Predicting Zoned Density Using Property Records

Emma Nechamkin
Urban Institute

Graham MacDonald
Urban Institute

January 2019

The authors welcome feedback on this working paper. Please send all inquiries to gmacdonald@urban.org.

We would like to acknowledge Solomon Greene for his insightful comments and feedback on initial drafts.

Urban Institute working papers are circulated for discussion and comment. Though they may have been peer reviewed, they have not been formally edited by the Department of Editorial Services and Publications. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders.

Copyright © January 2019. Emma Nechamkin and Graham MacDonald. All rights reserved.

Abstract

Local zoning codes affect a broad range of policies. For example, they can constrain housing supply and create affordability challenges (Ikeda and Washington 2015; Glaeser and Ward 2006), contribute to housing segregation (Greene at al. 2017; Pendall 2000), and determine how well a jurisdiction responds to changes in demographics and climate (Micklow and Warner 2014; Nolon 2013).

Despite the importance of local zoning regulations, research that measures the restrictiveness of zoning laws across places and the impact of restrictive zoning across a range of outcomes is remarkably thin. One reason is that zoning codes are long, technical, and difficult to access. Consequently, there is no up-to-date national dataset or reliable standard practice used to compare zoning codes.

To address this gap, we explore whether it is possible to merge property assessment records with granular data on zoning policies to generate a model that "predicts" zoning regulations. Using such a model, we could build an accurate, publicly accessible dataset of zoning regulations across the US.

In this paper, we test this hypothesis with a proof of concept in Washington, DC. We combine property assessment records provided by Zillow (ZTRAX) with data from the Washington, DC, Office of Zoning, and using a random forest regression, predict zoning characteristics in zones with residences.

We find that we are able to use property records to predict density limits with a relatively high degree of accuracy. Our model itself is exploratory and has many limitations: it uses a relatively small set of zones and depends greatly on how properties with missing zone designations are assigned to zones. Additionally, model predictions are largely directional, and vary in accuracy by zone type. We focus on one variable in the zoning code—density limits—and one city—Washington, DC—and thus more work is needed to determine whether the model can generalize to different jurisdictions and zoning regulations. Yet, preliminary directional success suggests that pursuing this work at a larger scale would be fruitful, and it would offer significant benefit to researchers seeking to study the effects of zoning across a wide array of policy domains.

Introduction

Land use regulations shape our communities, from our nation's largest cities to our smallest towns. Cities and counties throughout the US decide whether land will be designated for commercial, agricultural, industrial, or residential use, among others. Zoning codes specify permitted uses and limits such as the density of residential space, the ratio of building area to land, and the minimum number of parking spaces per unit, to name a few. Despite the importance of zoning codes, there is remarkably little research on the effects of zoning regulations, largely because no up-to-date, comprehensive, and comparable dataset on zoning regulations exists.

Zoning codes have wide-ranging effects on many public policy issues, from housing affordability to the ability of a city to adapt to climate change. And their impact is far-reaching: among major US cities, only Houston is without a formal zoning code (Pendall 2006).

Despite the importance and prevalence of zoning regulations, almost no systematic, comparable, and current data exist (Pendall et al. 2006). For instance, researchers do not have access to comparable data across jurisdictions for even the most basic zoning limits, like type of development permitted on a parcel, parking requirements, or density limits.

Part of the reason for this difficulty is that zoning data are difficult to collect. Zoning information is often locked in complex text descriptions in online HTML or PDF documents, or in physical publications at local planning departments. Consequently, rigorous research on the impacts of more granular zoning code regulation is remarkably thin. Though there are smaller, geographic-specific zoning data available, research on land use regulation at a national scale typically relies on two primary sources: the Wharton Residential Land Use Regulatory Index (Gyorko, Saiz, and Summers 2008) and Pendall's 1994 and 2003 surveys of local land-use regulations (Pendall et al. 2006). However, these data lack granularity and have only been sporadically updated, which can impede research.

In an attempt to work toward building a dataset of better zoning data, in this paper, we present an experimental approach to predicting zoned density limits using property assessment data in Washington, DC.

Filling Gaps with Administrative Data

Despite the lack of nationally comparable data on local zoning, we do have access to a national private-sector dataset on property assessments,[1] typically collected by local governments. These data often detail the size, location, and associated zoning classification of each property, among other characteristics, to facilitate the valuation of properties for tax assessment purposes. Though these data contain the zoning classification, they do not contain data on the actual limits of the zoning code in a jurisdiction; for example, the data may detail a given property is zoned R-1, but not how many units may be built on R-1 properties. These data include most properties in the United States, though we cannot say with certainty that all are included, as county-level techniques for assessment and recording of real property differ from jurisdiction to jurisdiction.

Although there is no national set of zoning regulations, in our experience, many zoning codes are publicly accessible and can be read and documented by hand. In this paper, we attempt to use these property assessment records, specifically, the characteristics of properties within each zone, to predict zoned density limits in Washington, DC. We focus on modeling density limits for our proof-of-concept because density limitations have the potential to affect a broad number of important policy domains from housing affordability, to educational access, to greenhouse gas emissions.

In the long run, we hypothesize that we can expand this approach to model selected cities and counties, so the model learns a more generalizable set of rules from which to predict density (and other zoning limits) across different jurisdictions with different built environments and regulatory structures. Ultimately, we hope this study acts as a first step to building an accurate, generalizable, easy-to-update, and publicly accessible national dataset of zoning regulations. This methodological approach is similar to Salganik's concept of "amplified asking" (2017), first used by Blumenstock et al. (2015), in that it uses a large, readily available "big data" source to estimate or predict data that would otherwise be extremely difficult and time-consuming to collect.

Ultimately, such data could inform research on how cities can best leverage zoning to respond to climate change, or identify best practices across cities for using zoning to reduce the

[1] Data were provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the authors and do not reflect the position of Zillow Group.

risk of flooding during national disasters. It could also empower communities: community organizations and neighborhood advocates would be able to combine tabular data on zoning with other neighborhood indicators to understand the effect of complex zoning regulations in their area, along with how changes to existing law might improve outcomes (Greene and Pettit 2016).

Data and Methods

Data

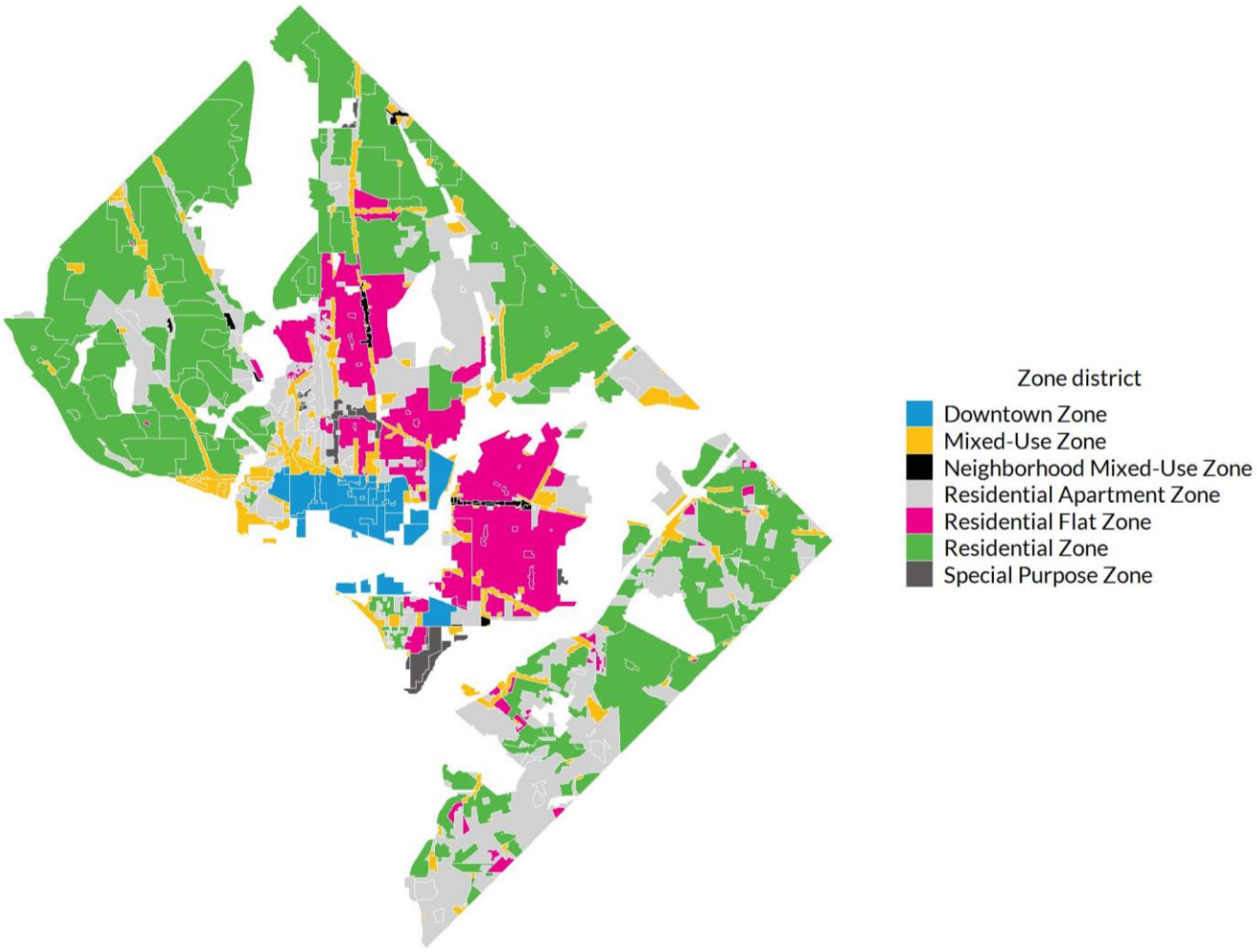
We use two datasets that cover the city of Washington, DC: manually coded information from the DC zoning code and property assessment data aggregated by zone. We provide summary information for these datasets in this section. For more detailed information, refer to appendix A.

Washington, DC, Office of Zoning (DCOZ) Data

We manually read and interpreted the DC zoning code to produce a tabular dataset of zoning regulations in DC. Washington, DC, has more than 100 zones across eight zoning districts: downtown; residential; residential apartment; residential flat; mixed-use; neighborhood mixed-use; production, distribution, and repair; and special purpose (figure 1). Zoning districts each have a distinct purpose and typically encompass zones with similar characteristics. Put simply, a single zone sets regulations on a specific geographical area, whereas a zoning district sets higher-level guidance on groups of similar single zones. For example, the residential district includes similar zones that primarily enforce low-density housing. Each of these zones has slightly different specific regulations but a similar purpose: to provide housing.

Figure 1. DC Has 139 Zones in 8 Districts

DC's zone districts, excluding production, distribution, and repair



Source: DC Office of Zoning zoning code.
Notes: Production, distribution, and repair zone districts are not pictured. Residential apartment, residential flat, and residential zones are separated for clarity.

The goal of our model is to predict the maximum floor area ratio (FAR) for the average residential property in each zone. FAR is a measure of density that specifies the area of building space permissible per unit area of lot space. However, zoning codes do not typically specify a single density limit for each zone. Instead, one zone may specify multiple density limits for different building heights or for different allowed building types or uses, such as buildings with affordable housing units. We simplify these varied density limits to derive a single permissible FAR per zone using the following assumptions:

- To simplify the model and our calculations of accuracy measures, for zones that have different regulations by subtype of buildings, we calculate the average of the regulations across the building subtypes. For example, if a zone has both 100-foot-tall buildings with a FAR maximum of 10 and 130-foot-tall buildings with a FAR maximum of 13, the FAR limit for the entire zone would be 11.5
- Historically, approximately 1 percent of properties in nearby Montgomery County have taken advantage of inclusionary zoning (IZ) bonuses, which allow additional density or height in exchange for ensuring a share of units are affordable (we use Montgomery County because DC's Inclusionary Zoning program was only initiated in 2009, during the Great Recession). For example, if a zone has a FAR maximum of 1 for a typical building and a FAR maximum of 1.2 for an IZ building, the FAR limit for the zone would be 1.002 (0.99 * 1 + 0.01 * 1.2). We derive the 1 percent figure from Montgomery County IZ statistics and US Census data.[2]
- We record by right regulations and ignore conditional use limits. Use by right refers to regulations in the zoning code that require only ministerial approval—that is, as long as the developer's drawings meet building code and other ministerial approvals and she pays the appropriate fees, the city must allow her to build. Conditional use, on the other hand, typically requires special approval by a local planning commission for the project to proceed.

