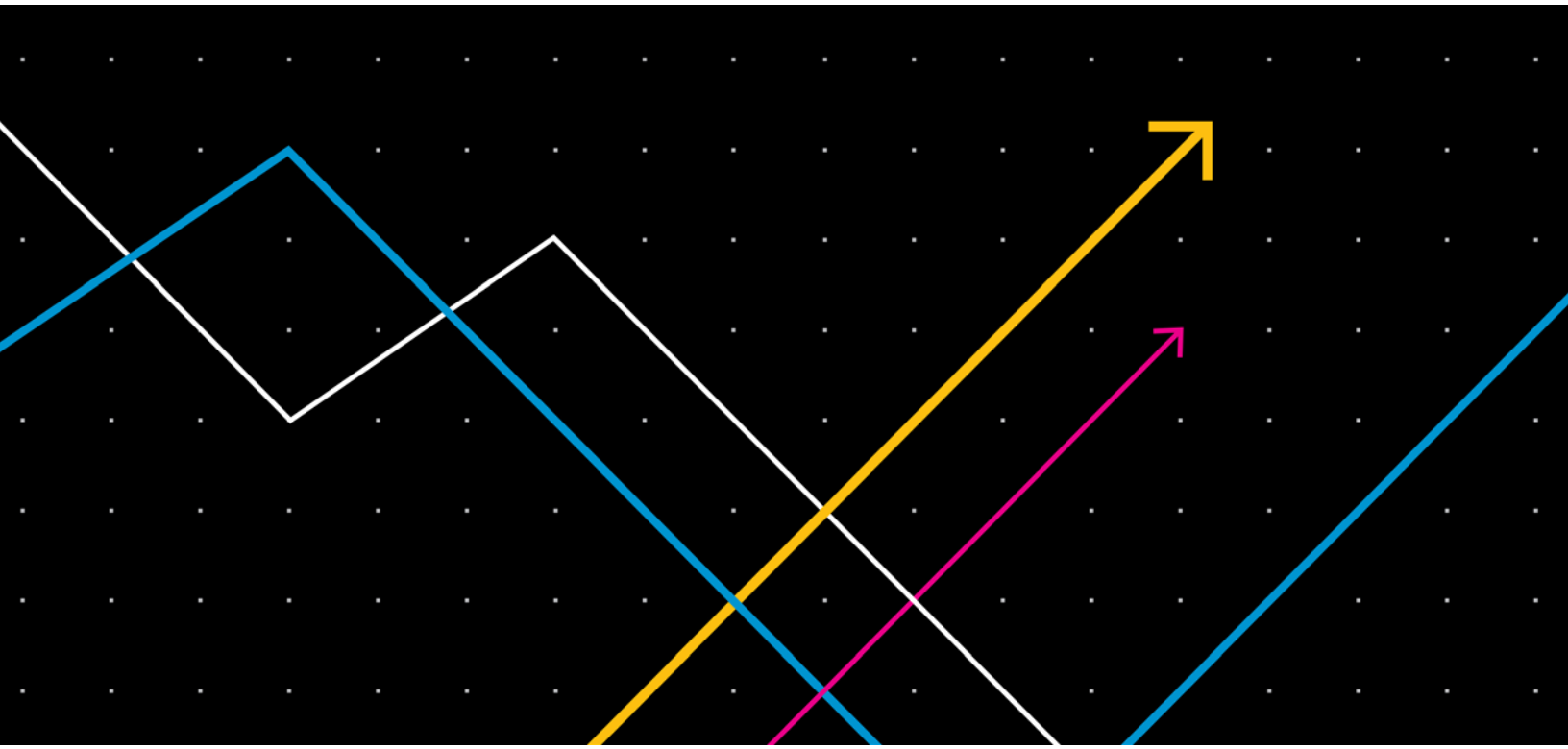


RESEARCH REPORT

Automating Zoning Data Collection

Results from a Pilot Effort to Automate National Zoning Atlas Methodologies

Judah Axelrod
URBAN INSTITUTE

Lydia Lo
URBAN INSTITUTE

Sara C. Bronin
CORNELL UNIVERSITY

February 2023

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization that provides data and evidence to help advance upward mobility and equity. We are a trusted source for changemakers who seek to strengthen decisionmaking, create inclusive economic growth, and improve the well-being of families and communities. For more than 50 years, Urban has delivered facts that inspire solutions—and this remains our charge today.

ABOUT THE LEGAL CONSTRUCTS LAB AT CORNELL UNIVERSITY

The Legal Constructs Lab at Cornell University leads interdisciplinary projects in property, land use, historic preservation, and energy, inquiring into how law can foster more equitable, sustainable, well-designed, and connected places. Further information can be found at https://labs.aap.cornell.edu/bronin.

Contents

Acknowledgments

This report is a product of the Urban Institute's Racial Equity Analytics Lab, which operates with the generous support of the Ballmer Group, the Bill & Melinda Gates Foundation, the Salesforce Foundation, Open Philanthropy, and Urban's general support donors. Lead funding for this report was provided by the Bill & Melinda Gates Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

This report uses data from the National Zoning Atlas, and the research team would like to thank the students, planning experts, geospatial experts, and volunteers who contributed to the Connecticut Zoning Atlas, all recognized at www.zoningatlas.org/connecticut.

The authors also are grateful to Yonah Freemark and Alena Stern for their advice during the project and their close technical review, both of which greatly improved the quality of the final product.

Automating Zoning Data Collection

Researchers and housing sector actors have recently called attention to the lack of transparent, updated, and standardized data on zoning laws across the United States.[1] Assembling these data is a monumental task given the number of local governments—as many as 30,000—that have independently adopted different rules, some under the guidance of state laws. Attempts to catalogue the contents of some zoning codes have included surveys of city planners,[2] statistical modeling that imputes zoning laws from satellite images (Nechamkin and MacDonald 2019; Mass GIS 2003), and laborious but limited manual documentation of zoning rules within or across a small set of jurisdictions or a single state (Bronin, forthcoming; Freemark et al. 2023; Gabbe 2019; Kok, Monkkonen, and Quigley 2014; Glaeser and Ward 2009). Despite these efforts, no group has been able to assemble comprehensive, standardized data on a national scale. In the absence of such data, researchers have difficulty evaluating the impacts of various zoning policies on racial equity, housing production and affordability, economic development, and environmental and climate outcomes across larger sets of jurisdictions. As a result, policymakers lack the evidence needed to guide improvement at scale. This report describes our effort—in partnership with the National Zoning Atlas—to make collecting zoning data across the country easier, lower cost, and more efficient.

The National Zoning Atlas aims to assemble, translate, and document zoning laws for jurisdictions across the nation in a publicly accessible, standardized format. Housed within Cornell University's Legal Constructs Lab, the National Zoning Atlas has a standard procedure that volunteer teams use to assemble and document zoning laws in their city, region, or state. However, the current approach to this process is entirely manual and, due to the volunteer nature of the collection teams, greatly inhibited by a lack of government or private resources. Reducing the burden of collecting and assembling data would go a long way toward scaling up this effort.

Methods from the fields of text analysis and natural language processing (NLP)—a branch of machine learning focused on written and spoken language—demonstrate the potential to help close the data gap and reduce the human effort required. These innovative methods may be able to automate parts of the material assembly, filtering, reading, and documentation of standard zoning characteristics from the haphazard and diverse state in which they are currently documented across the country. To

begin exploring this potential, we first mapped out the Atlas's manual process[3] of reading jurisdictions' legal specifications for different zoning districts and converting this information into structured, comparable data; then, we tested ways to automate different parts of the process.

This exploration highlighted several ways that NLP, optical character recognition, and text analysis can accelerate the process of populating the National Zoning Atlas. Namely, an approach blending manual review with computational tools could reduce the effort currently required to collect and digitize zoning codes, identify zoning district names, and (with further refinement) assemble relevant sections of the codes for human review and data extraction. At the same time, our study also revealed the limits of automation for capturing the diverse forms of expression within zoning's legal strictures. The automation task required a high level of accuracy to match what manual efforts can accomplish, but the combination of highly unstandardized zoning documents and the nuanced nature of the information being collected meant that a fully computational approach was highly unlikely to achieve that level of accuracy. Ultimately, we conclude that a hybrid data collection process combining both manual and automated techniques would result in higher-quality, more nuanced results achieved with greater efficiency than either approach would yield on its own.

Background

We began piloting zoning data automation methods motivated by a context of increasing awareness of the negative impacts of zoning and growing public interest in zoning data collection. The following sections lay out the policy context for this pilot as well as the prior methods and technical exploration that informed our research design.

Why Do Zoning Data Matter?

Zoning laws—created and implemented at the county, township, or municipal levels (which we collectively define as jurisdictions)—delineate how land may be used, what types of lots may be developed, and what dimensions and forms of buildings may be constructed on those lots. These laws use zoning maps to divide land into districts and offer corresponding texts ("code") that explain the rules for each district. Because they delineate the location, look, and feel of most new construction in the United States, zoning laws influence our economy, society, and environment in important but often unacknowledged ways. To ensure that these laws work to improve our lives, we must understand them in far greater detail then we do today.

