, South Carolina, Pennsylvania, and Virginia.<sup>37</sup>

We used case numbers and court numbers, the combination of which uniquely identifies each case, to merge LexisNexis data with state court data. We considered a case “matched” when two or more of the following variables corresponded in the LexisNexis data and state court data: defendant names, plaintiff names, court dates (filing dates, disposition dates, and/or judgment dates) and defendant addresses. According to these criteria, over 99% of LexisNexis cases accurately represent cases in the state court data.

Using the state court data, we also wanted to compare the outcomes found on the cases in LexisNexis data to another source of data. We attempted to parse the text of the outcome and case information to classify cases in our state court data. However, case disposition meanings lack uniformity across states and sources. In these cases, we searched for additional information contained in our data. For example, Nebraska, Philadelphia, and Virginia case records contained fields indicating whether a writ of restitution was executed. A writ of restitution authorizes a landlord to forcibly remove a tenant from the disputed property following a judgment for the plaintiff. If a writ of restitution was executed, we considered the case to have resulted in eviction.

---

<sup>37</sup> We did not perform this check in Pennsylvania because the date ranges of data available in both LexisNexis and State Court Data did not overlap. We were only able to do so with Philadelphia County, PA, because we received Philadelphia case records from a separate source.

Because information on writs was available in Nebraska and Virginia, we chose to compare data from these two states against LexisNexis data. Nebraska court data specify whether a judgment includes “Restitution of Premises,” and Virginia court data contain information on whether a possession was at stake or if a writ of possession was ever issued.

Our comparisons showed that in Nebraska, 97.9% of cases marked as having an eviction outcome in LexisNexis would also have been marked as such in the court data; in Virginia, 93.9% of cases marked as an eviction outcome in LexisNexis were marked as such in the court data. In Nebraska, non-eviction outcomes in the LexisNexis data (0.02% were dismissals and 8.43% did not have further outcomes collected) were marked as evictions in the state data; in Virginia, non-eviction outcomes in the LexisNexis data (0.23% were dismissals and 11.8% did not have further outcomes collected) were marked as evictions in the state data.

## V.2. State-Reported County Statistics

To validate our estimates of eviction case volume, we compared our counts directly to another source: the state-reported county-level statistics on eviction filings. In some jurisdictions, certain counties were not reported in the state aggregate report; if we did not have state-reported statistics for these areas, they were not included in our calculation.<sup>38</sup>

Table 10 represents the average ratio of the number of cases within our data to the number of cases reported by the state over all years with state-reported statistics at the state-level. This table does not include comparisons for states where we substituted in the state-reported court statistics to create our estimates of eviction filings. These ratios can vary within state and across years. For instance, Massachusetts’ case volume in our data increases over time, so in later years our estimates of eviction case filings are much closer to the reported volume by the state.

It is also important to note that we can only compare case volume for years and counties in which we have additional data sources. In some states, while we believe we see low counts in certain jurisdictions, we have no external numbers to validate. Data collection is ongoing – we will continue to gather more resources to help us validate our national estimates of eviction filings.

### **Table 9.** Ratio of Aggregated Individual-Level Cases to County-Level Cases

---

<sup>38</sup> Missouri lists out separate case types related to landlord/tenant. While two case types – *rent and possession* and *unlawful detainer* – were reported across all jurisdictions, landlord/tenant cases were not reported uniformly. We compared case volume for what we could see. As a result, we have more individual cases than are technically publically reported by the state. This results in the proportion for Missouri being higher than 100%.

<b>Region</b>	<b>Ratio</b>
California	0.42
Colorado	0.65
District of Columbia	0.57
Delaware	0.81
Florida	0.83
Georgia	0.85
Hawaii	0.07
Massachusetts	0.60
Maryland	0.13
Maine	0.70
Michigan	0.72
Missouri	1.08
North Carolina	0.82
Nebraska	0.84
New Jersey	0.92
New Mexico	0.86
Nevada	0.74
New York City	0.37

Ohio	0.94
Pennsylvania	0.94
Texas	0.86
Utah	0.81
Virginia	0.90
Vermont	0.05
Wyoming	0.25

---

*Note:* For the 23 states listed above, the District of Columbia, and New York City, we compared the case volume of aggregated individual-level eviction cases to case volumes reported by the states at the county-level. For example, ratio of aggregated individual-level cases to county-level cases in Ohio is 0.94, meaning that there were 94 individual-level cases for every 100 reported by the state. New York City represents the five counties housed in New York City: Kings, Queens, Bronx, Richmond, and New York.