Property Assessment Data

In this section, we describe two separate models:

- *General Model:* A model built with data that are generally available in the ZTRAX data across jurisdictions, and is therefore potentially replicable to other jurisdictions.
- *DC-Specific Model*: A model built with ZTRAX and manually collected DC-specific data we hypothesized might improve the performance of the model, but would also reduce the ability to replicate it across jurisdictions.

2. "Number of MPDUs Produced Since 1976," Montgomery County Department of Housing and Community Affairs, accessed September 7, 2018, https://www.montgomerycountymd.gov/DHCA/housing/singlefamily/mpdu/produced.html.

Creating the Dataset: General Model

We use 2016 Zillow ZTRAX data on property assessments collected from local assessors and recorders in the US and select only the records within Washington, DC. We clean the assessment data and aggregate it to a single record per zone—in other words, each row in the dataset represents summary information about all the properties in a single zoning designation in DC, such as R-1. After aggregating the data, we merge it with the manually transcribed density limits created from the DC Office of Zoning, then build a model that uses the aggregated zone-level characteristics to predict our transcribed density limits.

We use information from the ZTRAX dataset to inform our model. More concretely, we use build and remodel year, land use designations, and property counts to model larger zone characteristics. For example, we use the share of land by area used in each zone for high-, medium-, and low-density residences. We use the following ZTRAX fields to inform our data model:

- Lot Size: the area in square feet of each property
- Standard Land Use Code: a standardized code that describes how land is used; for example, the county code CR123 refers to a vintage vehicle service station
- Year Remodeled: the last major remodel to the building
- Year Built: the year a building was erected
- Geographic descriptors, like property longitude, latitude, legal block, address, and lot number

We use the following variables from the assessment data to construct our final dataset for modeling:

- The minimum build year and remodel year for all properties in the zone
- The maximum build year and remodel year for all properties in the zone
- The mean build year and remodel year for all properties in the zone
- The zone district type, determined by the most common land use type by area from commercial office, commercial retail, governmental, industrial, residential, residential multifamily, and vacant land; these land use designations are derived from standard county assessment records contained in the ZTRAX data

- The share of high-, medium-, and low-density homes by lot area, derived from the standard land use designations, which detail building type; see appendix C for more detail
- The total amount of residential lot area in the zone
- The average lot area in the zone per residential property
- The number of residential and total properties within the zone
- The share of neighboring zones categorized as high, low, or medium density based on the county-level land use code; see appendix C for more detail
- The number of neighboring zones in total and that were categorized as residential neighbors, using the standard land use code
- The share of residential land in the zone, defined as properties with residential land use category descriptions

When summarizing data at the property level before aggregating by zone, we make some simplifying assumptions. To address whether a record in the original ZTRAX dataset corresponds to a single property, we assume that all records with an identical full address (i.e., street number and street name) belong to the same property. Note that this methodology incorrectly categorizes single developments with multiple buildings numbered with different addresses. This method also incorrectly categorizes some properties with no address number, accounting for about 14,000 total records. Because we cannot easily verify or correct these records, we apply a few methods detailed in the following paragraphs to correct for the most obvious errors.

After aggregating to the property level, we use each property's latitude and longitude to assign it to a zone. Ideally, we would have used ZTRAX data alone and would have assigned properties that were missing zones to the correct zones based on the zones of their five nearest neighbors (for more details on the accuracy of this method, see appendix D). Unfortunately, because DC zones were reassigned in 2016 in a way that is incompatible with ZTRAX data, we instead have to match property latitude and longitude with zones using the DCOZ zoning map.[3]

[3] Table 1 shows average counts of residential and overall properties per zone district. We received ZTRAX data from Zillow as a one-time delivery before the assessment data for DC were updated with the new zoning codes. However, with updated assessment data, we could assign the missing zones based on their nearest neighbors without resorting to using nongeneralizable data like the DCOZ zoning map.

Because private data holders allow the purchase of up-to-date records, we do not believe this will be an issue for applying this approach more generally.

Table 1. Residential Districts Have the Highest Average Number and Share of Residential Properties

Average number of residential properties and all properties within each zone by DC zone district for generalized model

District type (zone count)	Average number of residential properties	Average number of properties
Downtown (9)	34.2	259.3
Mixed-use (25)	140.2	332.6
Neighborhood mixed-use (16)	15.3	49.9
Residential apartment (10)	1,597.7	1,842.2
Residential flat (3)	11,387.0	12,089.7
Residential (16)	3,798.3	4,102.3
Special purpose (20)	28.9	72.8

Sources: DCOZ and ZTRAX data.

To aggregate property records into zones, we have to first assign each record in the ZTRAX dataset to a single property. Often, buildings like condominiums cause assessment datasets to contain many records for a single property—one for each condo unit, in this example. If we fail to reduce these records into a single property, we might misrepresent a zone with one condo building (containing 100 condos, with one record per condo) and ten office buildings as predominantly residential, for example. We combine multiple property records into a single property records (e.g., the case of multiple condos aggregated to a single building) as follows:

- We use median values for lot number, latitude, and longitude from component records when aggregating records into a single building. In other words, if four records all pertain to the same physical building, our calculated lot number, latitude, and longitude for the building would represent the approximate center of those four records.
- We predict zone and land use categories by property by calculating the most common values of zone and land use category across all property records in a single building and extract the most likely zone and land use for a property. For example, if 10 records are associated with a given building, 9 of which are residential and 1 of which is commercial, we assign the building as residential. In cases of a tie, we use a stable

sorting algorithm to choose the first zone and land use category presented by the sort algorithm.

- We use the last possible year among year built and year remodeled to extract the most recent construction information. If a condo had five lots, for example, the last one to be remodeled would provide the remodel year for the building.
- We calculate lot size as the sum of lot size for condo lots and as the median lot size among records for all other lot types.

We then assign each property to its correct zone, which we need to calculate for two reasons. First, the assessment data include a field for property zone, but this field was left blank approximately 10 percent of the time. Second, even properties with assigned zones were outdated. As a result zoning updates in 2016, the assessment data would have mischaracterized or left out entire zones, such as R-20. We use the provided latitude and longitude, in combination with a map of DC's current zones, to assign each property to its correct zone. When we could not determine the latitude and longitude because of an incomplete address or missing data, we use an updated version of the ZTRAX zone by leveraging a DC Office of Zoning crosswalk of old zone designations to new zone designations.

We do not include zones in DC to which we could not assign residential properties. The majority of zones we exclude should not contain homes (for example, certain industrial zones). Other zones should have residential properties, but we are unable to properly assign them (they are primarily small, special purpose zones). As a result, we do not include every DC zone.

Finally, we impute any missing data and aggregate the data by zone, calculating the following variables: the average, maximum, and minimum build years and remodel years; the share of residences in high-, medium-, and low-density dwellings; the amount of total living space; the amount of space per property; the number of residential properties; the number of total properties; and the type of zone. These characteristics are calculated by computing either direct averages or averages weighted by total lot size. We impute values by designated group (either zone or land use code) using the best information possible for each variable. For example, we impute remodel year for residential buildings based on the average remodel year for all residential buildings, but we assign remodel year to the original build year if the build year is available. Our imputation methods are detailed in appendix B.

In addition to aggregating zone characteristics, we want to determine the predictive value of the characteristics of each zone's neighbors. We hope such a measurement would separate high-density downtown areas from low-density residential areas without additional geographic information—for example, without knowing the location of a city's downtown. To do this, we determine the boundary properties in each zone and use their latitude and longitude to determine all other zones within a small radius—about 250 feet for our general model. Then, we aggregate characteristics across neighboring zones to calculate the total number of zones nearby, the total number of residential zones nearby, the share of zones that are residential zones nearby, and the share of residential zones at each density level (high, medium, and low) nearby. For more information on our method for calculating statistics for zone neighbors, please see appendix E.

Creating the Dataset: DC-Specific Model

To test whether local data provide a more accurate prediction of density limits, we also run a model using publicly available data from DC from DC's Office of Zoning handbook and DC's open data portal's tax record and property lot descriptions. In our DC-specific model, we use a multistep process to assign properties to their correct zones using zone geometries that leverage DC's block, square, and parcel numbering system.

DC properties have identifying numbers to describe not only the geographic location of each property, but also the type of record. Each type of record has a target number range. For example, tax lot records, which exist as unique property identifiers for tax purposes only, have a different number range than the unique property records. Because each number is unique, we can identify which lot numbers referred to the physical lot of a property as opposed to the tax or condo lot for a given property. This additional information, for example, enables us to more easily identify properties in which multiple buildings are actually condos and not separate properties. In combination, we use these identifying numbers to consolidate property entries into a single entry per property.

After consolidating the dataset so each record corresponds to a single property, we use property lots, latitudes and longitudes and a map of zone boundaries from DCOZ to assign records to zones. For more information on zone assignment, please see appendix H. Table 2 shows average counts of residential and overall properties per zone district.

Table 2. Residential Districts Have the Highest Average Number of Residences

Average number of residential properties and all properties by DC zone district for DC-specific model

District type (zone count)	Average number of residential properties	Average number of properties
Downtown (9)	29.8	259.1
Mixed-use (24)	143.6	358.8
Neighborhood mixed-use (16)	15.5	52.1
Residential apartment (10)	1,625.7	1,960.8
Residential flat (3)	11,662.3	12,521.7
Residential (16)	3,860.5	4,257.0
Special purpose (15)	30.2	78.3

Sources: DCOZ and ZTRAX data.

Merging the Datasets to Create a Composite Data Source for Modeling

To produce a model-ready dataset, we merge our zone-level aggregated assessor data with our dataset of the DC zoning code and its associated density limits. We only include zones with residential properties. In this context, "zones with residential properties" refers to a zone that has both the legal capacity for residences (e.g., residences are allowed in the zoning code) and at least one residential property from ZTRAX. For our generalized model, the particular zones that we include, as well as their total number of properties, are shown in appendix A.

Predicting Zoning Regulations

Our model uses zone-level property assessment data to predict floor area ratio maximums for residential buildings in DC. It primarily serves as proof of concept, demonstrating that predicting density information is possible.

Our data have a small number of observations and many collinear features. We manually select the 27 features in the general model (changing zoning designations in the DC-specific model), described in the previous section, and use a LASSO regression to penalize features that have minimal impact on the outcome variable. We subsequently remove the most penalized features automatically to further reduce the total number of features used in the model.

Next, we run a random forest regression with the selected features to predict the FAR maximum for residential zones in DC. A random forest regression allows us to model nonlinear relationships in the data by using a collection of decision trees to predict a given outcome. A

single decision tree may not generalize well, but using a random forest of decision trees limits overfitting—and increases our ability to generalize the results—while decreasing overall error.

To increase the probability that our model will generalize outside our limited sample, the data are split into two separate datasets: a training set, on which the model is trained, and a test set, on which the model's predictive power is tested "out of sample." Testing ensures that the model is able to predict not only data it has seen but also new data. In the future, we plan to train our model on a nationally representative sample of cities to test whether it can produce accurate results for new cities that the model has not yet analyzed.

Because our dataset is small and heterogeneous, our model is extremely sensitive to the specific split between our test and our training sets. We use Leave One Out Cross Validation (LOOCV) to decrease the model's sensitivity. In LOOCV, we run many models, each with just one data point as test set. We run 94 models total, each with its own LASSO and random forest regression, and report our average results across these models. In other words, in each of our models, we predict a single zone's FAR, then average all predicted FARs for all models.

Measuring Accuracy

To determine the accuracy of our model, we must define how we measure error. As is typical in prediction problems that use regression, we report root mean squared error (RMSE) and mean absolute error (MAE). By default, these methods weight each zone equally. Because these metrics may not give us the best measure of "accuracy" that we care about, we present two alternative metrics to capture our model's accuracy and error that we believe are potentially more useful to researchers and policymakers. The mathematical formulas for these alternative metrics are included in appendix F.

Standard error metrics, like RMSE and MAE, measure absolute difference between predicted values and actual values. For FAR, however, relative deviation may be equally, if not more, important than absolute deviation. Consider two zones, one with an actual FAR of 10 and one with an actual FAR of 1. An absolute deviation of 1 is a much more significant error for the low-density zone than for the high-density zone in the case of a researcher looking to classify zones by density category. If our model predicts a FAR of 2 instead of 1, a researcher might categorize the zone as low-moderate density instead of low-density, while if the model predicted an FAR of 11 instead of 10, the zone would still be classified as high density in either case.

