URBAN INSTITUTE · ELEVATE · THE · DEBATE

# Predicting Zoned Density Using Property Records: Next Steps

## Technical Appendix

*Erika Tyagi and Graham MacDonald*

This appendix documents the technical steps and results that support our analysis using property assessment data to predict zoned densities in Washington, DC; Montgomery County, Maryland; and Arlington County, Virginia.[1]

## Data Sources

### Zillow Transaction and Assessment Dataset[2]

We relied on the Zillow Transaction and Assessment Dataset (ZTRAX), a national dataset of property assessments sourced from local county assessors' offices. We used the following ZTRAX fields for each property:

- Lot size: the property's lot size expressed in square feet
- Standard land-use code: the ZTRAX standardized land- and building-use code converted from local county codes and descriptions (e.g., RR101 refers to "single-family residential" across the entire dataset)
- Year built: the year the property was built
- Year remodeled: the year the property was remodeled
- Geographic descriptors: the property's street address, city, state, zip code, FIPS code, and longitude and latitude

## Local Zoning Ordinances

For Washington, DC, Montgomery County, and Arlington County, we manually read and interpreted the relevant zoning ordinances to produce tabular datasets where each row represents a zone (i.e., a unique combination of a zoning jurisdiction and a zoning code), and each column contains the relevant zoning restrictions for that zone.[3] The goal of our model is to predict the maximum floor area ratio (FAR) for the average residential property in each zone. In order to derive this single permissible FAR per zone, we use the following assumptions:

- We record by-right regulations and ignore conditional use limits that require special approval by a local planning commission.
- We record the base density and ignore bonuses permitted through optional development methods.[4]
- For zones with different regulations by residential building subtype, we calculate the simple average of the regulations across all subtypes (e.g., if a zone allows townhouses with a FAR maximum of 2.0 and condos with a FAR maximum of 4.0, the FAR limit for the entire zone would be 3.0).
- For zones where a maximum FAR limit is not specified but the maximum height limit and lot coverage are specified, the FAR limit is calculated as the product of the maximum-allowed stories and the maximum-allowed coverage (e.g., if a zone allows for buildings up to four stories and coverage up to 50 percent, its FAR would be 2.0).[5]

# Methodology

To create our modeling datasets, we use June 2019 ZTRAX data limited to records within the relevant areas (i.e., Washington, DC; Montgomery County, Maryland; and Arlington County, Virginia). At a high level, we aggregate this data so each row represents summary information about all of the properties in a single zoning designation. We then build a model that uses these aggregated zone-level characteristics to predict our transcribed FAR density limits.

### STEP 1: AGGREGATING PROPERTY-LEVEL DATA

We first aggregate ZTRAX data at the property level by assuming that all records with the same full street address belong to the same property. This attributes individual units (e.g., condos, offices, etc.)

within a single building to that building's record. We aggregate multiple ZTRAX records into a single property record as follows:

- Standard land-use code: We use the most common value (e.g., if 10 records are associated with a given building, 9 of which are residential and 1 of which is industrial, we assign the building as residential). We use a stable sorting algorithm to break ties.
- Longitude and latitude: We use the median value (e.g., if four records all pertain to the same physical building, our calculated latitude and longitude would represent the approximate center of those records).
- Year built and year remodeled: We use the most recent possible year (e.g., if a condominium had five lots, the most recent one to be remodeled would provide the remodel year for the building).
- Lot size: We calculate the sum of lot sizes across each individual record (e.g., the lot size of a condominium building would be the sum of each unit).

### STEP 2: IMPUTING MISSING DATA

We next imputed any missing data using the following methods:

- Lot size: We impute missing lot size with the mean value within a standard land-use code (e.g., if a single-family residential property did not have a lot area, we would impute that data using the mean of all other single-family residences).
- Year remodeled: For buildings with a year built but no year remodeled, we imputed the year remodeled with the year built (and assumed no major renovations had occurred). For all remaining missing values, we then imputed the year remodeled with the mean year remodeled within a standard land-use code (e.g., a single-family residential property without remodeling data would be assigned the average of all single-family residential properties).
- Year built: We impute a missing build year as the median for the zone code designation.[6]
- Latitude and longitude: We geocoded properties with full addresses but without longitudes and latitudes.[7]

We then assign properties to zones by geospatially matching each property's longitude and latitude to zoning boundary maps for the relevant jurisdictions.[8] We also use these geospatial zoning boundary maps to identify neighboring zones (i.e., zones with adjacent boundaries).