The content and scope of zoning codes differ from place to place, but the regulatory mechanisms are often similar. Use-based rules within zoning codes typically allow certain broad categories of uses (e.g., residential, commercial, or industrial) in particular districts and then add more specific provisions, such as how many unrelated people may reside in a single residential unit or whether a share of any development's residential units must be rented or sold within certain ranges of the area median income. Similarly, building form and density land-use laws have a wide range of limits such as minimum lot sizes, parking minimums, lot coverage ratios, minimum unit square footage requirements, impervious surface ratios, height limits, setbacks (the required distances from a lot's perimeter before building is permitted), floor-to-area ratios, and many others (see the glossary in Freemark et al. 2022 for definitions of these terms). Although states sometimes establish standards and goals that localities need to meet, the written forms these laws take are diverse and unstandardized, with information and restrictions divided across different chapters. presented in varied text and tables, and using various measurement conventions.

In some respects, zoning laws contribute positively to our quality of life. For example, they can helpfully order the location and shape of development, such as by separating schools and residences from industrial pollution. However, researchers have identified zoning laws as a significant impediment to the housing market's ability to respond to changes in demand and ensure equitable access to quality schools, safe environments, and healthy living conditions and environments (Hsieh and Moretti 2019; Swope and Hernandez 2019; Inturri et al. 2016; Rothwell 2012; Glaeser, Gyourko, and Saks 2003). Local policymakers sometimes support restrictive zoning laws because they preserve the economic conditions and lifestyles attractive to current residents and exclude others. Because these restrictions protect the status quo, they often ossify against any change, regardless of prevailing market forces and demands (Einstein, Glick, and Palmer 2019). Researchers have posited that unchanging zoning laws act as quasi home value insurance and a mechanism for "cartel"-like supply control, artificially restricting development of additional housing units in order to maintain or increase prices of existing units (Been, Madar, and McDonnell 2014; Fischel 2004; Dietderich 1996). This has serious implications for equity when placing zoning laws in the context of our nation's history around homeownership and racial segregation.

Federal policies and individual racist practices have historically advantaged white Americans in accessing homeownership as a wealth-generating engine, which has created and entrenched large racialized wealth, health, and education gaps over time (Ray et al. 2021; Turner et al. 2019). Even as post-World War II federal mortgage and lending policies excluded Black families from homeownership, racial covenants explicitly excluded them from single-family neighborhoods and from building the kind

of wealth through homeownership or gaining access to high-quality education that would allow them to move into these neighborhoods once the restrictions were removed (Rothstein 2017; Trounstine 2018). White homeowners have historically had an outsize influence on zoning policy, using their position to prevent or minimize the development of apartments or mixed uses that enable alternatives to expensive, car-based, large-lot, and sprawling lifestyles (Freemark et al. 2022; Einstein, Glick and Palmer 2019; Trounstine 2018; Been, Madar, and McDonnell 2014). The rise and distribution of class- and race-segregated neighborhoods has led to entrenched, multigenerational cycles of poverty and prevented many Americans from accessing social upward mobility (Chetty et al 2022; Massey and Rugh 2018).

Each of the three most recent presidential administrations has recognized that federal and local policies interact to create inequitable or undesirable housing outcomes, and each flirted with new zoning practices or policies (see Obama's 2016 Housing Development Toolkit, Trump's White House Council on Eliminating Regulatory Barriers to Affordable Housing, and Biden's Housing Supply Action Plan).[4] Yet none has yielded substantive change. To date, neither the White House nor Congress has asserted federal authority over zoning. The reluctance to intervene may result from uncertainty about the degree to which federal law can influence local zoning or the inability of federal policymakers to deeply understand local zoning laws and how to influence them. Lack of federal action may also be rooted in a fear of perceived overreach into what has, for the past century, been considered the province of local control.

Meanwhile, most state legislatures have similarly declined to reclaim their power over local zoning. States themselves, as guardians of police power, have clear constitutional authority to regulate land use. Over the course of the last century, all 50 states have delegated this authority to local governments through enabling acts. Today, only some states—Oregon, California, Connecticut, Massachusetts, and Washington being the most prominent—have begun to modestly reclaim the powers they previously delegated away (Coyle 1993; Stahl 2020). In the absence of strong federal or state guidance on zoning, local officials regulate in a vacuum, with rules that may have negative consequences outside the jurisdiction or that are not harmonized with the rules of their neighbors.

Existing and Emerging Zoning Data

There is no central repository for all US zoning codes, though the National Zoning Atlas described below aims to become such a resource. Instead, current data on zoning comes from secondary sources

such as surveys, automated imputations, rough NLP attempts, or manual standardization efforts. Each of these approaches has significant limitations.

Several scholars have attempted to collect data on zoning policies and practices through representative surveys of planning departments (Pendall, Lo, and Wegmann 2022; Glaeser, Hartley, and Krimmel 2019; Mawhorter and Reid 2018). Despite the value they provide in recording some data about surveyed jurisdictions' zoning laws, surveys have several distinct limitations. First, surveys are voluntary and therefore limited to a respondent's attention span, time, and capacity to share accurate legal information with the surveying entity, which can result in response bias (Pendall 2020). Additionally, these surveys have almost always solicited information at the jurisdiction level, meaning the questions ask whether a certain type of regulation occurs anywhere within the jurisdiction. The answers therefore lack spatial context, because they do not reveal the extent to which the regulation actually applies to land in the jurisdiction. These answers also obscure relationships that zoning has on demographic, economic, and environmental outcome patterns. Consequently, existing survey data provide indirect correlations between broad zoning characteristics within a jurisdiction and the jurisdiction's overall population trends rather than a direct link between types of policies and social outcomes. Additional research also has cast some doubt on the accuracy of survey findings, highlighting further limitations to the usefulness of these data (Lewis and Marantz 2019).

Other scholars have attempted to avoid the costs, errors, and biases associated with surveys by generating standardized datasets of localities' zoning laws themselves. The majority of these efforts are undertaken by private firms that capitalize on selling collected data (e.g., UrbanFootprint, GridX, and Cape Analytics). Academic manual collections have focused on specific regions or regulatory characteristics (e.g., Chriqui, Nicholson, and Slater 2016; Glaeser and Ward 2009; Kok, Monkkonen, and Quigley 2014; Gabbe 2019; Freemark et al., forthcoming). A handful of studies have attempted to generate standardized datasets through computational methods. These efforts have included text analysis techniques to delve into the zoning characteristics of property assessment records or zoning codes and the use of machine learning to understand growth controls, minimum lot sizes, open space requirements, and mandatory affordable housing requirements (Mleczko and Desmond 2020; Nechamkin and MacDonald 2019).

Recognizing the limitations of prior research on zoning, the National Zoning Atlas launched in 2022 to digitize key housing-related regulatory characteristics of the country's zoning codes. It draws from the Connecticut Zoning Atlas, completed in 2021 and inspired by the Desegregate Connecticut advocacy effort that used data compiled in the Zoning Atlas to push for more equitable, pro-development laws in the state. The National Zoning Atlas encompasses teams from 15 states, each

working to identify jurisdictions with zoning, discern each jurisdiction's zoning districts, and then record each district's restrictions using a common methodology.[5] This methodology is articulated in a guide published by the National Zoning Atlas Team, *How to Make a Zoning Atlas* (Bronin and Ilyankou 2022), on which our automated pilot heavily relied.

Successful reforms and initiatives have arisen from Zoning Atlas projects as policymakers, armed with data about the true landscape of housing exclusion and public service provision burdens, attempt to create more equitable policies. For example, the Desegregate Connecticut coalition used the Connecticut Zoning Atlas in advocacy pieces and news reports to motivate and justify the legislature's passage of H.B. 6107 in 2021, and Montana Zoning Atlas findings buttressed the recommendations of a statewide housing task force in 2022. The California Zoning Atlas has released a series of reports on its findings in the San Francisco, Los Angeles, and Sacramento regions, drawing attention from local policymakers and the state legislature, which continues to incrementally modify statewide zoning statutes.

Although the National Zoning Atlas promises to produce the most accurate, nuanced, and useful zoning data available in the public sphere, the methodology requires significant investments of human effort and skill. If some of the data collection were automated and streamlined, the project could more easily be completed, unlocking secondary research about zoning's impacts. In turn, this new research could enable policymakers across the country to make more informed decisions. The following sections describe our process of applying machine learning, natural language processing, and text analysis approaches to the process of gathering, inputting, and validating Zoning Atlas data in order to provide those automated streamlined supports.