As a result, we define an additional metric, the relative absolute error, to show our model's relative error. In this example, the high-density zone would have a relative absolute error of 10 percent, while the low-density zone would have a relative absolute error of 100 percent.

We also introduce weighted metrics to represent the importance of predicting each zone correctly. Because we predict residential FAR maximums, zones with many more residential properties may be more important—depending on the number of units per property—than zones with fewer residential properties, because they control comparatively more residential development. Consequently, we also provide an error statistic that weights the error by the number of residential properties per zone. We focus on the weighted relative mean average error, defined as the average of absolute deviations from the true FAR limits (weighted by number of residential properties), which we believe best captures accuracy for the purposes of researchers and policymakers that might use the model's predictions. For example, a zone with a true FAR of 1 and a predicted FAR of 2 would have an unweighted relative absolute error of 1. In this analysis, we weight each zone's relative absolute error by its share of residential properties to calculate the weighted relative mean average error. Ideally, we would weight the error metric by the number of units, not the number of properties, however number of units is not available in the ZTRAX data.

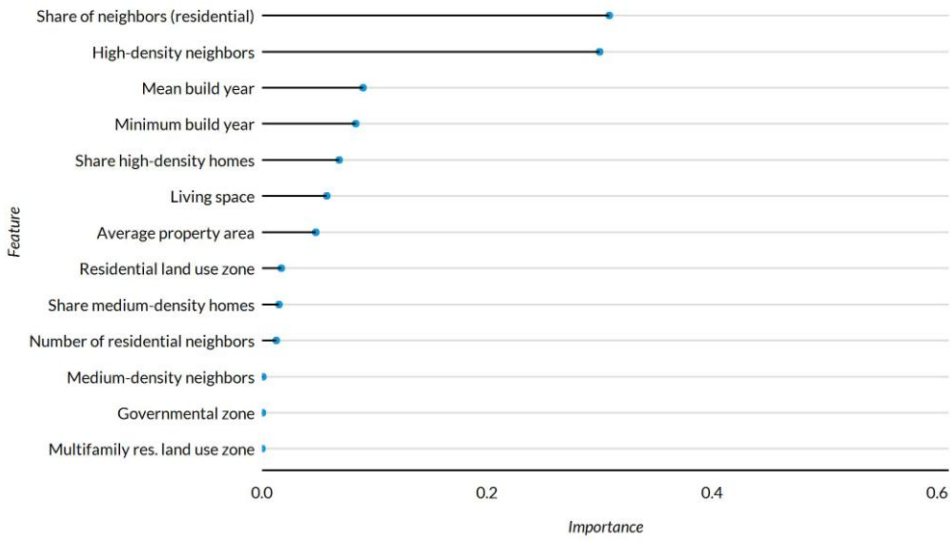
Results and Discussion

To test whether zoning characteristics can be predicted from property assessment records, we run a random forest regression using data from Washington, DC.

The average feature importance for each variable of across all of our models is shown below. Figure 2 shows the most important features for our generalized model. Note that our DC and generalized models have different features that represent similar information (figure 3). For example, the share of high-density housing among adjacent districts' residential property is reasonably analogous to a downtown classification, if we assume that downtown areas tend to be surrounded by mid- or high-density areas, as city planners are less likely to place large office buildings next to single-family homes. For both models, the final features shown in each graph have a negligible impact overall, as they are included in very few models.

Figure 2. General Model: Zone Neighbors, Zone Density, and Build Year Are Most Predictive of FAR Limits

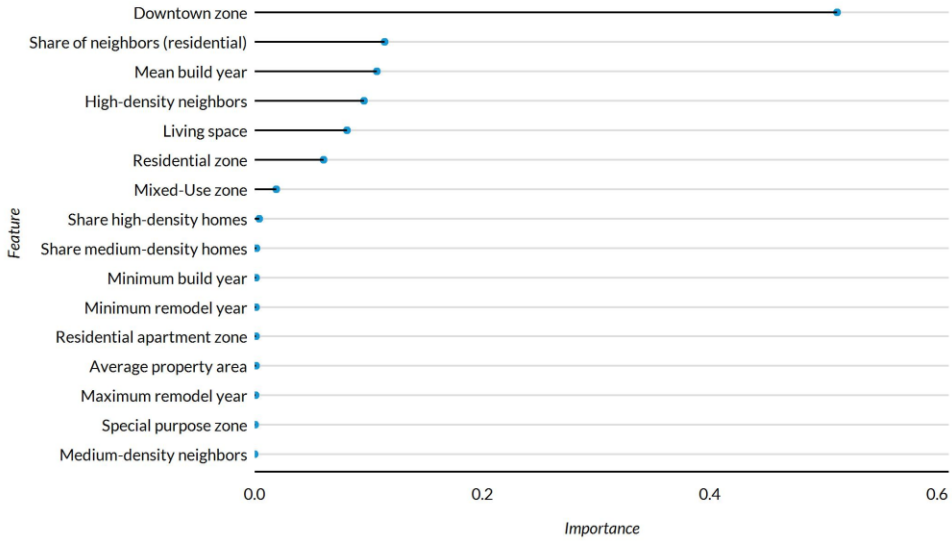
Average feature importance for the LOOCV random forest regression on our generalizable dataset



Sources: DCOZ and ZTRAX data.
Note: Feature names correspond with their descriptions above.

Figure 3. DC-Specific Model: Zone Neighbors, Zone Type, and Build Year Are Most Predictive of FAR Limits

Average feature importance for the LOOCV random forest regression on our DC-specific dataset



Sources: DCOZ and ZTRAX data.

We define the features in the graphs as follows. For more information on variable definitions, see the Data and Methods section.

- *Share of neighbors (residential):* the share of neighboring zones categorized as residential
- *High-density neighbors:* the share of neighbors categorized as high density
- *Minimum build year:* the earliest build year for the oldest property in the zone
- *Mean build year:* the average build year for a property in the zone
- *Share high-density homes:* the share of homes that are high density
- *Living space:* the total amount of living space
- *Average property area:* the average amount of land per property
- *Residential land use zone:* a zone categorized as "residential land"
- *Share medium-density homes:* the share of homes that are medium density
- *Number of residential neighbors:* the number of neighboring zones that are residential
- *Medium-density neighbors:* the share of neighboring zones with medium density
- *Governmental zone:* a zone categorized as "governmental"
- *Multifamily residential land use zone:* a zone categorized as "multifamily residential"

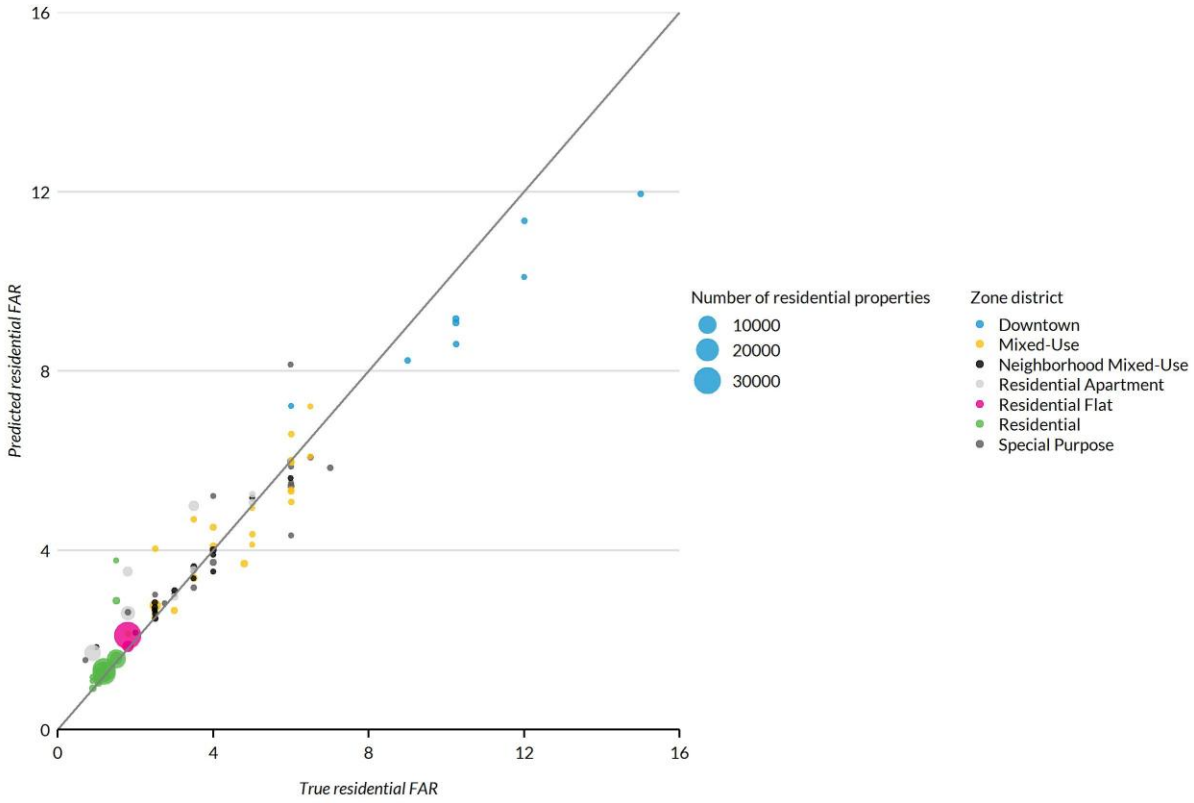
The features we find most important are in line with our expectations. We know downtown areas tend to be high density and that housing type tends to be clustered, so housing is often adjacent to similarly dense housing. And housing is typically next to housing, so it makes sense that the share of neighboring zones that are residential has a significant impact on the model. Furthermore, in DC, many downtown apartment complexes are new; our most important features reflect that. Build year is also logical; it likely helps the model understand which housing was built decades ago under older zoning law, and which was built while the current (or most recent) code was in effect. In the DC-specific model, because the downtown zone has such different density restrictions from other zones, and has a number of high-density neighbors, it appears to act as a proxy for the share of high-density neighbors in the general model.

Our general model predicts residential FAR limits across zone types very well. Figure 4 plots true (or actual) FAR limits against our model's predicted FAR limits for each zone. Data are colored by zone district and sized by the total number of residential properties. Perfectly predicted zones fall on the gray diagonal line; the farther points lie from the line, the higher the

error rate in our prediction. Our general model shows promise for creating granular geographic data, and predicts high- and low-density geographic areas accurately.

Figure 4. The Generalized Model Is Able to Predict FAR Well across Zones

Predicted residential FAR versus true residential FAR



Source: ZTRAX data, DCOZ zoning code.

Notice that the generalized model does not perform altogether that much worse than our DC-specific model, even though our DC-specific model is highly tailored to the DC area using many local, publicly available data sources (table 3).