## V.3. Low and High Rates

### Low Rates

As noted in Section V.2 and on the Eviction Lab map, we have indicated states with underestimated eviction counts. Some places have naturally low eviction rates because there are few renters living there or because there are very few evictions. Still, some states and the District of Columbia have underestimated eviction counts for a variety of other reasons.

In New York, records are often kept as “abstracted judgments,” meaning they are only in the public record if the plaintiff/landlord pays to have them placed there. That, plus the amount of town and village courts in the state, makes collection difficult.

In California, many cases that end in eviction are sealed and therefore not accessible by the general public. Moreover, it can be difficult to collect data from California as a whole, owing to restrictions on the number of records one can collect.

In most jurisdictions, the eviction process starts with an out-of-court notice delivered to a tenant, but in Maryland the process begins with an eviction filed in court. This means Maryland has a very high case volume. Because of that, it is difficult to collect data from Maryland, with the



exception of Prince George's county where we have consistency in case volume over time. (Thus, Maryland is in the ironic position of having an inflated eviction filing rate and an underestimated eviction rate.)

In New Jersey, while the number of eviction cases was collected reliably, information about the outcomes of those cases were not readily accessible. As a result, we believe our eviction rate there is an underestimate.

In Kentucky, Louisiana, Tennessee, and Texas, there is robust data coverage in urban centers, but there are other, more rural areas that are missing data, in light of collection difficulties.

After comparing our estimates to external data sources, we found that Hawaii, Connecticut, Vermont, Wyoming, and the District of Columbia have low counts in the Eviction Lab map and raw data, due either to the remoteness of some areas or to data collection difficulties. We also suspect that the numbers in Arizona, Idaho, New Hampshire, and Washington may be too low. To identify low case rates, we created an indicator that factored in renter households and external statistics.<sup>39</sup>

We also marked counties with low, or underestimated, eviction counts on our map. We used the individual-level data collected directly from state courts (Section II.2.) and the state-reported county statistics (Section V.2.) as a comparative external source of data. We assessed the breadth of coverage in counties each year by looking at the difference in the number of reported eviction cases in our data compared to the external source. We applied the following criteria to determine when case volume in a given county was low:

1. The proportion of external cases represented in our data was less than 85%.<sup>40</sup>
2. The absolute difference in the number of cases between our data and state or county court reported data was greater than 32 cases.<sup>41</sup>

<sup>39</sup> We generated external case rates using external court statistics, then found the tenth percentile of the external case rates in each renter household quartile each year. We marked our case rate estimate in that year as “too low” if it fell below the tenth percentile of the corresponding quartile. We placed a flag for a state if 25% or more counties in the state were identified as “too low” between 2009 and 2016.

<sup>40</sup> The difference in number of cases between our data and the external data source was right skewed (i.e. most counties had low to moderate differences in case volume while a few outliers had more extreme differences). For this reason, the average level of coverage across counties appeared to be less than 65% of cases. We then restricted our comparison to only those counties that were within 50% coverage of the external case counts to account for this. The median value for coverage after imposing these restrictions was 85%.

<sup>41</sup> This represented the 75<sup>th</sup> percentile in the absolute difference in the number of cases in our data and the external data when our coverage was greater than or equal to 85% of external data.

Counties that met both these criteria were marked as having a low case count on our map. In many instances, the counties marked as having low case counts are located in the states discussed above. However, coverage across counties also varies within states. In some cases, comparisons to the external data showed that some counties were well represented in our data, even if the state overall appeared to have low coverage. In other instances, we discovered that evictions appeared to be underrepresented in a few counties within states that were well represented overall in the data.

We used an alternative strategy to mark counties with low eviction case volume when direct, external sources of data from state or county courts were not available. We started by calculating “external” eviction case rates in all the counties where external data was available. We then grouped these counties into quartiles based on the number of renting households. In each quartile, we examined the distribution of external case rates to determine what a typical eviction case rate should look like in counties with comparable numbers of renting households. We used these distributions for the case volume comparison in the counties for which no direct, external source of data was available. We identified counties with a case rate in the 20<sup>th</sup> percentile or below of the external case rates within their renter household quartile each year as having a low eviction case count.<sup>42</sup> Since it is possible that few evictions occurred in counties with low numbers of renter-occupied households, we did not mark counties in the lowest (first) quartile of renting households. Table 10 lists the percentage of counties in a given year that were either flagged as low in our data or marked as “Unavailable” on our map. We also marked areas embedded in counties (e.g., Census tracts, block groups) if their corresponding counties received flags.<sup>43</sup> However, we only validated low counts at the county and state levels. Therefore, Census tracts and block groups that received a flag due to their location in low count counties may not in fact be undercounted.