### STEP 3: AGGREGATING ZONE-LEVEL DATA

Our final step in creating our modeling dataset is aggregating our property-level data to create characteristics at the zone level. We create the following features for each zone:

- 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 share of residential lot area occupied by high-, medium-, and low-density homes in the zone[9]
- the total residential lot area (or living space) in the zone
- the average lot area per residential property in the zone
- the number of residential and total properties within the zone
- the most common property land-use type in the zone[10]
- the average share of residential lot area occupied by high-, medium-, and low-density homes across all unique neighboring zones[11]
- the number of unique neighboring zones in total and that were categorized as residential[12]
- the share of total neighboring zones that were categorized as residential

We then merge this zone-level aggregated data with our datasets containing the permitted density limits from the transcribed zoning ordinances.[13]

TABLE 1

**Properties and Zones in Modeling Datasets**

| | Washington, DC | Montgomery County, Maryland | Arlington County, Virginia |
|---|---|---|---|
| Number of zones | 94 | 83 | 31 |
| Number of residential zones | 47 | 69 | 21 |
| | | | |
| Number of properties | 131,392 | 251,012 | 45,362 |
| Number of residential properties | 118,828 | 226,321 | 41,861 |

**Source:** ZTRAX and local zoning ordinance data. 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.

**Note:** Residential zones are those where the most common property land-use type is either residential or residential income (multifamily).

### STEP 4: MODELING

We first use a lasso regression to select a subset of the features generated for modeling. The lasso regression penalizes features that have minimal impact on the outcome variable, and we subsequently remove the most penalized features to further reduce the total number of features used in the model to prevent overfitting. We then run a random forest regression with the selected features to predict the FAR maximum for each zone.[14] When testing in-sample, we use leave-one-out cross validation (LOOCV), training one model for each zone in the dataset, each with its own LASSO and random forest regression. In the DC modeling dataset, for example, a LOOCV approach would mean that we train a model on 93 of the 94 zones, then test that model on the "left-out" zone, and then repeat this process for each of the 94 zones. When testing out-of-sample, we separate our training and testing datasets as specified in the results section below.

### STEP 5: EVALUATION

To evaluate the accuracy of our models, we consider five metrics. The first two are traditional prediction evaluation metrics that weight each zone equally: root mean squared error (RMSE) and mean absolute error (MAE). We also consider a weighted RMSE where we weight zones based on their number of residential properties (i.e., error in zones with more residential properties are weighted more heavily).

We consider relative deviation in predicted and actual FAR in addition to absolute deviation. Specifically, we note that predicting a FAR of 2.0 for a zone whose true FAR is 1.0 should be weighted more heavily than predicting a FAR of 11.0 for a zone whose true FAR is 10.0. Although the former might change a zoning classification from low to moderate density, the latter would likely still be labeled high density in either case. Using our relative MAE, rather than the traditional MAE, the former low-density zone would have a relative MAE of 100 percent, while the latter would have a relative MAE of 10 percent. We finally consider a weighted relative MAE that both weights by the number of residential properties in a zone and considers relative deviation. Mathematical formulas for the five metrics are specified as follows:

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: $\sqrt{\sum_{i=1}^{N} \frac{P_z*(\hat{y}-y)^2}{P}}$
- MAE: $\sum_{i=1}^{N} \frac{|\hat{y}-y|}{N}$
- relative MAE: $\sum_{i=1}^{N} \frac{|\hat{y}-y|/y}{N}$
- weighted relative MAE: $\sum_{i=1}^{N} \frac{P_z*|\hat{y}-y|/y}{N*P}$

# Results

We present our results for four model specifications. Two of these models use an in-sample LOOCV approach, as described in the previous section:

- predicting FAR for zones in Washington, DC
- predicting FAR for zones in Washington, DC, Montgomery County, and Arlington County

Two use an out-of-sample approach, training on zones in one set of jurisdictions and testing on another:

- predicting FAR for zones in Montgomery County, training on zones in DC and Arlington County
- predicting FAR for zones in Arlington County, training on zones in DC and Montgomery County

## Evaluation Metrics

We first present an overview of the evaluation metrics discussed in the previous section. The in-sample models consistently have lower error rates than the out-of-sample models—although this difference is reduced when limiting our analysis to just residential zones or just zones with a large number of residential properties. We present the metrics for these subsets of zones across the four models.

TABLE 2

**Error Metrics by Model**

| | In Sample | | Out of Sample | |
|---|---|---|---|---|
| | **DC** | **DC and Montgomery and Arlington Counties** | **Montgomery County** | **Arlington County** |
| **All zones** | | | | |
| *Traditional error metrics* | | | | |
| RMSE | 0.64 | 0.58 | 1.04 | 1.17 |
| MAE | 0.46 | 0.38 | 0.61 | 0.97 |
| *Adjusted error metrics* | | | | |
| Weighted RMSE | 0.28 | 0.25 | 0.33 | 0.77 |
| Relative MAE | 0.15 | 0.21 | 0.51 | 0.73 |
| Weighted relative MAE | 0.11 | 0.12 | 0.14 | 0.46 |
| **Residential zones** | | | | |
| *Traditional error metrics* | | | | |
| RMSE | 0.61 | 0.45 | 0.87 | 0.83 |
| MAE | 0.41 | 0.30 | 0.52 | 0.69 |
| *Adjusted error metrics* | | | | |
| Weighted RMSE | 0.27 | 0.24 | 0.33 | 0.77 |
| Relative MAE | 0.18 | 0.20 | 0.45 | 0.38 |
| Weighted relative MAE | 0.11 | 0.12 | 0.14 | 0.45 |
| **Zones with 100+ residences** | | | | |
| *Traditional error metrics* | | | | |
| RMSE | 0.41 | 0.36 | 0.61 | 0.76 |
| MAE | 0.26 | 0.23 | 0.41 | 0.61 |
| *Adjusted error metrics* | | | | |
| Weighted RMSE | 0.27 | 0.24 | 0.32 | 0.76 |
| Relative MAE | 0.15 | 0.19 | 0.36 | 0.39 |
| Weighted relative MAE | 0.11 | 0.12 | 0.14 | 0.45 |