Table 3. Models Have Roughly Comparable Error

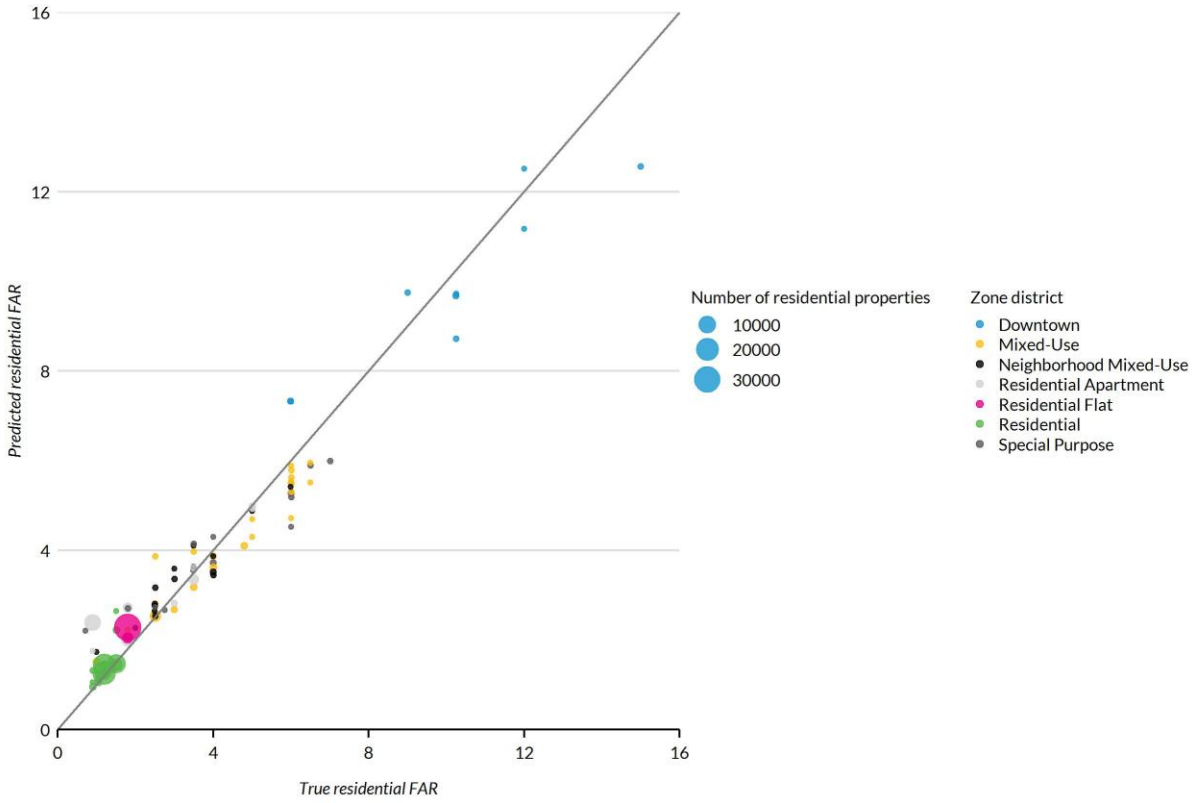
Error metrics for generalized and DC-specific model

	Traditional error metrics		Adjusted error metrics		
	RMSE	MAE	Weighted RMSE	Relative MAE	Weighted relative MAE
Generalized model	0.794	0.521	0.412	0.183	0.187
DC-specific model	0.655	0.482	0.500	0.186	0.255

Results for the DC-specific model are shown in figure 5. Unlike in our generalized model, we do not deterministically tie-break property-level aggregations in our DC model, so results vary slightly by run.

Figure 5. The DC-Specific Model Only Marginally Outperforms the Generalized Model

Predicted residential FAR versus true residential FAR



Sources: ZTRAX data and DCOZ zoning code.

Owing to the similarity between our two models' predictions and the more easily generalizable nature of our less-specific model, we will only discuss the general model results moving forward. For more results from the DC-specific model, see appendix G.

Our generalized model performs best in residential districts, and generally performs well in zones with a high number of residential properties. Overall, results are directionally correct across zones: zones with high density are predicted to have higher FAR limits. However, the model is less accurate for zoning districts with a wide range of FAR limits, such as downtown zones: it predicts average FAR per zoning district more accurately than the highest or the lowest

FARs within that district. We suspect this result is driven by the relative paucity of high-density zones, which could be solved by expanding the model to cover additional jurisdictions.

Our model has relatively low error rates by zone district, for most zones, and across DC by land area. It predicts FAR within 20 percent of its true value for most zones in our dataset (table 4). Our model performs worst in relative terms on residential apartment zones and on residential flat zones. About one-third of residential apartment zones have very few properties. There are only three residential flat zones and our model incorrectly predicts one of them, perhaps because a low share of residential properties exists in that zone. Our model also performs relatively poorly in mixed-use and special purpose zones, likely due to the wide range of zoning limits proscribed by these zones.

Table 4. Model Predictions Fall within 20 percent of True Values on Weighted Relative MAE

Error metrics for overall model and DC zone districts

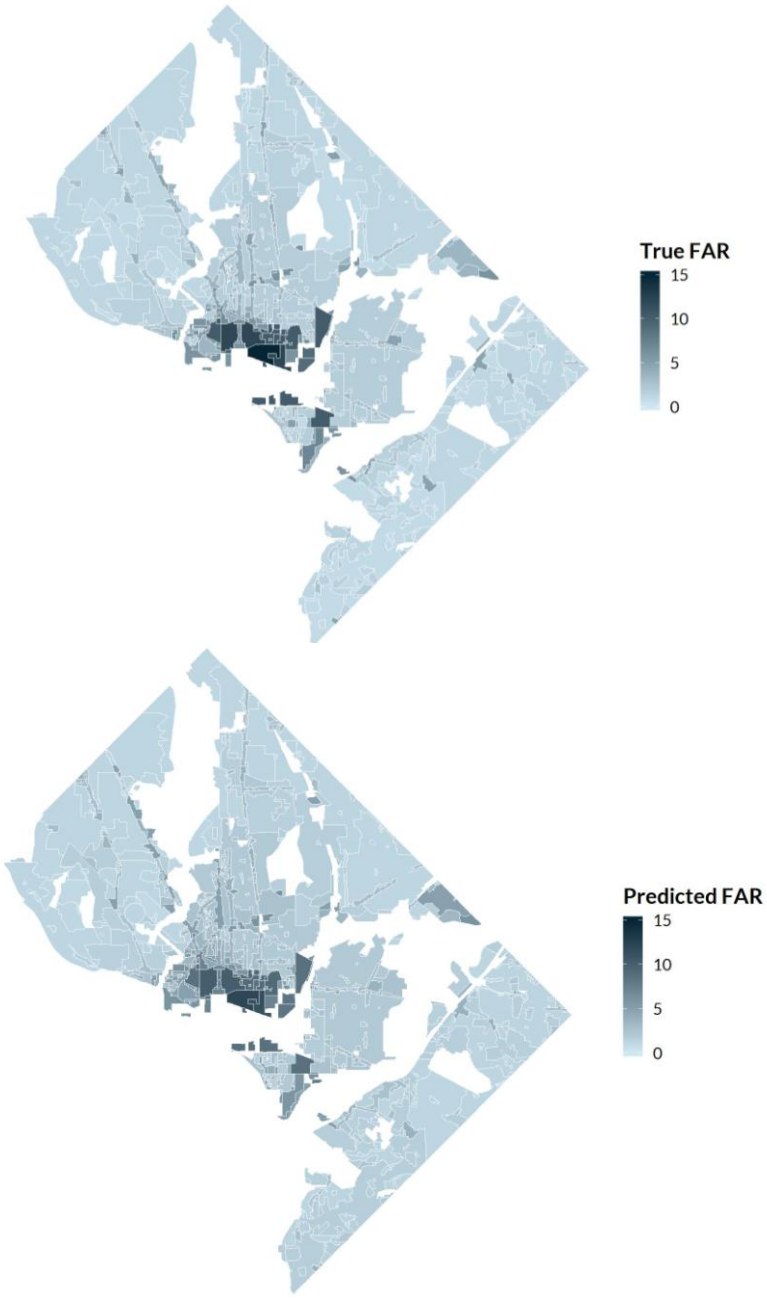
	Traditional error metrics		Adjusted error metrics		
	RMSE	**MAE**	**Weighted RMSE**	**Relative MAE**	**Weighted Relative MAE**
Overall model	**0.794**	**0.521**	**0.412**	**0.183**	**0.187**
Performance by zone					
Downtown	1.513	1.277	1.190	0.121	0.097
Mixed-use	0.619	0.459	0.388	0.119	0.113
Neighborhood mixed-use	0.296	0.212	0.200	0.102	0.055
Residential apartment	0.862	0.623	0.960	0.388	0.690
Residential flat	0.181	0.149	0.286	0.083	0.156
Residential	0.679	0.331	0.127	0.247	0.077
Special purpose	0.782	0.626	0.655	0.260	0.125

Sources: DCOZ and ZTRAX data.

Figure 6 shows predicted and true FAR in Washington, DC, geographically by zone for the generalized model. Dark zones have high residential FARs and density allowances, and light zones have low residential FARs and density allowances.

Figure 6. The Generalized Model Correctly Identifies High-Density and Low-Density Zones

FAR by DC zone



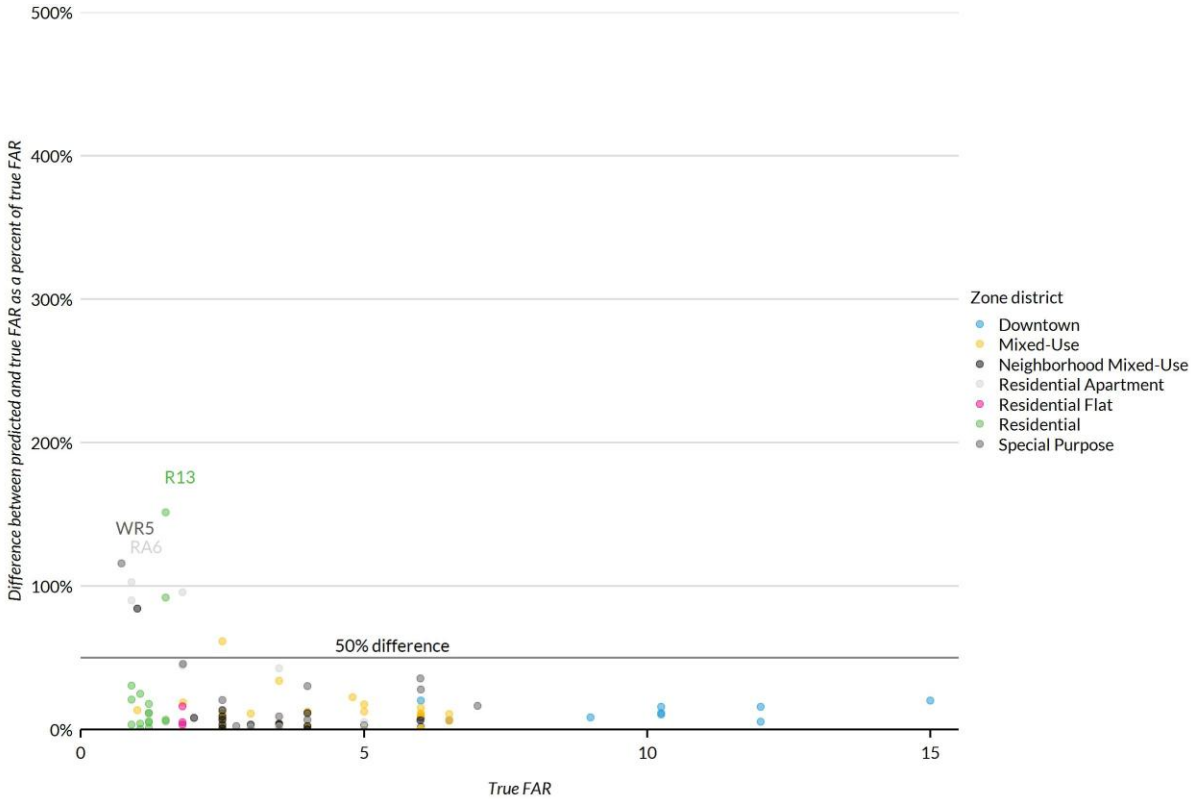
Sources: ZTRAX data and DCOZ zoning code.

Within zoning districts, most zones have very low relative error rates, with a few exceptions. Figures 7 and 8 show zone-level data on actual FAR plotted against both absolute error and relative error. Most zones' predictions are within 50 percent of true FAR; zones with a

higher relative deviation have smaller true FARs, largely because a small deviation produces a large relative difference. However, several zones with small true FARs, like residential zones, generally have extremely small relative error as well (figure 9).