Automation Pilot Process and Findings

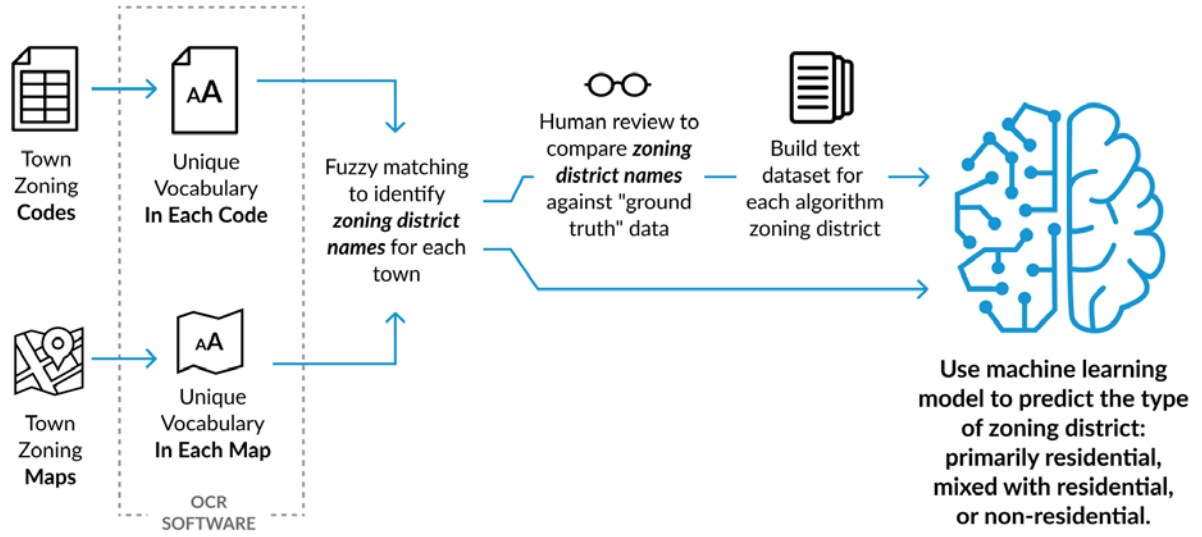
Given the importance of zoning and the inadequacy of hard data on these policies, the goal of this pilot was to assess our ability to automate, and thus accelerate, completion of the National Zoning Atlas. We hoped to identify areas in which machine learning, natural language processing, and text analysis might supersede or supplement human effort. We chose to use the only full-state dataset available—the Connecticut Zoning Atlas dataset[6]—which contains manually collected information about all zoning districts and their regulatory districts, to test our results against a "ground truth."[7] Our team's methodology consisted of the following four steps, also illustrated in figure 1:

1. Gather zoning codes and maps for all jurisdictions in Connecticut and process these documents to enable searchability.
2. Identify algorithmically the zoning district names that appear within those documents.
3. Build text datasets from identified districts that serve as model inputs.
4. Utilize machine learning and natural language processing to generate relevant information pertaining to each zoning district.

Below, we explain the processes we followed for each of these four steps, providing specific examples and case studies. We end with reflections on limitations and promises for putting these steps into practice.[8] For a more technical description of our work, please refer to our accompanying Data@Urban blog post.[9]

FIGURE 1

Diagram of Machine Learning Pilot Methodology



URBAN INSTITUTE

Step 1: Gathering and Processing Zoning Documents

The first step in our automated pilot project was to obtain copies of the zoning text and the zoning map, which together constitute the by-right zoning laws in any given jurisdiction. Typically, copies can be obtained through municipal websites, GIS repositories, and legal databases such as Municode and

eCode360, though in some cases, contacting jurisdiction staff may be necessary (Bronin and Ilyankou 2022). Many of these documents, such as Andover's code and map,[10] were directly accessible online with a URL to a PDF document. Others, such as New Haven's code,[11] needed to be manually appended together, as they were found in a piecemeal format across multiple web pages. Some zoning maps, such as Stamford's,[12] only existed in an ArcGIS online format, requiring manual screenshotting and appending into an image file. Wherever we found the documents, we needed to ensure that any text they contained could be read by our processing tool, so we saved or converted them into a readable file format, such as PDF, PNG, or JPEG.

Gathering these documents can take weeks or months, given the varied nature of how local governments store and document their zoning laws. As long as there is such broad variation, the need for human effort will remain. Even a flawless methodology would struggle to fully automate the process of collecting documents in a complete, authentic, and timely way. Nevertheless, box 1 offers recommendations for how web scraping—a tool our research did not deploy—can augment and speed up the document collection process in future projects.

BOX 1

Area of Promise: Web Scraping to Collect Zoning Information

For zoning codes and maps that are hosted online, web scraping could serve as an alternative to manually locating all of the documents. The following potential web scraping methodology could make this data collection process less onerous for research teams:

- **Using the Bing API to identify municipal webpages.** With a programming language such as Python, researchers can use the search feature within Bing's API to collect a list of municipal webpages where zoning maps and codes might be located.
- **Finding the right subpage.** From these home pages, researchers can use a "web crawler" to extract all subpages from municipal websites and filter to where zoning documents live on the site.
- **Extracting text.** Once the web scraping algorithm identifies the relevant documents, researchers can directly extract text for codes in an HTML format or through OCR software for codes and maps in a PDF, PNG, or JPEG format.

Researchers could employ a similar web scraping process for sites such as Municode or eCode360 that host many local zoning codes. Web scraping will be more complex or impossible for interactive online zoning maps or in cases where the documents are not available online. However, it could certainly reduce the number of jurisdictions for which manual efforts are necessary. Finally, users should ensure that any web scraping is done responsibly and adheres to website terms of service. Please refer to SiteMonitor, a tool created by the Urban Institute, for more information on responsible web scraping.[13]

After amassing all of the necessary documents for the 180 zoning jurisdictions in Connecticut (169 municipalities and 11 submunicipal districts with zoning authority) in readable file formats, we realized that while many documents were searchable, others were not. Typically, the nonsearchable documents came in the form of scanned images or hand-drawn maps. Our team thus decided to use optical character recognition (OCR) software to extract the full text of the documents. Deploying OCR software allowed us to prepare, transform, and clean the texts to ensure a standardized format across all jurisdictions.[14]

Step 2: Identifying Zoning Districts

With the completion of step 1, the team had gathered readable and searchable zoning documents and could proceed with substantive analysis. As articulated in *How to Make a Zoning Atlas*, the starting point

for accurate zoning data collection is the zoning district. These are the units by which local regulation of building processes and limitations vary. All zoning districts in a jurisdiction—both base zones and overlay zones—must be identified before researchers can account for each of the district's regulatory characteristics.

For the people who created the Connecticut Zoning Atlas, assembling a list of districts at first seemed simple. However, the Connecticut research team ultimately spent significant time returning to the zoning documents to review and cross-check district names. Reviewers had some intuitive sense of the types of names a zoning district might have. With this intuition, they could assess not only the "typical" zoning district names (e.g., single-family residential district, light industrial zone, floodplain overlay), but also idiosyncratic, jurisdiction-specific names. This observation underscores an important difference between humans and algorithms: unlike a human reader, no algorithm inherently knows what may or may not constitute the name of a zoning district. No publicly available comprehensive list of zoning districts exists, and while algorithms can learn through enough input data, the names of districts tend to be so idiosyncratic and jurisdiction specific that algorithms' amassed learning would produce little benefit and/or have little external validity.

Drawing from the manual cross-checking techniques developed by the Connecticut team, our team decided to rely on the fact that most zoning district names appear in two documents: the code and the map (in its legend or labels). We hypothesized that by comparing words and phrases that appeared in both documents, after filtering out irrelevant text, only the zoning district names would remain.

With this hypothesis established, we then set out to make cross-document comparison feasible. In terms of their written contents, zoning maps and texts diverge. Maps typically occupy just one page (and at most span a few pages), while zoning codes range from dozens to hundreds of pages. To enable comparisons, we needed to find the locations (i.e., the range of pages) within the text of zoning codes where lists of districts are most likely to be found. In so doing, we could isolate a small amount of the overall zoning code, which would enable comparison. Fortunately, many zoning codes list all their zoning districts in one place, whether in a table of contents or some other consolidated section. Researchers can often identify these lists by searching for specific phrases, some of which are included below with examples:

- Cromwell's zoning code list its districts in the subsections of the table of contents.[15]
- Stamford's zoning code lists its districts on page 45, identified by searching for the phrase "divided into."[16]

- West Haven's zoning code lists its districts on page 8, identified by searching for the phrase "are hereby established."[17]

Reviewing the table of contents or seeking phrases such as "divided into" and "are hereby established" works for most but not all jurisdictions. First, not all towns list their zoning districts in one place, and researchers will not be able to automate the process of identifying districts in those towns. Second, some maps are poor quality or hand drawn, making it more difficult or even impossible for OCR software to read the text legibly. Third, not all zoning districts will appear in both sources; these will be missed entirely by our algorithm. Finally, exact zoning district names could vary between documents, such as appearing with their full name in one document and their abbreviated name in another.

Our team accounted for some of these issues in the algorithm. For example, to address variation in district names, we used a technique known as "fuzzy matching" to identify near matches (e.g., "AL – Andover Lake District" and "Andover Lake District"). However, we were unable to overcome illegible maps or the issue of districts only appearing in one document. Nonetheless, we initially obtained 2,256 "algorithm district names"—that is, the zoning district names identified by the algorithm described above. We then needed a way to compare these results against the 2,385 mapped districts that the Connecticut team found through manual methods. There were surely false positives (irrelevant text that should have been filtered out) and false negatives (districts that the algorithm failed to capture). We implemented a process of manual validation of the algorithm district names, checking discrepancies against the list of districts provided by the Connecticut Zoning Atlas. (See appendix A for the data validation instructions that we devised for a human reviewer.) After this validation process, our team had 1,317 remaining algorithm district names as the list of known districts we could carry forward into step 3, for a match rate of 55.2 percent. This match rate would need to be improved substantially in order for the methodology to be as reliable as manual efforts, and the challenges noted above were the main reason why we see such limited results. Furthermore, extending this approach to zoning documents in other states could reveal additional issues beyond those present in Connecticut.