**Sources:** ZTRAX and local zoning ordinance data. 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.

**Notes:** MAE = mean absolute error; RMSE = root mean squared error. In the in-sample models, a leave-one-out cross validation approach is used for evaluation. In the Montgomery County out-of-sample model, we train on zones in DC and Arlington County. In the Arlington County out-of-sample model, we train on zones in DC and Montgomery County. Residential zones are those where the most common property land-use type is either residential or residential income (multifamily).

## Feature Importance

Next, we summarize the feature importance for each of the four models. Four features consistently have the highest predictive power across model specifications:

- The average share of residential lot area occupied by low-density and medium-density homes across all unique neighboring zones

- The share of total neighboring zones that were categorized as residential
- The mean build year for all properties in the zone

TABLE 3

**Feature Importance by Model**

| | In Sample | | Out of Sample | |
|---|---|---|---|---|
| | DC | DC and Montgomery and Arlington Counties | Montgomery County | Arlington County |
| Low-density neighbors | 0.32 | 0.24 | 0.08 | 0.42 |
| Share of residential neighbors | 0.25 | 0.32 | 0.41 | 0.27 |
| Medium-density neighbors | 0.14 | 0.22 | 0.24 | 0.12 |
| Mean build year | 0.17 | 0.10 | 0.15 | 0.08 |
| Medium-density share of homes | 0.00 | 0.06 | 0.07 | 0.05 |
| Number of residential neighbors | 0.02 | 0.03 | 0.04 | 0.02 |
| Residential living space | 0.03 | 0.00 | 0.00 | 0.00 |
| Minimum build year | 0.03 | 0.00 | 0.00 | 0.00 |
| Commercial office zone | 0.01 | 0.01 | 0.00 | 0.00 |
| Commercial retail zone | 0.00 | 0.00 | 0.00 | 0.01 |
| Vacant land zone | 0.01 | 0.01 | 0.00 | 0.01 |

**Sources:** ZTRAX and local zoning ordinance data. 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.

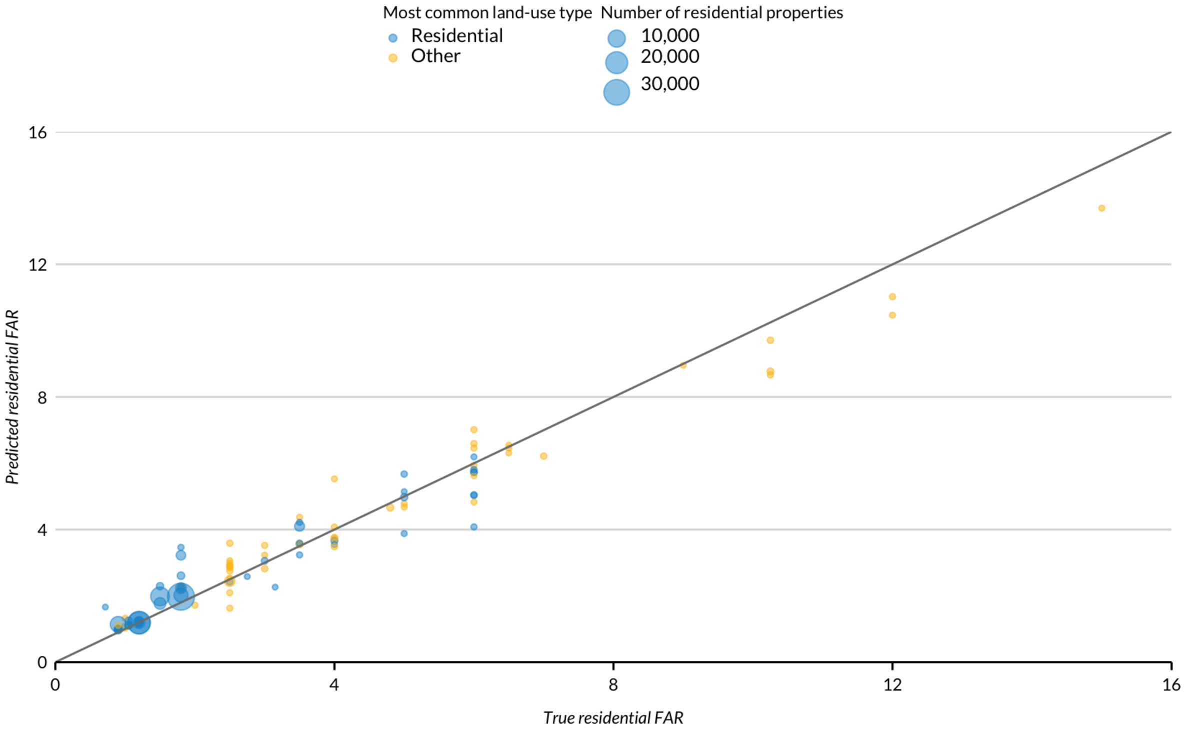
**Note**: In the in-sample models, a leave-one-out cross validation (LOOCV) approach is used for evaluation. In the Montgomery County out-of-sample model, we train on zones in DC and Arlington County and predict all zones in Montgomery County. In the Arlington County out-of-sample model, we train on zones in DC and Montgomery County and predict all zones in Arlington County. Feature importance reflects the simple average across all trained models in the LOOCV process. A feature not included after the LASSO feature selection step in building a single model is given an importance of 0.0. Features without an importance larger than 1.0% across any of the four models are not shown.