Figure 7. Most Predictions Are within 50 Percent of True FAR

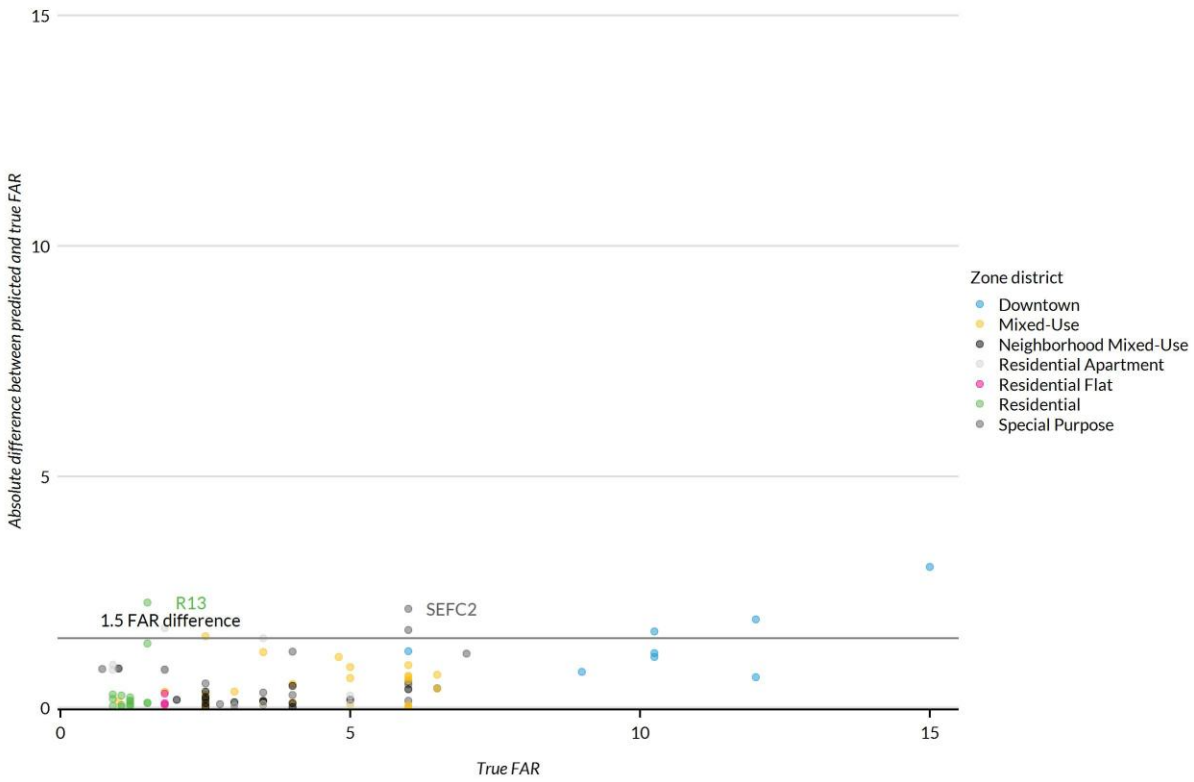
Share error for each zone's predicted FAR



Sources: ZTRAX data and DCOZ zoning code.

Figure 8. The Vast Majority of Zones Are Predicted Within 1.5 of True FAR

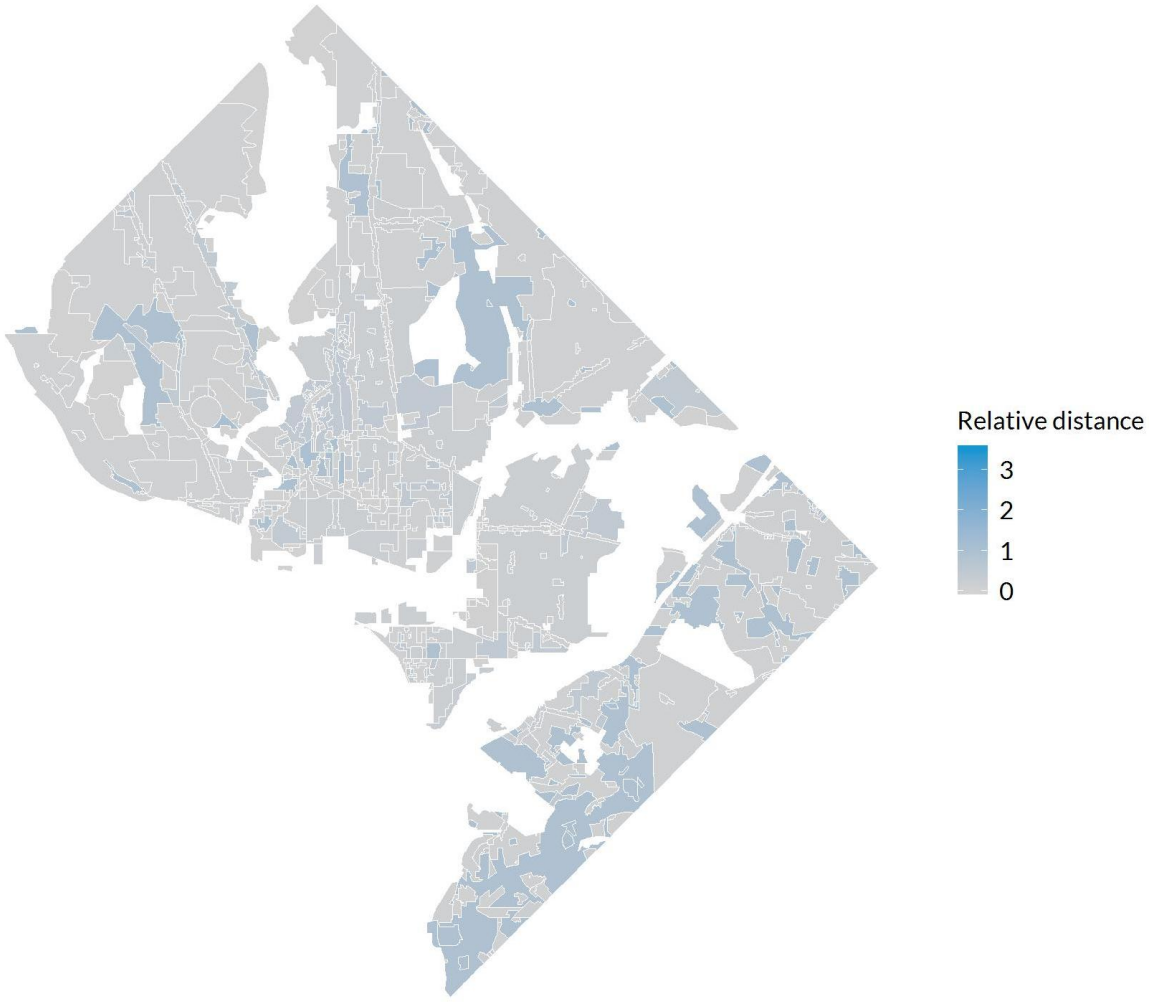
Absolute difference between predicted and true FAR



Sources: ZTRAX data and DCOZ zoning code.

Figure 9. The Generalized Model Has Low Relative Error, with a Few Exceptions

Relative distance from predicted to true FAR



Sources: ZTRAX data and DCOZ zoning code.

Model Limitations

Our model faces several issues, some of which we believe could be remedied by a dataset that spans multiple cities.

First and foremost, the model suffers from a lack of data. Our model is highly sensitive to small changes in zoning information, as there are few zones represented. Not only does the model have fewer than 100 data points (one for every allowed zone), many zones have few properties, amplifying any noise and inaccuracies in the data. Because our data come exclusively from DC, we are unable to increase the size of our dataset for this model. In the future, we hope

to develop a more robust, generalizable model by adding more cities. Expanding to additional cities will provide our model with more zones and allow us to better generalize across the US.

The type of data included in this model also present a limitation. Assessment data in DC do not reliably include dimensions—building height, for example, and lot area, which we found difficult to accurately calculate. We would hope to supplement our existing data with another dataset to remedy this. Satellite imagery, for example, could enable us to measure set-back per property or height for all properties. We believe that additional and reliable information about physical characteristics of buildings in zones could create a more robust model.

We depend heavily on the accuracy of some ZTRAX variables, such as latitude and longitude. Because we use latitude and longitude precise to five decimal places in our analysis, even small errors can negatively impact the quality of our data. The current model does not assign properties that are missing zones to zones based on the zones of their neighbors. We would like to make this change in future iterations of the model to make the results more generalizable. This means the accuracy of extant ZTRAX zone categorization, in combination with the rate of missing zones, is highly important. We have investigated alternative, private data holdings of property-level data and believe the regularly updated data provided by these groups, or regularly updated data provided by ZTRAX, would resolve this issue.

Because we require each property to have a zone designation, we eliminate records with potentially valuable information. In both our generalized and our DC-specific models, we dropped several records from our analysis, because they lacked either geographic information, address information, lot information, or existing zone information. Although the total number of dropped properties is small, we cannot measure how this effect is distributed across zones; it could disproportionately impact certain zones, especially those with fewer properties. Single zones with few properties were highly sensitive to dropped properties. More information on the characteristics of zones included in each model can be found in appendix A.

Additional limitations relate to our methods for assessing the accuracy of the data. It is almost impossible to measure the degree of error when assigning properties to zones because we did not have access to a perfectly labeled dataset of every property in DC. Most important, because of the DCOZ zoning code update, many assessor-assigned zones are now out of date. For example, had we strictly used ZTRAX zoning data, certain new zones, like R-20, would not even exist in the ZTRAX dataset, making it even harder to assess the accuracy of our data

cleaning procedures. And in our general model, we found that properties without street numbers (e.g., New York Ave., as opposed to 100 New York Ave.) were sometimes incorrectly labeled.

To obtain an approximate measure of our assignment error, we looked at a randomized subset of specific cases where the assessor-assigned zone did not match the zone we assigned each property and manually determined the proper zone. In both our models, our assignment was incorrect for about 10 percent of the mismatches; many mismatches were due to renamed zones with multiple designations. When ZTRAX and our dataset agreed on at least part of the zone description, we ignored the mismatch and treated it as a correct match. We found about 37,000 mismatching records in our generalized model and about 35,000 in our DC-specific model; we extrapolate these to 3,700 and 3,500 errors, respectively, using a 10 percent error rate. These totals represent less than 5 percent of the properties in our dataset. Across both our models, we found that error in zone assignment is most severe at zone boundaries.

Potential misclassifications are more damaging in small zones, which make up a substantial portion of the zones in DC. In our models, 46 of the 94 zones that we are able to include contain fewer than 100 properties. For small zones, a few properties can dramatically change overall zone characteristics. For example, our modelled D-2 zone has fewer than 10 residential properties, so an increase of a single residential property is nontrivial—a single property could change residential features significantly and increase the share of residential properties by 10 percent. We did not drop these zones from our analysis in order to maximize our coverage of zones in the city.

Incorrect assignment for boundary properties could also have more nuanced impacts on our model. For example, because we used boundary properties to determine neighbors for a given zone, we could incorrectly assign neighbors to each zone. We have calculated the approximate error for nearby zone assignment, however, and believe that our method is reasonably accurate for most zones: most zones correctly identify around 95 percent of their true neighbors and incorrectly classify about 25 percent of all neighbors. See appendix E for more information.

Because our DC-specific model required local zoning maps, which are not publicly available for every jurisdiction in the US, the DC-specific approach may not generalize nationally and could thus curtail our ability to generate a comprehensive national dataset. We address this limitation by running a generalized model, as discussed in the previous sections. The

generalized model may not be as accurate when predicting zoning in jurisdictions outside DC, as it likely learns DC-specific characteristics. Because we have not trained it outside DC, records in some states may have less complete information, and the model may not generalize to suburban or rural jurisdictions, or cities with significantly different built environments. Therefore, we recommend expanding our methodology to train the model on multiple different, diverse jurisdictions in the future.

In this study, we only predict one aspect of zoning—floor area ratio limits—and standardize other density limitations, such as height, so density is measured in a standard unit across zones. However, it may not be possible, or may be difficult, to translate limits in other jurisdictions to a standard unit like FAR, especially in areas that use form-based codes or incentive-based zoning; in such cases the model may be less reliable. In addition, zoning limitations reach far beyond just density limits: other important variables may include parking, mixed-use, and set-back requirements. It is unclear whether this model will work equally well to predict these important zoning features, and thus we recommend expanding our methodology to additional zoning limits in the future.

Conclusion

Our results demonstrate that FAR limits in most zones in DC can be predicted from property assessment records. Our model predicts over 90 percent of zones to within 50 percent and most zones (about 76 percent) to within 20 percent of the actual FAR limits established in the zoning code. Well-predicted zones account for the majority of the city by land area and residential property count, and our model is largely directionally accurate: zones with higher density limits are predicted as such. Moreover, the model provides granular, zone-level data on density that is difficult to collect and is not available in zoning data commonly used for research, such as Pendall's (2006) surveys and the Wharton Land Use Regulatory Index (2007).

The similarities between our generalized model and our DC-specific model suggest that our exploratory work may generalize on a national level using information from property assessment records alone. Our work suggests that creating a national database of zoning information from property assessment data may be feasible, given additional data collection and modeling. However, we must note that this work is exploratory, and that we only modeled

density through FAR limits and focused solely on DC. More work is needed to ensure that this model can generalize outside Washington, DC, to different types of jurisdictions and zoning regulations.