We conclude that future efforts to identify district names would require a combination of automated district name identification paired with extensive manual validation. Box 2 discusses how a human reviewer might search code text and what such a hybrid process might look like.

BOX 2

Area of Promise: Searchable Text and Narrowing the Search Field

As noted above, it is impossible to know which algorithm district names are correct without a "ground truth" dataset. Thus, automation aiming to identify zoning districts should be paired with manual review. Ideally, manual reviewers would be working from searchable documents that make the identification of relevant information far easier than text in the format of scanned images.

Because districts are listed in one place in most zoning documents, the following hybrid approach could work to identify districts within a jurisdiction:

- Automated text analysis helps identify a small number of pages in which the zoning districts are likely to appear, based on the keywords above (e.g., "divided into," "table of contents," etc.), and pulls an initial list of potential zoning districts in that jurisdiction.
- A human reviewer reviews the selected pages and the preliminary list of zoning districts manually, cross-checking against the zoning map to see if any others are missed.

Automated text analysis with a dictionary of search terms could also help identify special districts and overlays that are less likely to appear alongside other districts. For example, terms such as "flood," "park," "historic," "overlay," and "soil" could identify these additional districts.

Step 3: Building Text Datasets

In step 2, the team developed a strategy for discerning the number and names of zoning districts in Connecticut. We could then use these collected district names to build a dataset or "corpus" of extracted, compiled text from the numerous sections describing each districts' regulatory characteristics. This corpus would consist of excerpts from the zoning codes themselves, which would then be fed into our machine learning model. A portion of these excerpts would serve as the training data that allow the model to learn how to make the predictions necessary to generate accurate information for districts it has not seen (i.e., the test data). Ideally, if a model yielded strong results in this pilot setting, it could be extended to generate data on zoning districts in other states.

To establish the appropriate scope for the corpus, we reviewed the *How to Make a Zoning Atlas* guide, which listed and described all of the district-specific regulations the National Zoning Atlas aims to collect. These regulations cover all aspects of the typical zoning laws described above, including land uses, the size and development of lots, and the construction of structures (though the fields mostly pertain to residential construction characteristics). Among other types of information, collected regulatory data include the following information for each zoning district: permitted uses; districts that

have specific dedications or requirements for affordable or elderly housing; one-, two-, three-, and/or four-or-more-family housing units permitted; minimum lot size; maximum density; floor-to-area ratio; and minimum unit size. We knew we could not train our model to cover all of these categories of collected information. As described in the section on step 4, we ultimately focused on one type of information—the type of zoning district—which can be primarily residential, mixed with residential (i.e., commercial or industrial mixed with residential), or nonresidential. Yet in step 3, we created a dataset that could encompass most of the regulatory data collected in the National Zoning Atlas.

Our methodology to create this dataset prioritized portions of the zoning code that both mentioned a zoning district and offered substantive, district-specific text relating to one or more categories of information the National Zoning Atlas aims to collect. The exact search criteria were complex and entailed the use of "regular expressions," or special patterns used to match character combinations in text. Appendix B includes some of the specific search terms we used to locate relevant excerpts of text. These criteria come directly from sections V–VII of *How to Make a Zoning Atlas*.

This methodology was most successful at identifying excerpts that were descriptive in nature. For example, we were able to extract paragraphs of information describing the purpose of specific zoning districts, including information that could help an Atlas team identify whether a district allows affordable or elderly housing. The algorithm did less well when search terms did not co-occur in close proximity to where the name of a known zoning district appeared or where the zoning codes failed to separately offer information for each of the zoning districts. In these cases, machine learning approaches would have to either parse through hundreds of pages of zoning code or make predictions to generate data without the benefit of knowing which zoning district they were for, both of which were intractable problems.

The methodology was least successful at discerning information provided in tabular form, which was particularly unfortunate given that laws typically offer a significant amount of information about zoning districts exclusively in tabular form. We were able to read the text from those tables, but interpreting spatial relationships between row and column entries—a trivial task for humans—was beyond the scope of our text extraction capabilities. One jurisdiction, Hartford, offered bespoke symbols in its table of principal uses[18] (see appendix figure 3.2A) that our algorithm could not read at all.

As a result of these issues, many text datasets did not pass the eye test—meaning that reading them offered little relevant information to either a human being or a machine learning model. Our reliance on searching for co-occurrence of zoning district names with relevant zoning text, as well as our inability to process tabular data, left some text datasets with little to no text whatsoever. For others, the opposite

problem occurred: so many relevant search terms were matched that the dataset text was long and unwieldy, effectively creating a "needle in a haystack" problem. Calibrating these text datasets is a crucial area of future efforts, as natural language processing tasks rely on having enough relevant data to detect signal but also require limiting irrelevant data to avoid unwanted noise. Apart from further investment in research to address technical shortcomings at this step, box 3 lays out a potential compromise to reduce the amount of manual review needed through the identification of relevant text for each column of the Zoning Atlas data.

BOX 3

Area of Promise: Flagging Relevant Excerpts of Text

Our methodology of searching for co-occurrences of zoning district names and substantive regulations faced significant limitations due to the unstructured and unstandardized nature of zoning codes, combined with the propensity of jurisdictions to assemble their regulatory data in tables. But we do see a role for an algorithm to reduce the amount of time required for National Zoning Atlas teams to identify portions of the text covering district-specific regulatory characteristics. These excerpts may not be robust or clear enough to support machine learning efforts, but they should effectively signal to human reviewers where the most important sections of code are located for the zoning district in question. For tabular information, text analysis may not be able to help interpret the meaning of a table, but it can identify and pinpoint the locations of words and phrases which human reviewers can more easily interpret. This theme of machine *recognition* and human *interpretation* comes up repeatedly in these areas of promise because it plays to the comparative strengths of each in this setting.

Step 4: Using Machine Learning and NLP Techniques to Generate Data

The limited scope of this pilot did not allow our team to test whether our algorithm could read through all the collected excerpts to extract the many district-specific regulations collected by the Connecticut Zoning Atlas. Instead, we decided to test our algorithm's ability to search for and extract just one piece of information: the type of zoning district. The National Zoning Atlas requires each zoning district to be characterized as either primarily residential, mixed with residential, or nonresidential.

We chose to focus on this piece of information for three reasons. First, as explained in *How to Make a Zoning Atlas*, understanding the type of zoning district offers baseline insights about the jurisdiction's regulatory scheme and general constitution. For example, a jurisdiction with most or all districts characterized as primarily residential will likely be a suburban town without much commercial activity,

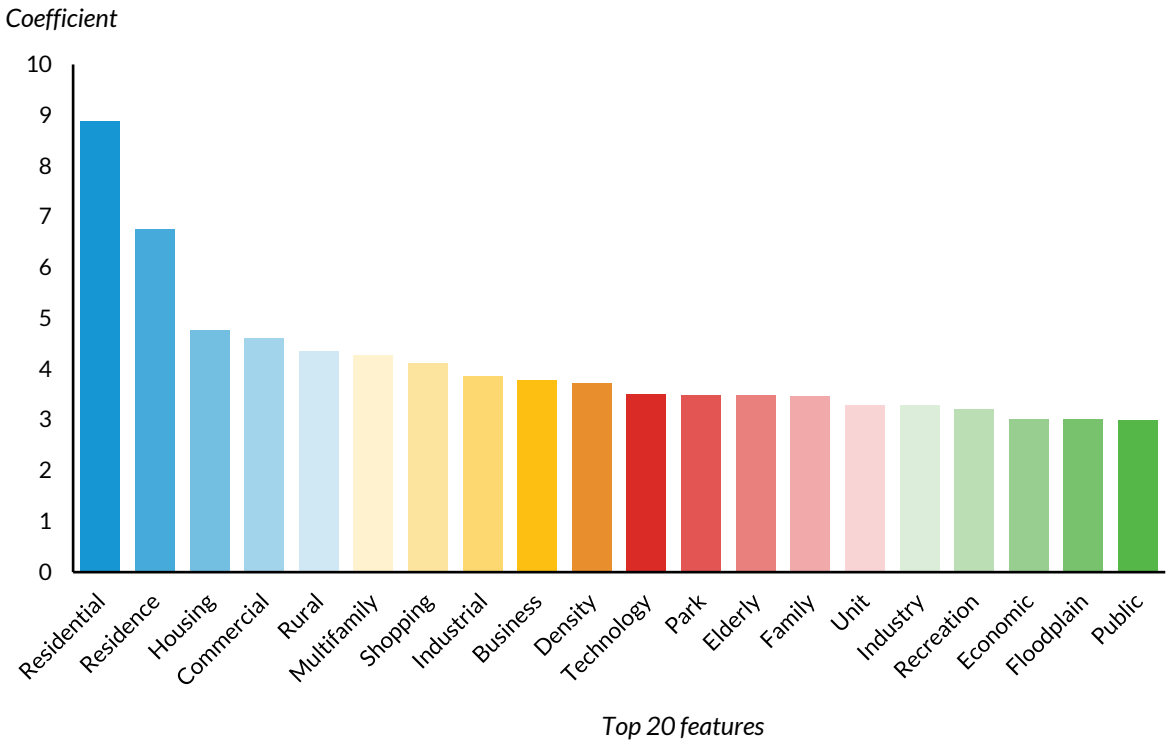
while a jurisdiction with more mixed-with-residential zoning districts will likely be an urban core. Second, unlike other more open-ended inquiries, district type information is a more structured classification with only three options. This application is better suited for machine learning, which has an entire branch devoted to such classification problems. Third, the name of a zoning district often makes this classification easier. For example, it is straightforward to assume that a district named "R-10 One Family Residence" will fall under the primarily residential type, though other cases are far more nuanced.