## Comparing Predicted and True Density

Finally, we present figures comparing our predicted densities with the actual FAR limits as specified in the zoning ordinances for each of the four models. The total number of residential properties and the most common type of property in each zone are reflected by the size and color of each point, respectively.

FIGURE 1

**Predicted and True FAR: In-Sample DC Model**



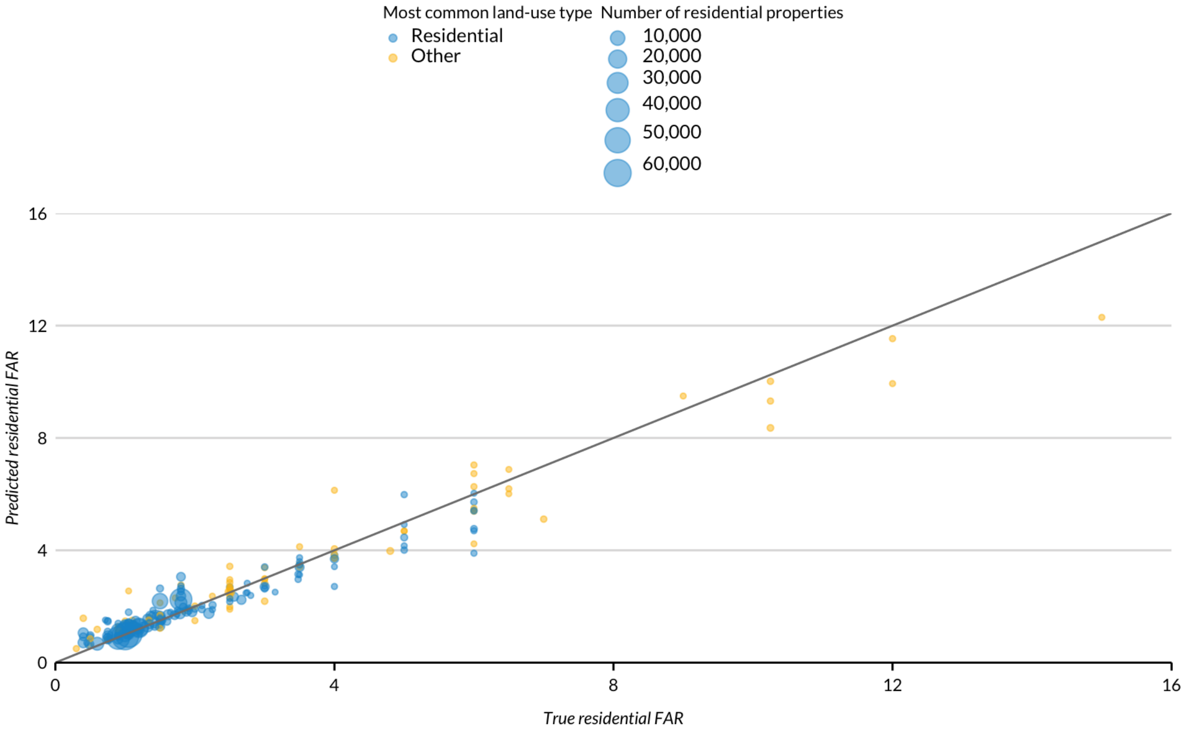
URBAN INSTITUTE

**Sources:** ZTRAX and local zoning ordinance data. 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.

**Note:** FAR = floor area ratio.

FIGURE 2

## Predicted and True FAR: In-Sample DC, Montgomery County, and Arlington County Model

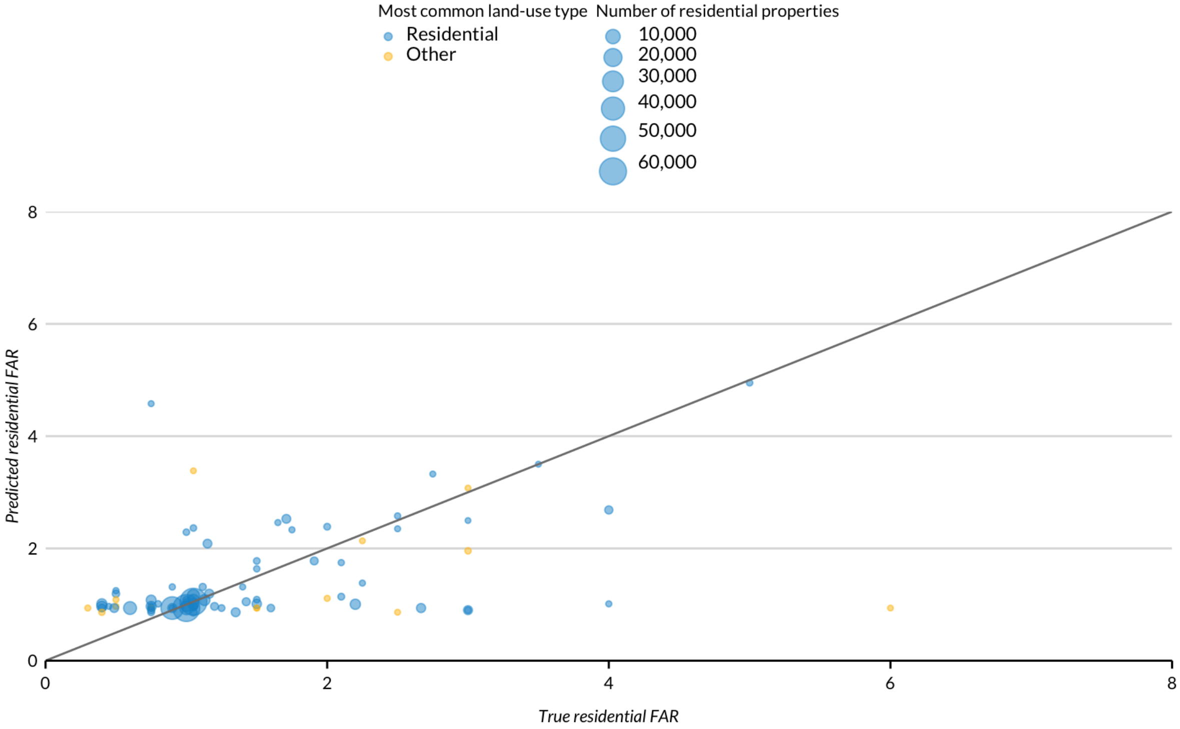