Appendix A: Information about the Modeling Datasets FTB

The final dataset for modeling is summarized below, as well as the true number of zones with residences provided by DCOZ. Table A.1 shows which zones with fewer than 100 properties were included in the generalized model, as well as how many residential properties and total properties were included in each zone, to demonstrate which zones may be particularly sensitive to small changes in assumptions.

Table A.1. Generalized Model: Most Zones Have Few Properties, and Few Zones Have Most Properties

Zones with fewer than 100 properties

Zone	Number of properties	Number of residential properties	Zone	Number of properties	Number of residential properties
ARTS1	30	<10	NC15	43	<10
CG1	<10	<10	NC16	96	25
CG2	23	<10	NC17	<10	<10
CG3	12	<10	NC2	42	<10
CG5	24	<10	NC3	46	<10
CG7	<10	<10	NC5	40	19
D1R	81	54	NC6	43	<10
MU14	56	17	NC9	42	<10
MU16	40	<10	R10	13	11
MU17	33	14	R11	14	<10
MU18	63	<10	R12	89	72
MU19	25	<10	R13	<10	<10
MU2	46	15	RA10	45	35
MU20	73	<10	RA6	12	<10
MU21	80	<10	RA9	59	51
MU22	12	<10	RC1	43	31
MU24	69	29	RC3	26	15
MU27	54	<10	SEFC2	<10	<10
MU29	<10	<10	WR3	<10	<10
MU8	35	<10	WR5	<10	<10
MU9	18	<10	NC12	<10	<10
NC1	18	<10	NC13	<10	<10
NC10	38	<10			
NC11	<10	<10			

Source: DCOZ and ZTRAX data

Notes: Data truncated to <10 due to data use restrictions.

Table A.2. Zoning Districts Include Similar Zones

Zones by zoning district in the Washington, DC zoning code

Zone district	Includes residential	Zones
Special Purpose Zone	Yes	ARTS1, ARTS2, ARTS3, ARTS4, CG1, CG2, CG3, CG4, CG5, CG6, CG7, HE1, HE2, HE3, HE4 RC1, RC2, RC, SEFC1, SEFC2, SEFC3, SEFC4, StE1, StE10, StE11, StE12, StE13, StE14, StE15, StE16, StE17, StE18, StE19, StE2, StE3, StE4, StE5, StE6, StE7, StE8, StE9, USN, WR1, WR2, WR3, WR4, WR5, WR6
Downtown Zone	Yes	D1R, D2, D3, D4, D4R, D5, D5R, D6, D6R, D7, D8
Mixed-Use Zone	Yes	MU1, MU10, MU11, MU12, MU13, MU14, MU15, MU16, MU17, MU18, MU19, MU2, MU20, MU21, MU22, MU23, MU24, MU25, MU26, MU27, MU28, MU29, MU3, MU4, MU5A, MU5B, MU6, MU7, MU8, MU9
Neighborhood Mixed-Use Zone	Yes	NC1, NC10, NC11, NC12, NC13, NC14, NC15, NC16, NC17, NC2, NC3, NC4, NC5, NC6, NC7, NC8, NC9
Production, Distribution, and Repair Zone	No	PDR1, PDR2, PDR3, PDR4, PDR5, PDR6, PDR7
Residential Zone	Yes	R10, R11, R12, R13, R14, R15, R16, R17, R19, R1A, R1B, R2, R20, R21, R3, R6, R8, R9
Residential Apartment Zone	Yes	RA1, RA2, RA3, RA4, RA5, RA6, RA7, RA8, RA9, RA10
Residential Flat Zone	Yes	RF1, RF2, RF3

Note: "UNZONED" is also a designation DCOZ recognizes.
Source: DCOZ and ZTRAX data.

Appendix B: Imputation methods

For our generalized model, we had the following share of missing data per property (Table B.1):

Table B.1. Zoning Districts Include Similar Zones

Zones by zoning district in the Washington, DC zoning code

Variable	Share missing data	Variable	Share missing data
Zoning code	6.8%	Remodel year	62.2%
Lot area	0.2%	Build year	12.1%
Land use code	0.0%	Longitude	4.1%
Standard land use code	0.0%	Latitude	4.1%

Source: ZTRAX data.

We imputed data at the property record level as follows:

- *Lot size:* We imputed missing lot areas with the mean lot are of the zone district. For example, if a single-family residence in the generalized model did not have a lot area, we would impute that lot area using the mean of all other single-family residences.
- *Remodeling year:* We imputed remodeling year in two steps. First, for all buildings that had a build year but no remodel year, we imputed the remodel year with the build year (and assumed no major renovations had occurred). Then, we imputed the remodel year for all remaining missing homes with the mean remodel year for all other properties in the dataset with the same county land use designation (e.g., a vertical residential condo with a missing remodel year would get the average missing year of all vertical residential condos).
- *Build year:* We imputed missing build years by the median for the zone. We intentionally imputed build and remodel year slightly differently: some zones exist for historical preservation reasons, and we wanted to capture that.
- *Longitude and latitude:* Where properties had a full address, we used the census's geocoding API to identify longitude and latitude.
- *Zoning code:* As discussed in the paper, in our generalized model, we used DC's zoning map to assign properties to zones based on their latitudes and longitudes. See appendix H for information on assigning zones within our DC-specific model. For cities with up-to-date zoning information, we could use K-Nearest Neighbors to assign missing properties to zones with great accuracy (see appendix D).

Appendix C: Categorizing Single Record Density

We used county-level land use categorizations to determine the density of single property records, as shown in the table below. ZTRAX includes at minimum two types of land use categories per jurisdiction—a local categorization and a standardized categorization across the entire dataset. We used DC's county-level land use designations because of their granularity; a similar approach could be replicated across every county in the United States or standardized categories could be used.

It is also important to note that land use category was, after identifiers like Assessor's Parcel Number (APN), the most reliably complete field for the ZTRAX data in Washington, DC and two other states we examined.

Table C.1. Land Use Categories in DC Include Information about Density

Categories by density level for Washington, DC

Low density	Medium density	High density
▪ Single family row homes	▪ Horizontal condos	▪ Elevator apartments
▪ Single family semi-detached homes	▪ Horizontal condos for investment	▪ Vertical condos
▪ Single family detached homes	▪ Walk up apartments	▪ Dormitory
▪ Residence conversions with fewer than 5 units	▪ Residential conversions with more than 5 units	▪ Vertical combined condos
▪ Residential flats with fewer than 5 units	▪ Residential multifamily miscellaneous housing	▪ Vertical residential cooperatives
▪ Non-conforming single family residences	▪ Residential conversions with exactly five units	▪ Vertical mixed use cooperatives
▪ Miscellaneous single family residences	▪ Mixed use residential space	▪ Vertical condos
	▪ Transient residential space	▪ Vertical condos for investment
	▪ Multifamily non-conforming residences	

Source: ZTRAX data.

Appendix D: Using K-Nearest Neighbors to Determine Missing Zones

DC Updated their zoning code in 2016, and our dataset contained the old zoning code's designations. In order to ensure our dataset matched the current interpretation of the DC Zoning Code, we needed to match properties to their zones by latitude and longitude. We believe that this will not be necessary in other jurisdictions where ZTRAX zoning data is up-to-date. In these cases, we can use extant zoning data in combination with latitude and longitude to assign properties that are missing zones to zones using k-Nearest Neighbors.

As proof-of-concept, we undertook this process for DC. We would only be able to find zones this way for records that had latitude and longitude, including those that had been geocoded using the census address lookup geocoding API (for DC, this includes fewer than 20 records). Zones that did not have either an existing zone designation or latitude and longitude were dropped from our dataset.

We used the data that had assigned zones to train and validate our model. First, we split this data into a training and testing set. We then used the training data and calculated the 5-Nearest Neighbors using haversine distance, weighted by distance, using a leaf size of 30 in order to identify which zones were geographically closest to each target point. We validated the model using the test data. Our model was accurate 88 percent of the time in our testing set, and 97 percent of the time overall across our entire known dataset. These results support the feasibility of using k-Nearest Neighbors to impute zone designations at the property level.

Appendix E: Determining adjacent zones

Each ZTRAX's record has both a latitude and longitude. We used these latitudes and longitudes to overlay each zone with a very fine grid in order to determine the neighbors of a given zone, as follows:

1. We grouped all records within a zone by their latitudes rounded down to the fifth decimal place (about 4 feet), so if two records had latitudes of 38.908075 and 38.912459, we would represent the latitudes as 38.90807 and 38.91246 respectively.
2. We then found, per grouped latitude, the maximum and minimum associated longitudes per zone. In order for this to be effective, all points cannot be identically stacked.
3. From this reduced set of points, we found all non-self-neighbors within about 250 feet. We chose our radius by comparing our accuracy metrics across a range of potential values. Note that we used a smaller radius of about 180 feet for the DC-specific model.

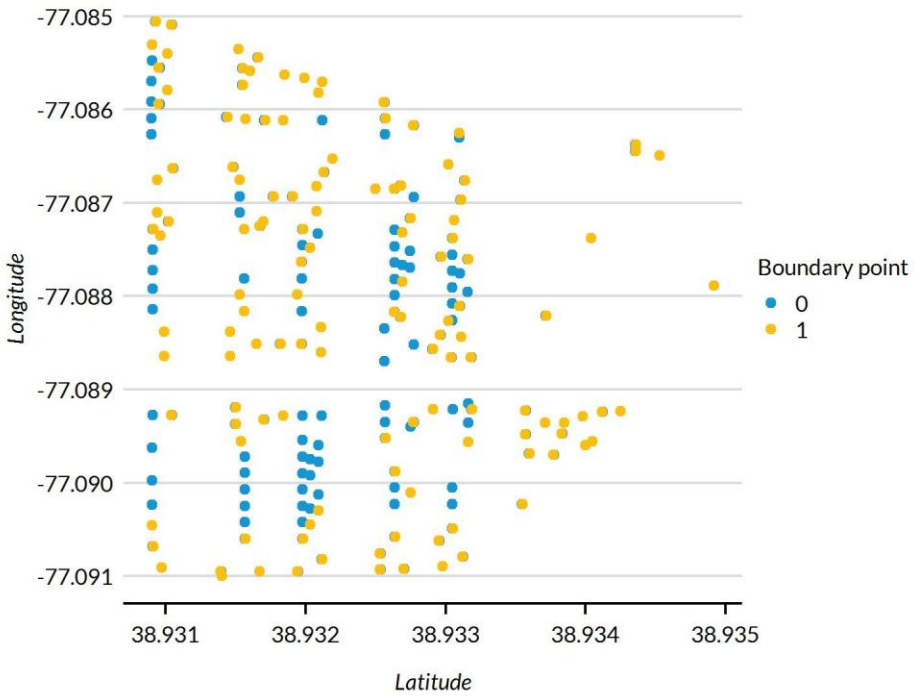
We considered our data in several dimensions. We will use an example zone, R-17, to guide our explanation. R-17's true neighbors are RA-2, MU-6, and RA-5. Our method predicted neighbors RA-2, MU-6, RA-4, MU-2, RA-5. As context, R-17 contained the most error of all residential, residential flat, and residential apartment zones (along with RF-2).

- Number of predicted neighbors: how many neighbors our method assigned to a zone. For R-17, a total of 5 neighbors were assigned to the zone.
- Number of true neighbors: how many neighbors a zone truly has. R-17 has 3 true neighbors.
- Number of same neighbors: the number of neighbors that our method assigns to a zone that are that zone's true neighbors. For R-17, three total zones that are assigned as neighbors are true neighbors: RA-2, MU-6, and RA-5.
- Number of missed neighbors: the number of neighbors that our method should have assigned to a zone, but did not. For R-17, we did not miss any zones.
- Number of incorrect neighbors: the number of zones that our method assigned as neighbors, but are not true neighbors. For R-17, there are two: RA-4 and MU-2.
- Share of all wrong neighbors: the wrong neighbors is the number of missed neighbors plus the number of incorrect neighbors as a share of the true neighbors. For R-17, the share of all wrong numbers is 2/3, or 66.67 percent

- Share of accurate neighbors: the share of the total predicted neighbors that are true neighbors. For R-17, three out of five, or 60 percent, of neighbors were accurate.

This method allowed us to reduce our set of latitude and longitudes by about half. Sample zone boundary points are shown in Figures E.1, E.2, E.3, and E.4.