With this rationale in mind, we used machine learning to predict the type of zoning district for each district in our dataset. Our models relied on two sources of information: the 2,385 zoning district names collected by the National Zoning Atlas team and the text datasets created in step 3. Following best practices in machine learning processes, we divided the zoning districts into a training set and a test set, the sizes of which are arbitrary but often set at 70–80 percent and 20–30 percent, respectively, by convention.[19] Our training set thus consisted of 1,669 zoning districts, and our test set consisted of 716 districts. The training set serves to teach the model how to make predictions. To avoid allowing the model to "cheat" by making predictions on the same data it has already seen and learned from, the test set then allows for us evaluate how well the model performs on unseen data. This is worth mentioning because any discussion of the accuracy of the machine learning model is referring only to accuracy on the 30 percent, or 716 unseen zoning districts in the test set. If the test set is too large, we may not be giving the model enough training data from which to learn. But if it is too small, we may not have enough data from which to glean anything useful about the accuracy of the model.[20]

Using the training set, we created variables that measure "term concentration" for each of the three types of zoning districts. To do so, we collected counts of the number of times words or phrases appeared in each text dataset in the training set that related to residential, mixed residential, and nonresidential districts. In order to scale these counts, we then divided them by the average length of text datasets among all districts in a given town, leaving us with a measure of how highly concentrated those terms are. The specific search terms were informed by *How To Make a Zoning Atlas* and are described in more detail in appendix C. We then reviewed the importance of each of these variables to the model's ability to make predictions. The 20 most important variables are included in figure 2 below. We find that the appearance of words such as "residential" or "residence" in the zoning district name are the most valuable pieces of information used by the model.

FIGURE 2

Variable Importance for Predicting Zoning District Type



URBAN INSTITUTE

Table 1 below displays four key evaluation metrics for the machine learning exercise: accuracy, precision, recall, and F1 score.[21] Accuracy is simply the proportion of zoning districts in the test set whose type the model correctly predicts. Precision refers to the proportion of zoning districts that the model predicts as falling within one of the three categories (e.g., nonresidential) that actually fall into that category. Conversely, recall refers to the proportion of zoning districts that actually fall into one of the three categories that the model predicts correctly. For example, let's say there are 40 nonresidential districts and the model correctly identifies 30 of them as nonresidential, incorrectly identifies the other 10 as either primarily or mixed residential, and also incorrectly identifies 20 other districts as nonresidential. In this case, the precision for nonresidential districts would be 30/(30+20), or 0.6. The recall for nonresidential districts would be 30/(30+10), or 0.75. These two figures are helpful for better understanding specific strengths and weaknesses of a model with more specificity than we can gather from overall accuracy alone. Finally, the F1 score averages the precision and recall into one aggregated number.

TABLE 1

Model Results

Zoning district type	Accuracy	Precision	Recall	F1 Score	Sample size (test set)
Primarily residential	-	87%	91%	89%	277
Mixed with residential	-	75%	61%	68%	297
Nonresidential	-	53%	69%	60%	142
Overall total weighted average	**74%**	**75%**	**74%**	**74%**	**716**

In the first three rows of table 1, we can see the results for each of the categories. They show that the model performs extremely well in precision, recall, and F1 score for the primarily residential districts, though not as well for the mixed and nonresidential districts. This is most likely because the residential categories are simply easier to predict; as shown in figure 2, the two most important variables in terms of the predictive power in the machine learning model were the presence of "residential" and "residence" in the district name. The fourth row shows the overall accuracy of the model across all three residential categories, as well as the total sample size of the test set. The final row computes aggregated metrics; precision, recall, and F1 score represent averages of the preceding rows, weighted by the number of zoning districts in each category. All four metrics yield nearly identical results, with the simplest interpretation coming from accuracy (in the fourth row): the model is able to correctly classify 74 percent of the 716 zoning districts in the test set.

This is a promising result, although one that is perhaps buoyed by the 277 primarily residential districts (which, as noted above, benefited from predictive terms used in their titles). The results also indicate that this same methodology would not be able to extend to other district-specific regulatory characteristics without significant changes. The variables for term concentration were among the very *least* important variables in the best model, suggesting that almost all of the predictive power came from the district names, rather than the text datasets. In other words, if figure 2 were extended to show results beyond the top 20 variables, nearly every other word found in the zoning district titles was more predictively important than the term concentrations that came from the text datasets in step 3.

Building input data that contain more signal and less noise will be crucial to improving model results in the future. Our best model significantly outperformed a baseline "rules-based" approach that did not use machine learning techniques but instead relied on logic outlined in the *How to Make a Zoning Atlas* guide. For instance, the guide advises, "Any district full name including the words 'planned residential', 'planned residence', 'planned unit development', or 'active adult' will almost certainly be Primarily Residential," and we instructed our baseline model to classify those districts as such. This logic does not

hold in every case but serves as a useful heuristic, and any model that could not outperform the classification accuracy of this baseline method would not be worth considering. While the algorithm may not yet be ready to apply and rely on for standardization of yet-undocumented zoning data, these results demonstrate that the model is learning from the information that it is given.

A Proposal for a Hybrid Zoning Atlas Methodology

Automated zoning data collection techniques are rife with both promise and limitations. Our exploration yielded several clear conclusions. Foremost among them is the revelation that relying fully on algorithmic methods to digitize and standardize zoning data would be impractical given the complex and esoteric nature of the task at hand. Accordingly, we propose that the National Zoning Atlas and other partners working on the project of creating a national zoning database evolve their data collection strategy to combine manual and automated steps into a more efficient, hybrid process.

Below, we lay out step by step what such a hybrid methodology might look like for a given Zoning Atlas project, drawing on the areas of promise identified above and laying out areas where targeted exploration is most needed.

Step 1: Gathering and Processing Zoning Documents

The first step of hybrid methodology would be to gather zoning codes and maps and render them legible, as employed in step 1 of the methodology used for our automated pilot project. In the hybrid approach, Zoning Atlas teams would implement web scraping techniques (see box 1) to search municipal websites and code databases, collecting links to zoning codes and maps into a spreadsheet and then extracting documents or HTML text directly from these links. The research team would need to manually collect documents for any missing spreadsheet entries or entries for which the links do not lead directly to documents. Web scraping output might be erroneous for certain jurisdictions, but even in these cases might lead to intermediate results that allow reviewers to more quickly navigate to the proper subpage within a website than if they were starting from scratch.

After the researchers collect documents for all covered jurisdictions, the team would extract text from these documents using optical character recognition software, exactly as we did for this pilot project. Support from web scraping and OCR software will reduce time and effort spent in this data gathering phase.

Step 2: Identifying Zoning Districts

In the second step of the hybrid methodology, Zoning Atlas teams would implement the same methodology laid out in step 2 of our automated pilot. They would algorithmically generate a list of district names for each jurisdiction, as well as a subset of page numbers of the zoning code where the entire list of districts is likeliest to appear. Reviewers then could compare the algorithm district names with what they see in the zoning code, filtering out irrelevant text and adding names the algorithm missed. The team would finish by cross-checking that list a final time against zoning map legends and labels for districts that might only be present in one of the two documents.

For many jurisdictions, the algorithm district names will be entirely or nearly comprehensive, saving reviewers valuable time with manual data entry for some or all of the districts. For others, the algorithm's output might be entirely unhelpful, forcing reviewers to disregard results and revert to a manual process. Even in the unhelpful cases, the page numbers returned by the algorithm should enable researchers to more quickly filter relevant text from hundreds of pages.

Note that an algorithm could also assist with gathering both the full and abbreviated names of a zoning district (e.g., "IPD" and "Industrial Park District"). The National Zoning Atlas methodology requires the collection of both names, as they are often used interchangeably. The algorithm used in the pilot could capably match both names, particularly when the abbreviated name is a perfect acronym of the longer name. An updated algorithm could achieve this goal more comprehensively, which could further save reviewers search and entry time per jurisdiction.

Step 3: Building Text and Table Databases

In our automated pilot, step 3 consisted of building a dataset, or corpus, of extracted zoning code text that could be read by the algorithm for specific regulatory data. In step 4 of the automated pilot, we generated data for just one piece of information—the type of zoning district—which required answers to take one of three specified forms (primarily residential, mixed with residential, or nonresidential). We believe that the methods used in the pilot to generate data could be extended to a few other pieces of information required by the National Zoning Atlas to have answers within a similarly specified menu. Good candidates for this include the "elderly housing district" and "affordable housing district" designations. These are simple, binary categories (requiring "yes" or "no" answers) that can generally be answered after searching for a few key phrases in the zoning text or by reviewing the name of the zoning district alone.