**Sources:** ZTRAX and local zoning ordinance data. Data were provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the authors and do not reflect the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

**Note:** FAR = floor area ratio.

FIGURE 3

### Predicted and True FAR: Out-of-Sample Montgomery County



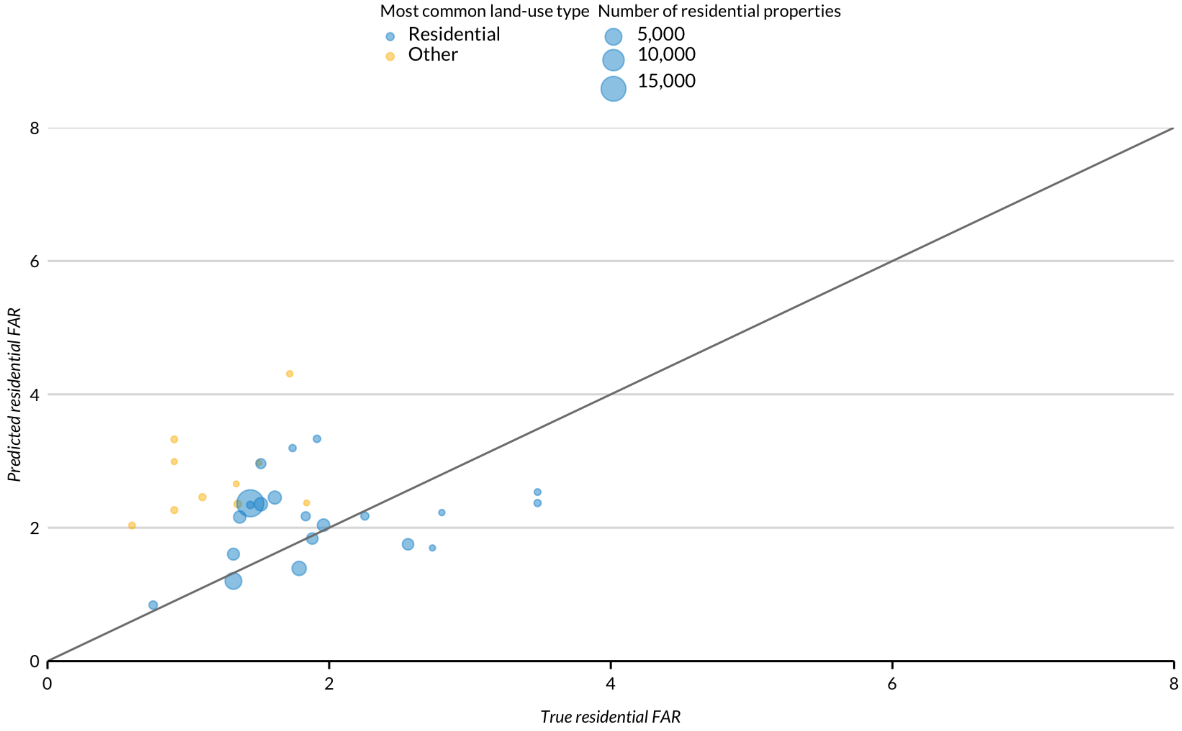
URBAN INSTITUTE

**Sources:** ZTRAX and local zoning ordinance data. 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.

**Note:** FAR = floor area ratio.

FIGURE 4

**Predicted and True FAR: Out-of-Sample Arlington County**



URBAN INSTITUTE

**Sources:** ZTRAX and local zoning ordinance data. 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.
**Note:** FAR = floor area ratio.