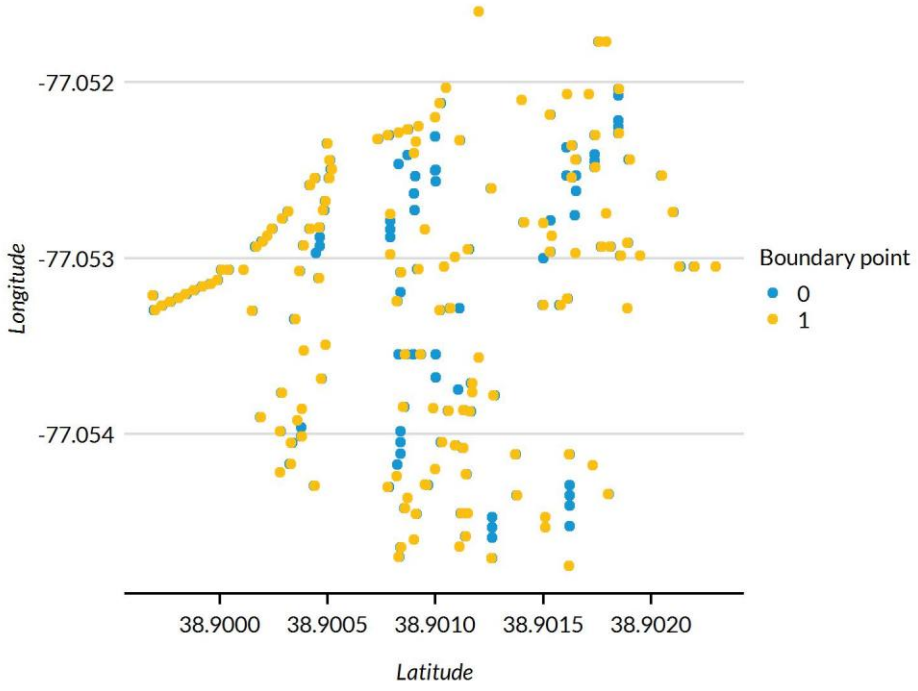
Figure E.1. R-15's Boundary Points



Source: DCOZ and ZTRAX data.

Notice that most center points of R-15 are left out of our boundary point subset. Although some points on horizontal boundaries are not included, our search radius is sufficiently large to encompass them. Our method predicts R-15's neighbors quite well: it predicts 5 of the 6 true neighbors of R-15 and does not incorrectly add any zones.

Figure E.2. R-17's Boundary Points



Source: DCOZ and ZTRAX data.

Notice that many center points of R-17 are left out of our boundary point subset, but R-17 is irregularly shaped. Our model predicts R-17's neighbors the worst of all residential zones, potentially because R-17 only has three true neighbors.

The results and accuracy of this method are summarized in Table E.1. Figures E.5 and E.6 similarly show summary information. We would expect smaller and more irregularly-shaped zones, like the Special Purpose zones, to be predicted less well. Note that these error metrics represent worst-case scenarios—the reality is, most of the cases in which our method determined incorrect errors, the errors were not significant enough to change the characteristics overall of the neighborhood metrics in our model.

Table E.1. Most Zones Retain Most True Neighbors and Inaccurately Classify About 1/3 of Total Neighbors

The share of accurate and inaccurate neighbors across all DC zones

	Share of accurate neighbors		Total share of wrong neighbors	
	Mean	**Median**	**Mean**	**Median**
All Zones	**86.62%**	**95.55%**	**31.77%**	**25.00%**
Downtown Zones	98.18%	100.00%	16.98%	11.11%
Mixed-Use Zones	85.94%	89.29%	28.93%	25.00%
Neighborhood Mixed-Use	89.16%	100.00%	20.05%	12.50%
Residential Apartment Zones	86.82%	92.06%	28.46%	26.37%
Residential Flat Zones	79.79%	88.46%	35.47%	21.57%
Residential Zones	87.44%	87.75%	29.70%	25.00%
Special Purpose Zones	78.72%	100.00%	54.48%	50.00%

Source: ZTRAX and DCOZ data.
Notes: Average weights all zones equally (that is, it is not weighted).

There are many ways we could tweak and refine these methods: we could change how latitudes are rounded, how points are grouped, and the radius used to search for neighborhoods. Yet, this preliminary method to determine neighbors for a given zone is promising and suggests that, with refinement, we could determine the neighborhoods around a zone on a larger scale.

Appendix F: Granularity on generalized model results

We created additional metrics to assess the accuracy of our model. RMSE and MAE weight each zone equally, but we believe that zones with more residential properties are more important to accurately predict. In addition to RMSE and MAE, we created metrics to take into account the relative importance of each zone. The formulas for each additional metric are below, followed by additional granularity and segmentation of the generalized model results.

Mathematical formulas

- We will define $\hat{y}$ as the value that we have predicted, y as the true value, N as the number of zones, P_z as the number of residential properties in a zone, and P as the total number of residential properties
- *RMSE:* $\sqrt{\sum_{i=1}^{N} \frac{(\hat{y}-y)^2}{N}}$
- *Weighted RMSE*: The weighted RMSE aims to assess how big of an error there is across residential properties—that is, how much error for zones by the number of residences exists. We defined weighted RMSE as: $\sqrt{\sum_{i=1}^{N} \frac{P_z*(\hat{y}-y)^2}{P}}$
- *MAE:* $\sum_{i=1}^{N} \frac{|\hat{y}-y|}{N}$
- *Relative MAE:* A relative difference between predicted and actual far of 1.0 is quite different for a high density zone with a FAR of 15 and a low density zone with a FAR of 0.8. To capture this difference, we revised the MAE formula (more specifically, the difference $\hat{y} - y$ in the MAE formula). We defined relative MAE as: $\sum_{i=1}^{N} \frac{\frac{|\hat{y}-y|}{y}}{N}$
- *Weighted relative MAE:* We further modified MAE by weighting it according to which zones had more residential properties. We defined the weighted relative MAE as: $\sum_{i=1}^{N} \frac{P_z*|\hat{y}-y|/y}{N*P}$

Additional results for the generalized model

The statistics described above are shown in Table F.1 for additional segments of the data: quantile of residential properties per zone, quartile of total number of properties per zone, and quartile of remodel year.

Table F.1. Residential Districts Have the Highest Average Number of Residences

Average number of residential properties and all properties by DC zone district for generalized model

	Traditional error metrics		Adjusted error metrics		
	RMSE	**MAE**	**Weighted RMSE**	**Relative MAE**	**Weighted Relative MAE**
Overall Model	**0.794**	**0.521**	**0.412**	**0.182**	**0.186**
Residential property count					
Fewer than 5	0.912	0.651	0.885	0.320	0.320
5 to 13.8	1.041	0.707	1.082	0.132	0.136
13.8 to 52.6	0.708	0.489	0.684	0.131	0.117
52.6 to 193	0.608	0.416	0.634	0.130	0.159
More than 193	0.597	0.342	0.403	0.198	0.187
Total property count					
Fewer than 35	0.921	0.657	0.903	0.311	0.277
35 to 108	0.495	0.358	0.402	0.106	0.114
108 to 304	0.875	0.534	0.571	0.118	0.139
More than 304	0.807	0.530	0.408	0.189	0.187
Average remodel year					
Before 1962	0.534	0.347	0.165	0.159	0.068
1962 to 1971	0.765	0.537	0.294	0.158	0.138
1971 to 1978	1.067	0.762	0.840	0.248	0.681
After 1978	0.729	0.450	0.691	0.168	0.129

Appendix G: Results for the DC-specific model

The results below follow the format of the generalized model results presented in the paper.

Table G.1. Predictions for the DC-Specific Model Are Largely Accurate

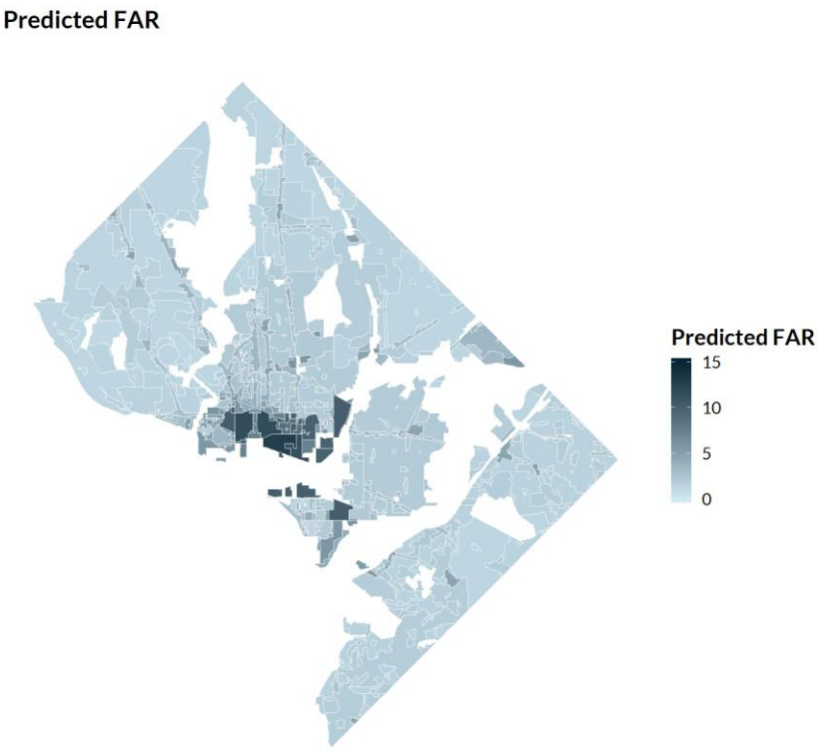
Error metrics for overall model and DC zone districts

	Traditional error metrics		Adjusted error metrics		
	RMSE	**MAE**	**Weighted RMSE**	**Relative MAE**	**Weighted Relative MAE**
Overall Model	**0.655**	**0.482**	**0.500**	**0.186**	**0.255**
Performance by Zone					
Downtown	1.240	1.088	1.113	0.117	0.111
Mixed-Use	0.576	0.464	0.305	0.130	0.096
Neighborhood Mixed-Use	0.434	0.374	0.398	0.148	0.118
Residential Apartment	0.632	0.426	1.087	0.354	0.904
Residential Flat	0.347	0.329	0.473	0.183	0.261
Residential	0.372	0.221	0.152	0.177	0.094
Special Purpose	0.760	0.613	0.617	0.261	0.120

Source: DCOZ and ZTRAX data.

Figure G.1. The DC-Specific Model Properly Identifies Areas of High Density

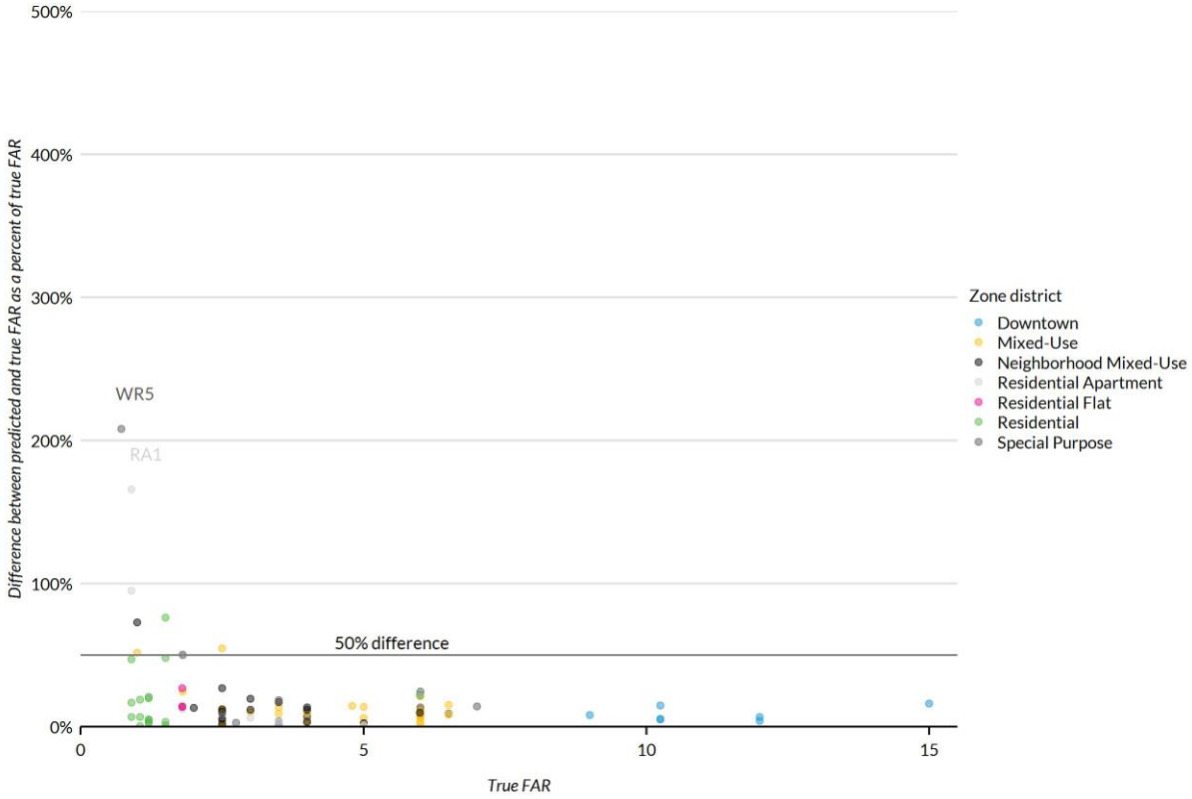
FAR by DC zone



Source: ZTRAX data, DCOZ zoning code.