However, most pieces of information that the National Zoning Atlas requires are not such ideal candidates for algorithmic approaches. Pieces of information allowing open-ended answers require a more complex approach. For a hybrid approach, then, we suggest that for the latter category of data, the algorithm extract relevant text for humans to manually review. At least in the short term, we suggest that machine learning applied to the National Zoning Atlas focus less on complete data generation and more on the construction of databases of relevant zoning law excerpts (i.e., strategically collecting rather than analyzing zoning information).

First, we suggest tuning the algorithm to search for occurrences of zoning district names within the same window of text as certain search terms. These search terms would correspond to each piece of information that must be collected (e.g. "lot size" for minimum lot size, "F.A.R." for floor-to-area ratio, etc.). The algorithm can either provide page numbers or extract relevant text. From this output, team members can more efficiently find, read, and pull relevant data from these pages. Such text analysis will not handle every case, as sometimes the information needed to fill out a column is more subtle and nuanced than what can be captured in short search terms and phrases (e.g., restrictions that include conditions or that refer to other portions of the zoning text), but for the majority of cases, teams should be able to limit long zoning texts to just the relevant portions needed for manual review. These include information living in relevant tables about uses and dimensions. As noted above, it may be difficult to train an algorithm to recognize how rows and columns intersect to parse and analyze the information contained by the tables. But if human reviewers had easy access to these tables, they could enter the data more efficiently.

Step 4: Data Validation

As noted above, it is important to ensure that data included in a Zoning Atlas is free of errors. One benefit of an algorithm is the ability to establish a scope of possible answers. In the automated pilot, the algorithm was incapable of producing responses beyond the three specific preprogrammed answers (e.g., it could not categorize a zoning district as "orange" or "superstar"). Other pieces of information the National Zoning Atlas is required to collect might similarly have a fixed set of answers, such as the binary yes/no answers required by the affordable housing and elderly housing district categories. Still others may require answers within a given range, such as a numerical range. For example, zoning laws use a number, typically between 1 and 100, for the maximum height of buildings or the minimum setback width. Data that come in letter form, data that exceed 100 (likely an accurate entry only in larger cities), or data entered in fractions or percentages could be flagged as potential errors.

An algorithm could be trained to search for specific types of information and discard potentially erroneous information, but a well-designed data entry system could bypass the need for such an algorithm. As we were drafting this report, the National Zoning Atlas team opted to develop the latter. Thus, in future data generation efforts, the National Zoning Atlas will be using a data entry system that places parameters on possible answers relevant to the type of information being collected. This system will use provided parameters to check data as they are entered. Entries that violate logical rules will be rejected or require editing.

Step 5 and Beyond: Further Exploration of Machine Learning and NLP

Any serious attempts to automate the creation of text datasets and use predictive modeling to fill out data columns need to address the fact that much important information is presented in tabular form. Experimenting with tabular machine learning methods was beyond the scope of our team's pilot project, but future work to improve upon model accuracy and build out further time-saving automation for the National Zoning Atlas process should start with investment in this area.

Even with better input data to develop a more refined algorithmic process, we think it is unlikely that a computational model on its own will achieve near-100 percent accuracy, which is the ultimate goal of a National Zoning Atlas. Further, without ground truth data, there will be no way to validate which entries are correctly or incorrectly generated absent human review. Future efforts will need to be tested and trained against existing National Zoning Atlas data and new data assembled across more jurisdictions. The model will need to be refined and tailored to new state contexts and improved collectively over time with additional use. Incorporating human review into the pipeline allows for data validation, even in cases where no ground truth data exist.

Still, collecting certain types of information is likely easier to automate, and we believe a hybrid approach can improve expediency without a significant loss of data quality. As our analysis demonstrated, promising candidates include data columns that have a finite set of categories and for which other data, such as the zoning district name, can add information. While removing human review from the process altogether is impractical, if a model can correctly predict enough entries that a reviewer only needs to spot check results (rather than enter them all manually), this represents a major time save for Zoning Atlas teams.

Future investigations would also benefit from expanding test data to incorporate manually generated zoning data from at least two to three states from other regions in the United States in order to ensure that the final hybrid methodology takes into account potential regional differences in zoning

code formats. Our project's focus on Connecticut-only data may have resulted in either false restrictions or overly optimistic projections about the usefulness of automation, and future efforts will need to test the generalizability of our results.

Conclusion

There are several benefits to ensuring we have a standardized dataset about zoning rules throughout the country and the spatial maps to which those rules apply. With these data, researchers and regional, state, and federal government authorities would have a much easier time assessing and understanding the impact of zoning laws. And we could finally monitor and identify the exact policies that hinder or improve racial equity, climate resilience, access to occupational and educational opportunities, and environmental health, and with that knowledge create more effective land-use policies to maximize health, well-being, and equity for all residents.

Governments seeking to ensure the equitable distribution of public goods and maximization of economic growth while preserving environmental integrity and neighborhood quality have an interest in investing in the creation of zoning atlases. To that end, they could allocate funds or staffing to support their creation. State and federal government requirements could also establish transparency and reporting standards for jurisdictions with zoning, aligning such standards with those of the National Zoning Atlas already underway. These reporting requirements could be tied to funding opportunities, as the Biden administration recently proposed. Or, they could be justified as a necessary means for determining whether local governments are satisfying their obligations under the federal Fair Housing Act and state equivalents to affirmatively further fair housing. Given the pressing need for these data, the method of support for enhanced zoning data collection matters less than the building of momentum, creative pathways, and political will to see greater democratic oversight in land use.

Appendix A. Zoning District Validation Instructions

The goal of this task is to match zoning district names identified algorithmically through text analysis with the true list of districts identified manually for the Connecticut Zoning Atlas. This will help with (1) understanding how well the algorithm performs at identifying zoning districts in Connecticut; and (2) subsequent text analysis/machine learning which will try to fill out the other columns in the Atlas data.

For some towns, the algorithm will have done a good job at identifying the district names, and this will be very simple. For others, there will be duplicates and extraneous text to sort through, or the algorithm may have failed to identify any district names and the town can just be skipped. Below is a walkthrough of how to conduct the matching process. Thanks very much for your assistance!

Data You Need

Two spreadsheets:

- Zoning Atlas Data, "Mapped Districts" tab: (Only columns A–C are relevant here - Town, Abbreviated District Name, and Full District Name)
- *algorithm-district-names.csv*: contains a list of zoning districts identified by the algorithm.

Methodology

1. Copy columns A–C over to a new spreadsheet, which you can call matched-districts.csv.
2. Create new columns D, E, and F in matched-districts.csv called "Algorithm Abbreviated Name," "Algorithm Full Name," and "Match Plus Extra Text," respectively. This will be where you paste the zoning district names from the algorithm that match up with districts in the Atlas data.
    a. For example, the first district in the Atlas data is "Andover Lake." This matches the full name "andover lake" that was identified by the algorithm. (Unfortunately, it isn't always this easy!)
3. For each town:
    a. Filter to that town in both spreadsheets using the "Filter" option in Excel.

i. Shortcuts for Windows users to save time (can otherwise just use your mouse to accomplish the same thing):
    1. If you click on a column's name in row 1, CTRL + SHIFT + L allows you to quickly filter and unfilter that column.
    2. After filtering, ALT + the down arrow key will display the filter options for that column. Once the options appear, typing "E" on your keyboard will jump you right down to the search menu, where you can type the name of the town you wish to filter to.

b. Start with the town's first district in the Atlas data.
c. For that same town, search in *algorithm-district-names.csv* for a zoning district name that matches *either* column B or C in the Atlas data. **See part D below for how to identify a match.**
d. If you find a match, copy and paste the algorithm name into the corresponding cell in column D or E. It's important that there are no typos, which is why we ask you to copy and paste rather than manually enter.
    i. Sometimes the algorithm will find duplicates (e.g., both the abbreviated and full names) for a district. In this case, copy them both over into columns D and E.
    ii. Sometimes there can even be multiple duplicates beyond this (e.g., "flood prone district" and "flood plain district"). If so, use your judgment to choose the most informative name. In other words, there should be a maximum of two names (one short and one long) for each district that are copied over.
e. If you do not find a match, leave columns D and E blank.
f. Go on to the next district for that town and repeat this process

4. Example of what *matched-districts.csv* would look like for Andover, the first town in the Atlas Data. Note the following:
    a. "Soil and water conservation" (identified by the algorithm) does not appear in the "Mapped Districts" tab of the Atlas data, so it should be ignored here:
    b. "Flood" (identified by the algorithm) is a duplicate for "flood prone" and does not match the abbreviated name "FP," so it can also be ignored.
    c. In this case, the algorithm identifies all of the full names and none of the abbreviated names, but sometimes there will be overlap.

Town	Abbreviated district name	Full district name	Algorithm abbreviated name	Algorithm full name	Match plus extra text
Andover	AL	Andover Lake		andover lake	
Andover	ARD	Andover Rural Design		andover rural design	
Andover	B	Business		Business	
Andover	FP	Floodplain		flood prone	
Andover	I	Industrial		industrial	

Defining a Match

Sometimes the algorithm will identify the abbreviated name, the full name, or both names in the Atlas data. If the algorithm name matches the abbreviated district name, it should be copied over to column D. If it matches the full district name, it should be copied over to column E. For each algorithm name, the following conditions constitute a match to one of the Atlas columns:

Comparing to Abbreviated District Name (Column B)

- Exact match with identical spelling
- Exact match, besides any of the following exceptions:
  - » Punctuation, capitalization, or spacing differences between the names (e.g., "r1," "R 1," and "R-1" should all be considered identical)
  - » Trailing "Z" or "D" at the end (which could just be because the word "zone" or "district" was removed from one but not the other).