# Discussion

Our approach suffers from several limitations. First, our models are built on an extremely small number of zones—207 at most in the broadest in-sample model. This is inherently capped by the number of distinct zones within the three jurisdictions, and this number is further reduced as we are forced to exclude zones where the zoning ordinances do not provide sufficient information to compute a "true" maximum-allowed FAR. Additionally, many zones contain few properties and even fewer residential properties. In Washington, DC, 46 of our 94 zones contain fewer than 20 residential properties. In Montgomery and Arlington Counties respectively, 7 of 24 and 24 of 83 zones contain fewer than 20 residential properties. Though our predictions are remarkably accurate for such a small sample size, they remain highly sensitive to the underlying accuracy of the property assessment data and our imputation process.

Our process for calculating a "true" FAR also merits further consideration. As previously noted, we choose to compute the maximum-allowed by-right floor area ratio for the average residential property in each zone. In jurisdictions where optional and conditional development methods (e.g., optional Moderately Priced Dwelling Unit (MPDU)development in Montgomery County and special exception development in Arlington County) are prevalent, however, choosing to predict by-right FAR limits may not be as meaningful. Additionally, our decision to use a simple average of density limits across subtypes to compute a single maximum-allowed FAR for each zone is imprecise. For example, if a zone allows townhouses with a FAR maximum of 2.0 and condos with a maximum of 4.0, the FAR limit would be 3.0—even if the vast majority of properties in that zone are actually townhouses, and the technical maximum FAR limit for the zone is 4.0.

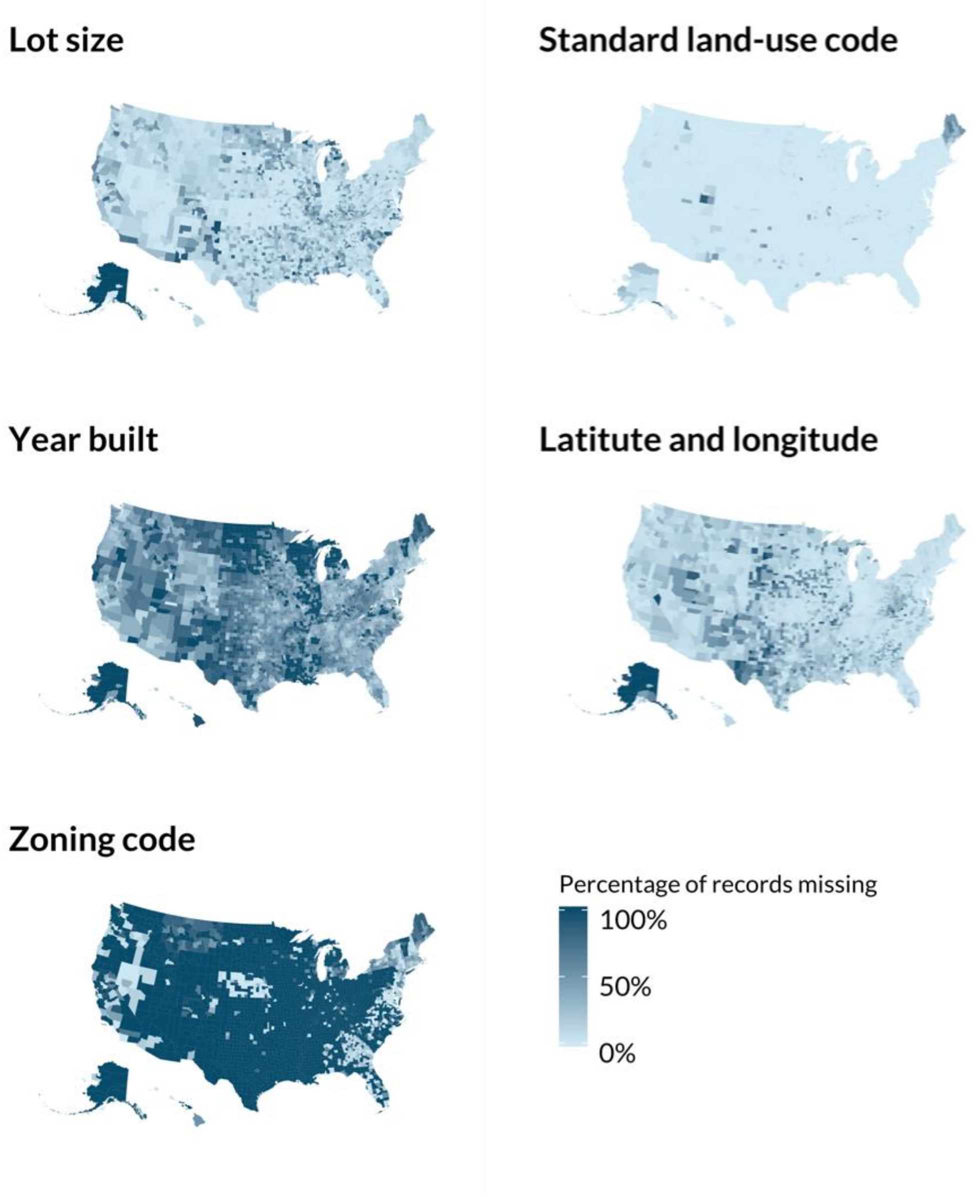
By expanding this work and refining our approach, we believe this methodology could be broadly applied to both new jurisdictions and additional zoning restrictions (e.g., lot setbacks or parking restrictions). To begin to understand the feasibility of expanding to additional jurisdictions, we explored the availability (i.e., nonmissingness) of ZTRAX data nationally for the five foundational fields used in our approach:

- Lot size: the property's lot size expressed in square feet
- Standard land-use code: the ZTRAX standardized land- and building-use code
- Year built: the year the property was built
- Latitude and longitude: the property's coordinates
- Zoning code: the property's local zoning code (e.g., R-2)

As noted, ZTRAX data are typically sourced from local county assessors' offices. Thus, we compute the percentage of ZTRAX records with missing data for each of these five fields at the county level.

FIGURE 5

**Percentages of ZTRAX Records with Missing Data Nationally**



URBAN INSTITUTE

**Source:** ZTRAX data. 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.
**Note**: Percentages are aggregated at the county level.

We also summarize the share of missing data across these five fields for the 20 most populous metropolitan areas.[15]

TABLE 4

**Percentage of ZTRAX Records Missing Data for the 20 Most Populous Metropolitan Statistical Areas**

| | Lot size (%) | Standard land-use code (%) | Year built (%) | Latitude and longitude (%) | Zoning code (%) |
|---|---|---|---|---|---|
| San Diego-Carlsbad, CA | 67 | 0 | 17 | 10 | 2 |
| Washington-Arlington-Alexandria, DC-VA-MD-WV | 11 | 0 | 14 | 3 | 21 |
| Phoenix-Mesa-Scottsdale, AZ | 3 | 0 | 16 | 6 | 22 |
| Philadelphia-Camden-Wilmington, PA-NJ-DE-MD | 7 | 0 | 13 | 2 | 25 |
| Los Angeles-Long Beach-Anaheim, CA | 12 | 0 | 12 | 4 | 29 |
| Boston-Cambridge-Newton, MA-NH | 14 | 0 | 13 | 1 | 38 |
| Atlanta-Sandy Springs-Roswell, GA | 23 | 0 | 14 | 4 | 38 |
| Tampa-St. Petersburg-Clearwater, FL | 19 | 0 | 14 | 2 | 42 |
| New York-Newark-Jersey City, NY-NJ-PA | 16 | 0 | 20 | 2 | 42 |
| Seattle-Tacoma-Bellevue, WA | 6 | 0 | 13 | 4 | 48 |
| Miami-Fort Lauderdale-West Palm Beach, FL | 28 | 0 | 7 | 2 | 49 |
| St. Louis, MO-IL | 31 | 1 | 27 | 5 | 58 |
| Detroit-Warren-Dearborn, MI | 18 | 2 | 24 | 1 | 64 |
| Denver-Aurora-Lakewood, CO | 8 | 0 | 16 | 4 | 69 |
| Riverside-San Bernardino-Ontario, CA | 10 | 0 | 33 | 21 | 83 |
| San Francisco-Oakland-Hayward, CA | 16 | 0 | 12 | 4 | 87 |
| Minneapolis-St. Paul-Bloomington, MN-WI | 22 | 1 | 20 | 2 | 90 |
| Dallas-Fort Worth-Arlington, TX | 14 | 0 | 18 | 4 | 94 |
| Chicago-Naperville-Elgin, IL-IN-WI | 17 | 0 | 24 | 1 | 100 |
| Houston-The Woodlands-Sugar Land, TX | 14 | 0 | 25 | 8 | 100 |

**Source:** ZTRAX data. 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.

As the table demonstrates, property assessment records lack a zoning code across most counties nationally—although this field is fairly complete for several large metropolitan statistical areas, such as DC, Los Angeles, Philadelphia, and San Diego. In jurisdictions where these data are reasonably complete, we could use a nearest-neighbors approach to impute a property's zoning code based on the zoning code of its nearest properties with nonmissing codes.[16] We also found that we can leverage previous releases of ZTRAX data to impute missing data in the current release, which could reduce the share of properties missing a zoning code in the table above. In Arlington County and Washington, DC, for example, previous releases of ZTRAX data included zoning codes for most properties, although these data were entirely missing in the June 2019 version.[17] Finally, geospatial files delineating zoning boundaries are often made publicly available by municipalities. These files could be used to assign properties to their zone and identify neighboring zones to fill these gaps. For example, across the central Washington, DC, metropolitan statistical area (including five cities and nine counties), we found that these geospatial files were publicly available for all but two localities.[18]

Where the necessary data are available, the approach described here should be fully generalizable across jurisdictions in the US. Moreover, we think it is very likely that the accuracy of our model will improve as additional jurisdictions are incorporated. Supplementing property assessment data with additional sources—such as satellite imagery on building footprints and heights—also has the potential to improve the robustness of our predictions. By continuing to expand and develop this approach, we hope that machine-learning methods can unlock comparable zoning data across several regions and perhaps one day create a national database of predicted zoning restrictions.