Figure G.2. Most Predictions Are within 50 Percent of True FAR in the DC-Specific Model

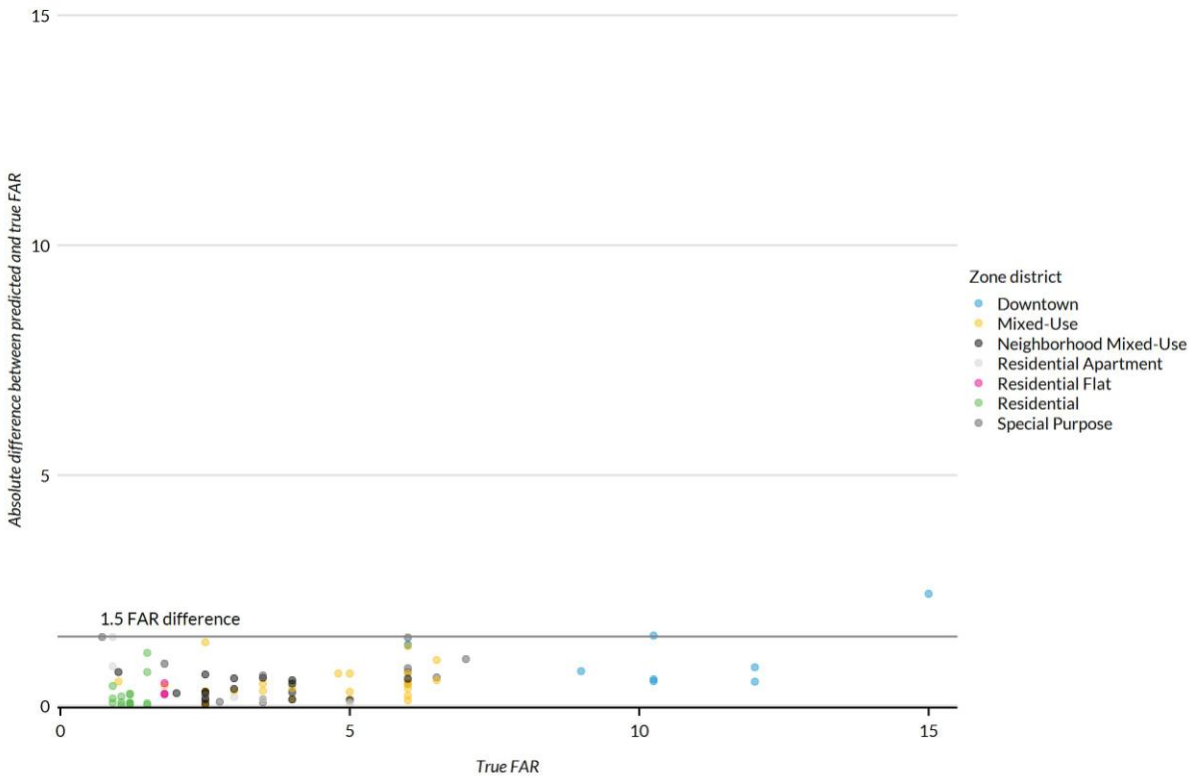
Share error for each zone's predicted FAR



Source: ZTRAX data, DCOZ zoning code.

Figure G.3. Almost Every Zone Is Predicted Within 1.5 FAR of True FAR

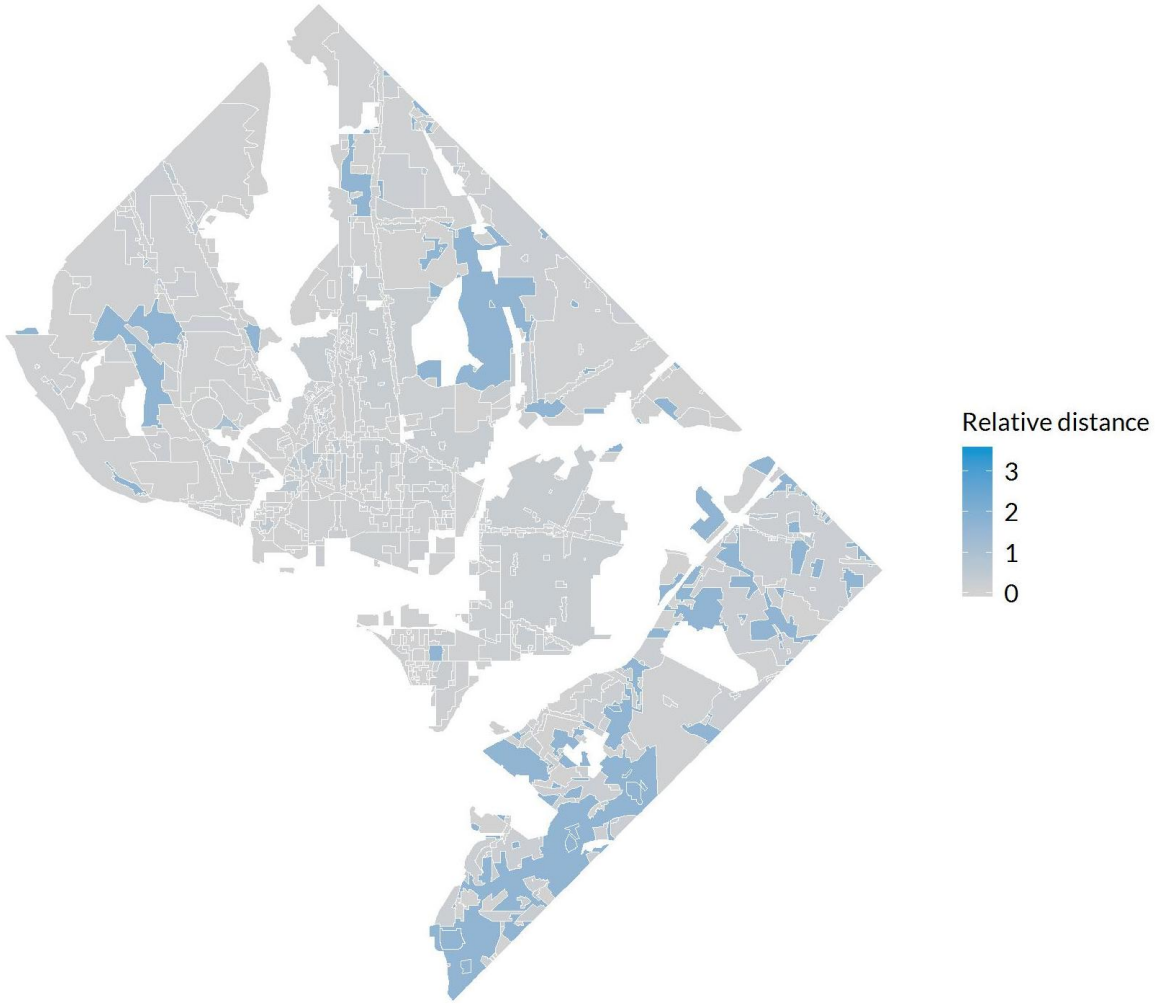
Absolute difference between predicted and true FAR



Source: ZTRAX data, DCOZ zoning code.

Figure G.4. The DC-Specific Model Has Low Relative Error with a Few Exceptions

Relative Distance from Predicted to True FAR



Source: ZTRAX data, DCOZ zoning code.

Appendix H: Additional information on zone assignment for the DC-specific model

For our DC-specific model, we used domain knowledge about lot numbering and lot boundaries in order to best assign properties to zones. At a high level, we identified where each lot was using a map of lots, and then matched each lot's centroid to the DCOZ zoning map. More specifically, we:

1. We split all records by lot number into tax lots, condominium lots, and record lots, as described in the body of the paper.
2. Found the geometry of each lot using DCOZ data and took the centroid of that lot. In cases where we could not find the geometry, we used the US Census geocoding API to find a latitude and longitude for the address provided, or ZTRAX's provided latitude and longitude.
3. We used these coordinates to merge our property records with DCOZ's zoning map and determine the current zone per property.
4. We supplemented our final data in two ways:
    a. Properties with no latitude and longitude but a ZTRAX assigned zone were assigned to the closest 2016 update version of that ZTRAX zone.

Residential properties that were assigned by our method to industrial zones were reassigned to the closest 2016 update of their ZTRAX zone.

References

Blumenstock, J., G. Cadamuro, and R. 2015. “Predicting Poverty and Wealth from Mobile Phone Metadata.” *Science*. 350(6264): 1073-6.

Furman, Jason. 2015. “Barriers to Shared Growth: The Case of Land Use Regulation and Economic Rents.” Remarks at the Urban Institute.

Ganong, Peter and Daniel Shoag. 2017. “Why Has Regional Income Convergence in the U.S. Declined?” *Journal of Urban Economics*. 102:76-90.

Glaeser, Edward and Joseph Gyourko. 2018. “The Economic Implications of Housing Supply.” *Journal of Economic Perspectives* 32(1): 3-30.

Glaeser, Edward, L. and Bryce A. Ward. 2006. “Myths and Realities of American Political Geography.” *Journal of Economic Perspectives*, 20(2): 119-144.

Greene, Solomon, and Kathryn L.S. Pettit. 2016. “What if Cities Used Data to Drive Inclusive Neighborhood Change?” Washington, DC: Urban Institute.

Greene, Solomon, Margery Turner, and Ruth Gourevitch. 2017. *Racial Residential Segregation and Neighborhood Disparities*. Washington, DC: Urban Institute.

Gyourko, J., A. Saiz, and A. Summers. 2007. “A new measure of the local regulatory environment for housing markets: The Wharton Residential Land Use Regulatory Index.” *Urban Studies* 45(3): 693-729.

Hsieh, Chang-Tai and Enrico Moretti. 2015. “Why Do Cities Matter? Local Growth and Aggregate Growth.” NBER Working Paper No. 21154.

Ikeda, Sanford and Emily Washington. 2015. *How Land-Use Regulation Undermines Affordable Housing.* Arlington, VA: Mercatus Research, Mercatus Center at George Mason University.

Micklow, A. C., Warner, M. E. 2014. “Not Your Mother’s Suburb: Remaking Communities for a More Diverse Population.” *The Urban Lawyer* 46 (4): 729–51.

Nolon, John R. 2013. “Shifting Paradigms Transform Environmental and Land Use Law: The Emergence of the Law of Sustainable Development.” *Fordham Environmental Law Journal*.

Pendall, Rolf. 2000. “Local Land Use Regulation and the Chain of Exclusion.” *Journal of the American Planning Association.* 66(2): 125-142.

Pendall, Rolf, Robert Puentes and Jonathan Martin. 2006. *From Traditional to Reformed: A Review of the Land Use Regulations in the Nation’s 50 Largest Metropolitan Areas.* Washington, DC: Brookings Institution Metropolitan Policy Program.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age.* Princeton, NJ: Princeton University Press. Open review edition.

Schleicher, David. 2017. “Stuck! The Law and Economics of Residential Stability.” *Yale Law Journal*, Vol. 127.

Tach, Laura, Rolf Pendall, and Alexandra Derian. 2014. *Income Mixing across Scales: Rationale, Trends, Policies, Practice, and Research for More Inclusive Neighborhoods and Metropolitan Areas.* Washington, DC: Urban Institute.