**Table 10.** Percentage of Counties Marked as Having a Low Count of Cases or Unavailable Data, by Year

<b>Year</b>	<b>Not Marked</b>	<b>Low</b>	<b>Unavailable</b>	<b>Total</b>
2000	47.12	16.00	36.88	100
2001	57.46	18.80	23.74	100

<sup>42</sup> We chose the 20<sup>th</sup> percentile threshold by calculating the median case rate in the counties that were marked low when external data was available for direct comparison. The median case rate in those counties was consistent with the 20<sup>th</sup> percentile of external case rates.

<sup>43</sup> We also assigned flags to Census Designated Places (CDPs). Unlike Census tracts and block groups, CDPs are not always fully contained within counties. Therefore, we used area allocation factor files from the [Missouri Census Data Center](#) to assign CDPs (or parts of CDPs) to counties. When a CDP was contained entirely within one county, as was the case for 96.6% of CDPs, we added a low count flag to the CDP when we had also added one to the county. We added low count flags to the remaining 3.4% of CDPs if 15% or more of the CDP’s area was contained within a low count county. For example, if a CDP’s area was split evenly among three counties (i.e., approximately 33% of the CDP’s area was contained in each county) and we marked one of those counties with a low count indicator, then we also marked the CDP as having a low case count.

2002	63.25	17.72	19.03	100
2003	63.98	18.29	17.72	100
2004	64.91	17.37	17.72	100
2005	64.91	17.79	17.31	100
2006	65.22	17.50	17.28	100
2007	60.90	21.22	17.88	100
2008	62.04	21.13	16.83	100
2009	64.52	19.15	16.32	100
2010	65.45	21.60	12.95	100
2011	65.03	21.99	12.98	100
2012	65.51	21.57	12.92	100
2013	64.84	22.14	13.01	100
2014	63.98	22.34	13.68	100
2015	66.47	18.84	14.70	100
2016	63.12	23.10	13.78	100
<b>Average</b>	62.87	19.80	17.34	100

## High Rates

The Eviction Lab map also indicates Census tracts and block groups with filing and eviction rates in the top 1%. Filing and eviction rates are calculated by dividing the number of filings or evictions in an area by the number of renter homes. That means that very high rates could be the result of (1) a large number of evictions or filings or (2) a small number of renter homes in an area.

High eviction and filing rates should be interpreted with caution, for a few reasons. First, they may be driven by small denominators found in areas with few or underestimated renter homes. For instance, some Census tracts incorporate both industrial zones and a small residential section. If the residential section had 40 renter homes, and 20 households were evicted residential section that would generate an eviction rate of 50%. Here, the eviction rate is high because the number of renter homes in the area is small. A residential neighborhood surrounding a college with 10 estimated renter homes and 20 evictions would report an eviction rate of 200%. In this case, the high eviction rate is explained by the fact that the Census often undercounts seasonal renters (like college students) in its estimate of the number of renter homes in an area.

In most jurisdictions, the eviction process starts with an out-of-court notice delivered to a tenant. In Maryland, however, the process begins with an eviction filed in court. Many landlords file against their tenants every month, resulting in a very high case volume. Here, the number of filings is inflated because of unique court procedures, resulting in a filing high rate, but not necessarily a similarly high eviction rate.

## VI. Future Research

The Eviction Lab's efforts have produced the most comprehensive data set of evictions in America to date. We have published aggregated eviction data for years 2000 to 2016 at the Census block group-, tract-, place-, county-, and state-level. Our data are publicly-available at [www.evictionlab.org](http://www.evictionlab.org).

The Eviction Lab will continue our endeavor to provide the most comprehensive data on evictions in America and we invite you to stay tuned for updates to our data set by signing up for our email list at [www.evictionlab.org](http://www.evictionlab.org). We will continue to update our data set and this Methodology Report as new data are collected and as we improve upon our methods. At our website, researchers interested in merging data at the individual-level can also submit a [Data Merge Application](#).

If you have eviction data we do not or if you use our data, please contact us and tell us about your findings by emailing [research@evictionlab.org](mailto:research@evictionlab.org). Fully understanding America's housing crisis, including the dynamics of eviction, requires a collective and interdisciplinary endeavor.

v.1.1.0 (5/7/18)