Comparing to Full District Name (Column C)

- Exact match with identical spelling
- Exact match besides any of the following exceptions:
  - » Punctuation, capitalization, or spacing differences between the names (e.g. "r1", "R 1", and "R-1" should all be considered identical)
  - » Minor typos or misspellings (which may just be due to poor-quality scans of zoning codes/maps that are misread by the software). **Please do not correct any typos—copy the algorithm name over as is.**

    - The presence of the words "zone" or "district," which can be ignored
    - The presence of the word "overlay" if it is clear that the other words in the name point to this overlay district **and no other district in that town.**
    - A few words being abbreviated (e.g., "Plan. Dev." instead of "Planned Development"), where abbreviated terms clearly point to the accurate name of the district **and no other district in that town.**

- Exact match based on the conditions above, except for some irrelevant text at the beginning or end of the name (e.g., *"incentive housing overlay zone (adopted 12/18/13"*). Such cases should be copied over to column F, "Match plus extra text."

Notes

- If it is ambiguous which district in the Atlas data matches an algorithm name (e.g., "PARD" could match to any of "PARD #1," "PARD #2," or "PARD #3"), these should **not** be considered matches.
- Algorithm names that are subsets of Atlas names but do not satisfy the criteria above should **not** be considered a match. For example, "residential" should not be considered a match for the more specific Atlas name "r1 residential."
- There will be some occurrences where an algorithm name is a combination of the abbreviated and full district names (e.g., "co – corridor overlay"). This should be considered a match for the "Full District Name" only.
- If you identify any other cases not covered by these conditions that you still think are matches, please do copy them over, but highlight the row so that we know to double-check. We want to leave room for your own judgment.
- Some towns may be missing entirely from *algorithm-district-names.csv* (e.g., Bolton), or the zoning district names may all be totally irrelevant (e.g., for Branford - Pine Orchard, all the district names are false positives). This is a known issue with the algorithm, and in these cases, you can just skip to the next town.
- We hope to eventually use this methodology as one piece of a hybrid how-to guide for other states that leverages the things humans and machines can each do well to complement one another. Your feedback on these instructions and your help are both hugely appreciated! You can refer any questions or comments to jaxelrod@urban.org.

Appendix B. Search Terms for Building Text Datasets

Topic	Some relevant search terms
Permitted use	“uses permitted,” “permitted uses,” “use regulations,” “use table,” “allowed use”
Affordable housing	"affordable,” “opportunity,” “workforce,” “incentive housing,” “specialty housing”
Elderly housing	“elderly,” “age restricted,” “senior,” “active adult,” “older,” “golf,” “planned”
One-family housing	“single family,” “one family,” "mobile," "manufactured,” “residential use”
Two-family housing	“two family,” “duplex,” “no more than two units,” “multifamily”
Three-family housing	“three family,” “triplex,” “no more than three units,” “multifamily”
Four-family housing	“four family,” “quadplex,” “multifamily,” “apartment”
Minimum lot size	“lot,” “lot and building,” “area and bulk,” “dimensional,” “minimum lot”
Maximum density	“units per acre,” “density per acre,” “dwellings per acre,” “maximum density”
Floor-to-area ratio	“floor to area,” “FAR,” floor area ratio”
Minimum unit size	“unit size,” “minimum size,” “floor area”

Appendix C. Search Terms for Calculating Term Concentration

Topic	Some relevant search terms
Primarily residential	"residential," "housing," "mobile home," "agriculture," "active adult," "elderly," "planned unit development"
Mixed with residential	"mixed," "village," "central business," "main street," "college" or "university" near "dormitory"
Nonresidential	"open space," "land conservation," "airport," "office park," "public utility," "cemetery," "commercial," "business," "industrial"

Notes

[1] See Bronin (2022); Erika Tyagi and Graham MacDonald, "We Need Better Zoning Data. Data Science Can Help," Greater DC—Urban Institute, October 15, 2019, https://greaterdc.urban.org/blog/we-need-better-zoning-data-data-science-can-help; Lydia Lo, "Who Zones? Mapping Land-Use Authority across the US," *Urban Wire* (blog), Urban Institute, December 9, 2019, https://www.urban.org/urban-wire/who-zones-mapping-land-use-authority-across-us; Jenny Schuetz, "Is zoning a useful tool or a regulatory barrier?" Brookings Institution, October 21, 2019, https://www.brookings.edu/research/is-zoning-a-useful-tool-or-a-regulatory-barrier/; "The Data Challenge in Cities," Symbium (blog), July 13, 2022, https://symbium.com/blog/the-data-challenges-in-cities.

[2] See, for example, Pendall, Lo, and Wegmann (2021) and Gyourko, Hartley, and Krimmel (2019). Additional surveys going back further in time include Robert W. Burchell and Michael L. Lahr, "A National Survey of Local Land-Use Regulations" (New Brunswick, NJ: Edward J. Bloustein School of Planning and Public Policy, Rutgers University, 2008); David D. Foster and Anita A. Summers, "Current State Legislative and Judicial Profiles On Land-Use Regulations in the U.S." (Philadelphia: Zell/Lurie Real Estate Center, University of Pennsylvania, 2005), http://realestate.wharton.upenn.edu/wp-content/uploads/2017/03/512.pdf; Ned Levine, "The Effects of Local Growth Controls on Regional Housing Production and Population Redistribution in California," *Urban Studies* 36 (2047) (1999); Madelyn Glickfeld and Ned Levine, "Regional Growth, Local Reaction: A Review of Empirical Evidence on the Effectiveness and Efficiency of Local Government Land Use Regulation" (1992); Peter Linneman et al., "The State of Local Growth Management" (Philadelphia, PA: Wharton Real Estate Center, 1990).

[3] Sara Bronin and Ilya Ilyankou, "How to Make a Zoning Atlas: A Methodology for Translating and Standardizing District-Specific regulations," National Zoning Atlas, accessed December 22, 2022, https://www.zoningatlas.org/how.

[4] The White House, "Housing Development Toolkit" (Washington, DC: The White House, 2016), https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/images/Housing_Development_Toolkit%20f.2.pdf; US Department of Housing and Urban Development, "White House Council on Eliminating Regulatory Barriers to Affordable Housing; Request for Information," *Federal Register*, November 22, 2019, https://www.federalregister.gov/documents/2019/11/22/2019-25388/white-house-council-on-eliminating-regulatory-barriers-to-affordable-housing-request-for-information; The White House, "President Biden Announces New Actions to Ease the Burden of Housing Costs," News release, May 16, 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2022/05/16/president-biden-announces-new-actions-to-ease-the-burden-of-housing-costs/.

[5] The Zoning Atlas only records a small share of the overall elements of the zoning code, and its ability to describe elements that are conditional is limited, but the standardized fields it does require for reporting still cover essential restrictions on a basic level.

[6] "Connecticut Zoning Atlas," National Zoning Atlas, accessed December 22, 2022, https://www.zoningatlas.org/connecticut.

[7] Connecticut's governance structure is different than that of most states because it has no unincorporated land and county governments have no zoning power. Instead, all zoning decisions and land-use authority are at the town or municipality level. Given the early stages of the National Zoning Atlas data collection process, we are not yet able to compare Connecticut's land-use data to that of other states, though we should not expect that our automation pilot would transfer immediately to other contexts without additional testing.

[8] In a forthcoming publication, we will provide a full technical appendix and public GitHub repository for others interested in replicating or building on any aspect of our methodology.

[9] Judah Axelrod, "The Potential of Machine Learning for Compiling Standardized Zoning Data," *Data@Urban* (blog), Urban Institute, February 27, 2023, https://medium.com/@urban-institute/the-potential-of-machine-learning-for-compiling-standardized-zoning-data-c0f4697a9f0.

[10] Town of Andover Planning and Zoning Commission, "Zoning Regulations" (Andover, CT: 2019), https://www.andoverconnecticut.org/sites/g/files/vyhlif5346/f/uploads/zoning-regulations-effective-7.15.2019.pdf; "Zoning Districts, Andover, CT" (Branford, CT: New England Geo Systems, 2017), https://www.andoverconnecticut.org/sites/g/files/vyhlif5346/f/uploads/appendix_1_zoning_map.pdf.

[11] "Zoning Supplement 27," Zoning Ordinance City of New Haven, Connecticut, updated April 5, 2021, https://library.municode.com/ct/new_haven/codes/zoning.

[12] "Stamford CT Zoning Map," Stamford Planning Board, accessed January 2023, https://www.arcgis.com/apps/webappviewer/index.html?id=1ea803054a1f4d049f3bcc7035d2c20c.

[13] Jeffrey Levey and Graham MacDonald, "SiteMonitor: A Tool for Responsible Web Scraping," *Data@Urban* (blog), Urban Institute, April 16, 2019, https://urban-institute.medium.com/sitemonitor-a-tool-for-responsible-web-scraping-e759042e296a.