# Notes

[1] The methodology and underlying code were largely adapted from earlier work done by Emma Nechamkin and Graham MacDonald, available at https://www.urban.org/research/publication/predicting-zoned-density-using-property-records.

[2] 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.

[3] In Montgomery County, the following municipalities do not fall under the county's zoning ordinance: Barnesville, Brooksville, Gaithersburg, Laytonsville, Poolesville, Rockville, and Washington Grove. We read the zoning ordinances for Gaithersburg and Rockville separately, but we excluded the other jurisdictions from our analysis.

[4] Density bonuses allowed through Washington, DC's Inclusionary Zoning program, Montgomery County's Moderately Priced Dwelling Unit and Cluster Development programs, and Arlington County's Special Exception allowances are not included.

[5] For zones where a maximum height is specified in feet, we convert this measurement to stories using the following formula: $height_{stories} = \ ceiling[\frac{(height_{feet}-10)}{10}]$.

[6] We intentionally imputed build year differently from remodel year. Some zones exist for historical preservation reasons, and we wanted to capture that.

[7] This was done using the Census Geocoder API: https://geocoding.geo.census.gov/.

[8] We follow this approach rather than rely on the property zoning description field in the ZTRAX data because this field is entirely missing for DC and Arlington County. Note that the property zoning description field contained data in previous ZTRAX releases that we could have used, but we used actual zoning data to ensure we produce accurate model test results of the ability to generalize for this project. Zoning boundary geospatial data were available from the following jurisdictions:

Arlington County: https://gisdata-arlgis.opendata.arcgis.com/datasets/zoning-polygons.
Gaithersburg:https://gaithersburgmd.maps.arcgis.com/apps/webappviewer/index.html?id=e03b0687ef6c401381efb0355915a4b4;
Montgomery County: https://montgomeryplanning.org/tools/gis-and-mapping/data-downloads/;

Rockville: https://data-rockvillemd.opendata.arcgis.com/datasets/rockville-zoning-districts?geometry=-77.343%2C39.04%2C-76.852%2C39.133;
Washington, DC: https://opendata.dc.gov/datasets/zoning-regulations-of-2016.

[9] Properties are assigned to a density level (high, medium, or low) based on their standard land-use code description. The following types of properties are categorized as high-density: high-rise apartment, apartment (generic); the following types are categorized as medium-density: condominium, residential income general (multifamily), residential general, multifamily dwelling (generic any combination 2+), cooperative, garden apartment, court apartment (5+ units); the following types are categorized as low-density: single-family residential, townhouse, rural residence, inferred single-family residential, row house, duplex (2 units, any combination).

[10] This is based on the least granular aggregation of the standard land-use code and includes the following categories: residential, residential income (multifamily), commercial office, commercial retail, governmental, industrial, exempt and institutional, and vacant land.

[11] This is based on a simple average (i.e., each neighboring zone is weighted equally).

[12] Residential zones are those where the most common property land-use type is either residential or residential income (multifamily).

[13] We exclude zones where a by-right maximum-allowed FAR cannot be calculated (e.g., zones where the density limits are established at the time of development approval, rather than delineated in the ordinance). We also exclude zones with no residential properties in the ZTRAX data.

[14] We use a lasso optimization tolerance of 0.1 when removing features. For each random forest regression model, we determine the number of trees and maximum depth of each tree to minimize root mean squared error. Other model hyperparameters are not tuned and reflect sklearn.ensemble.RandomForestRegressor default values.

[15] These areas reflect the 20 metropolitan statistical areas with the largest populations as of July 1, 2018, based on the Census Bureau's annual estimates: https://factfinder.census.gov/bkmk/table/1.0/en/PEP/2018/PEPANNRES/0100000US.31000.

[16] This approach is described in detail in the first publication of this methodology, available at: https://www.urban.org/research/publication/predicting-zoned-density-using-property-records.

[17] ZTRAX property assessment data contain a unique identifier that can be used to match parcels in the current assessor data to previous assessor parcel records.

[18] The central Washington, DC, metropolitan area is defined by the US Census Bureau using the approach described here: https://www2.census.gov/geo/pdfs/reference/GARM/Ch13GARM.pdf. This includes the following entities: Alexandria City, Arlington County, the District of Columbia, Fairfax City, Fairfax County, Falls Church City, Fauquier County, Loudon County, Manassas City, Manassas Park City, Montgomery County, Prince George's County, Prince William County, and Stafford County.

## ABOUT THE URBAN INSTITUTE

URBAN INSTITUTE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.