[14] Several OCR offerings exist with various pros and cons, but we chose Amazon Textract, a cloud-based offering from Amazon Web Services. For teams looking for additional guidance on which OCR offering will work best for them, please refer to Judah Axelrod, "Choosing the Right OCR Service for Extracting Text Data," *Data@Urban* (blog), Urban Institute, March 25, 2022, https://urban-institute.medium.com/choosing-the-right-ocr-service-for-extracting-text-data-d7830399ec5.

[15] Town of Cromwell, "Zoning Regulations, Town of Cromwell" (Cromwell, CT: 2015), https://www.cromwellct.com/sites/g/files/vyhlif2976/f/uploads/zoning_regulations_effective_8-25-15.pdf.

[16] City of Stamford, "Zoning Regulations, City of Stamford, Connecticut" (Stamford, CT: 2023), https://www.stamfordct.gov/home/showpublisheddocument/5847/637699103943570000.

[17] City of West Haven, "West Haven, Connecticut, Zoning Regulations" (West Haven, CT: 2022), https://www.cityofwesthaven.com/DocumentCenter/View/4504/Zoning-Regulations-Revised-to-03-31-22-PDF.

[18] City of Hartford, "City of Hartford Zoning Regulations" (Hartford, CT: 2020), https://www.hartfordct.gov/files/assets/public/development-services/planning-zoning/pz-documents/zoning-regulations/zoning-regulations-06052020.pdf.

[19] See "Splitting the Data into Training and Evaluation Data," Amazon Web Services, accessed January 2023, https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-the-data-into-training-and-evaluation-data.html.

[20] We evaluated a number of classification models, including a baseline logistic regression, random forest, and support vector classifiers with different types of kernels. After tuning hyperparameters using cross-validation, we found that a support vector classifier using a linear kernel achieved the best results on the test set, though most of the other tuned models performed similarly. We did not explore deep learning models that can fit even more flexibly to the data, which we leave to future research.

[21] For a more thorough description of the machine learning models applied in this section, see the *Data@Urban* blog post describing our technical methodology. Axelrod, "Choosing the Right OCR Service for Extracting Text Data."

References

Been, Vicki, Josiah Madar, and Simon Thomas McDonnell. 2014. "Urban Land Use Regulation: Are Homevoters Overtaking the Growth Machine?" *Journal of Empirical Legal Studies* 11 (2): 227–65.

Bronin, Sara C. Forthcoming. "Zoning by a Thousand Cuts." *Pepperdine Law Review* 50. *SSRN* 3792544.

Bronin, Sara C, and Ilya Ilyankou. 2022. "How to Make a Zoning Atlas: A Methodology for Translating and Standardizing District-Specific Regulations." *SSRN* 3996609.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, and Sara Gong. 2022. "Social Capital I: Measurement and Associations with Economic Mobility." *Nature* 608 (7921): 108–21.

Chriqui, Jamie, Lisa M. Nicholson, Emily Thrun, Julien Leider, and Sandy J. Slater. 2016. "More Active Living–Oriented County and Municipal Zoning Is Associated with Increased Adult Leisure Time Physical Activity—United States, 2011." *Environment and Behavior* 48 (1): 111–30.

Coyle, Dennis. 1993. *Property Rights and the Constitution: Shaping Society through Land Use Regulation.* New York: SUNY Press.

Dietderich, Andrew. 1996. "An Egalitarian's Market: The Economics of Inclusionary Zoning Reclaimed." *Fordham Urban Law Journal* 24 (1): 23–104.

Einstein, Katherine, David Glick, and Maxwell Palmer. 2019. *Neighborhood Defenders: Participatory Politics and America's Housing Crisis.* Boston: Cambridge University Press.

Freemark, Yonah, Lydia Lo and Sara Bronin. Forthcoming. "Bringing Zoning into Focus: A Fine-Grained Analysis of Zoning's Relationships to Housing Affordability, Income Distributions, and Segregation in Connecticut." Washington, DC: Urban Institute.

Freemark, Yonah, Lydia Lo, Olivia Fiol, Gabe Samuels, and Andrew Trueblood. 2023. *Making Room for Housing Near Transit: Zoning's Promise and Barriers*. Washington, DC: Urban Institute.

Freemark, Yonah, Lydia Lo, Eleanor Noble, and Ananya Hariharan. 2022. "Cracking the Zoning Code: Understanding Local Land-Use Regulations and How They Can Advance Affordability and Equity." Washington, DC: Urban Institute.

Gabbe, C.J. 2019. "Changing Residential Land Use Regulations to Address High Housing Prices: Evidence from Los Angeles." *Journal of the American Planning Association* 85 (2): 152–68.

Glaeser, Edward, and Bryce Ward. 2009. "The Causes and Consequences of Land Use Regulation: Evidence from Greater Boston." *Journal of Urban Economics* 65: 265–78.

Glaeser, Edward, Joseph Gyourko, and Raven Saks. 2003. "Why is Manhattan So Expensive? Regulation and the Rise in House Prices." Working Paper 10124. Cambridge, MA: National Bureau of Economic Research.

Gyourko, Joseph, Jonathan Hartley, and Jacob Krimmel. 2019. "The Local Residential Land Use Regulatory Environment Across U.S. Housing Markets: Evidence from a New Wharton Index." Working Paper 26573. Cambridge, MA: National Bureau of Economic Research.

Gyourko, Joseph, and Raven Molloy. 2014. "Regulation and Housing Supply." Working Paper 20536. Cambridge, MA: National Bureau of Economic Research.

Hsieh, Chang-Tai, and Enrico Moretti. 2019. "Housing Constraints and Spatial Misallocation." *American Economic Journal: Macroeconomics* 11 (2): 1–39.

Inturri, Giuseppe, Matteo Ignaccolo, Michela Le Pira, Salvatore Caprì. and Nadia Giuffrida. 2017. "Influence of Accessibility, Land Use and Transport Policies on the Transport Energy Dependence of a City." *Transportation Research Procedia* 25: 3273–85.

Kok, Nils, Paavo Monkkonen, and John Quigley. 2014. "Land Use Regulations and the Value of Land and Housing: An Intra-Metropolitan Analysis." *Journal of Urban Economics* 81: 136–148.

Massey, Douglas, and Jacob Rugh. 2018. "The Intersection of Race and Class: Zoning, Affordable Housing, and Segregation in U.S. Metropolitan Areas." In *The Fight for Fair Housing: Causes, Consequences, and Future Implications of the 1968 Federal Fair Housing Act*, 1st ed., edited by Gregory Squires. New York, NY: Routledge, pp. 245–265.

Mawhorter, Sarah, and Carolina Reid. 2018. "Local Housing Policies Across California: Presenting the Results of a New Statewide Survey." Berkeley, CA: Terner Center at the University of California, Berkeley.

Mleczko, Matthew, and Matthew Desmond. 2020. "Using Natural Language Processing to Construct a National Zoning and Land Use Database." In *2020 APPAM Fall Research Conference*. Washington, DC: Association for Public Policy Analysis and Management.

Nechamkin, Emma and Graham MacDonald. 2019. *Predicted Zoned Density Using Property Records*. Washington, DC: Urban Institute.

Pendall, Rolf. 2020. "Knowing What Land Use Regulations Localities Have 'On the Books' Can Reveal Regulatory Stringency—And Much More." *Journal of the American Planning Association* 86 (2): 264–265.

Pendall, Rolf, Lydia Lo, and Jake Wegmann. 2022. "Shifts Toward the Extremes." *Journal of the American Planning Association*, 88 (1): 55–66.

Ray, Rashawn, Andre Perry, David Harshbarger, Samantha Elizondo, and Alexandra Gibbons. 2021. *Homeownership, Racial Segregation, and Policy Solutions to Racial Wealth Equity*. Washington, DC: Brookings Institution.

Rothstein, Richard. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America*." New York, NY: Liveright Publishing.

Rothwell, Jonathan. 2012. "Housing Costs, Zoning, and Access to High-Scoring Schools." Washington, DC: Brookings Institution.

Stahl, Kenneth. 2021. "Home Rule and State Preemption of Local Land Use Control." *The Urban Lawyer* 179 (50): 1–34.

Swope, Carolyn, and Diana Hernandez. 2019. "Housing as a Determinant of Health Equity: A Conceptual Model." *Journal of Social Science and Medicine* 243: 112571.

Trounstine, Jessica. 2018. *Segregation by Design: Local Politics and Inequality in American Cities*. Boston, MA: Cambridge University Press.

Turner, Margery, Solomon Greene, Corianne Scally, Kathryn Reynolds, and Jung Choi. 2019. *What Would It Take to Ensure Quality, Affordable Housing for All in Communities of Opportunity?* Washington, DC: Urban Institute.

About the Authors

Judah Axelrod is a senior data scientist at the Urban Institute. He works in collaboration with Urban's Technology and Data Science team and Racial Equity Analytics Lab to provide technical analysis and support for policies that strive to mitigate structural racism.

Lydia Lo is a research associate in the Fair Housing, Land Use and Transportation division of the Urban Institute's Metropolitan Housing and Communities Policy Center. A quantitative and qualitative researcher, her research focuses on land use and zoning, racial equity, community development, and systems change. She helps run the Land Use Lab at Urban.

Sara C. Bronin is a professor at Cornell University on the planning, law, real estate, and architecture faculties and the director of the Legal Constructs Lab. A Mexican American architect and attorney, Professor Bronin's research focuses on how law and policy can foster more equitable, sustainable, well-designed, and connected places.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org