Using Machine Learning to Generate New, Valuable Zoning Data

Graham MacDonald, Chief Data Scientist
What is zoning?
What is zoning?

“From the start, zoning has separated more than just land uses. It also separates people.”
There are no (comprehensive, comparable, and current) zoning data?
There are no (comprehensive, comparable, and current) zoning data?

306 REAR YARD

306.1 A minimum rear yard of twenty-five feet (25 ft.) shall be provided in the R-1-A and R-1-B zones.

306.2 A minimum rear yard of twenty feet (20 ft.) shall be provided in the R-2 and R-3 zones.

306.3 Notwithstanding Subtitle D §§ 306.1 and 306.2, a rear wall of an attached or semi-detached building shall not be constructed to extend farther than ten feet (10 ft.) beyond the farthest rear wall of any adjoining principal residential building on an adjoining property.

306.4 A rear wall of an attached or semi-detached building may be constructed to extend farther than ten feet (10 ft.) beyond the farthest rear wall of any adjoining principal residential building on an adjoining property if approved as a special exception pursuant to Subtitle X, Chapter 9 and as evaluated against the criteria of Subtitle D §§ 5201.3(a) through 5201.3(d) and §§ 5201.4 through 5201.6.

How can data science help?
How can data science help?

1. Use natural language processing to extract zoning rules directly from local zoning codes.

2. Use machine learning to predict zoning rules based on property assessment data.
Property assessment data?

*Use machine learning to predict zoning rules based on property assessment data*
Property assessment data?

*Use machine learning to predict zoning rules based on property assessment data*

- Lot size
- Year built/remodeled
- Land use description
- Geographic information
Zoning rules?

Use machine learning to predict *zoning rules* based on
*property assessment data*
Zoning rules?

*Use machine learning to predict *zoning rules* based on property assessment data*

Maximum allowed by-right floor area ratio (FAR)

\[ \text{FAR} = 1.0 \]

- 100% LOT COVERED
- 50% LOT COVERED
- 25% LOT COVERED

\[ \text{FAR} = 9.0 \]

- 100% LOT COVERED
- 100% LOT COVERED (COMBINATIONS)
Machine learning?

Use *machine learning* to predict zoning rules based on property assessment data
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

**Step 1:** Transform property assessment data into meaningful features

**Step 2:** Build a predictive model

**Step 3:** Evaluate our model
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

Step 1: Transform property assessment data into meaningful features
Machine learning?

Use *machine learning* to predict zoning rules based on property assessment data

Step 1: Transform property assessment data into meaningful features

<table>
<thead>
<tr>
<th>Property-Level</th>
<th>Zone-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lot size</td>
<td>Average lot size per home</td>
</tr>
<tr>
<td>Land use description</td>
<td>Share of low-density homes</td>
</tr>
</tbody>
</table>
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

Step 1: Transform property assessment data into meaningful features

Step 2: Build a predictive model
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

Step 2: Build a *predictive model*
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

Step 1: Transform property assessment data into meaningful features

Step 2: Build a predictive model

Step 3: Evaluate our model
Machine learning?

*Use machine learning to predict zoning rules based on property assessment data*

Step 3: **Evaluate** our model

How close are our predicted FARs to the true FARs?

- In-sample? Out-of-sample?
- Weighted RMSE? Weighted Relative MAE?
How do we do?
How do we do?

1. Predicting Washington, DC in-sample?
2. Predicting DC, Montgomery, and Arlington County in-sample?
3. Predicting Montgomery County out-of-sample?
4. Predicting Arlington County out-of-sample?
How do we do?

Predicting Washington, DC in-sample?

<table>
<thead>
<tr>
<th></th>
<th>Weighted RMSE</th>
<th>Weighted Relative MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>0.26</td>
<td>0.12</td>
</tr>
<tr>
<td>Non-Residential</td>
<td>0.39</td>
<td>0.04</td>
</tr>
</tbody>
</table>
How do we do?

Predicting Washington, DC in-sample?
How do we do?

Predicting DC, Montgomery, and Arlington County in-sample?

<table>
<thead>
<tr>
<th>Number of residential properties</th>
<th>Most common land use type</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>Residential</td>
</tr>
<tr>
<td>20000</td>
<td></td>
</tr>
<tr>
<td>30000</td>
<td></td>
</tr>
<tr>
<td>40000</td>
<td>Other</td>
</tr>
<tr>
<td>50000</td>
<td></td>
</tr>
<tr>
<td>60000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Weighted RMSE</th>
<th>Weighted Relative MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>0.21</td>
<td>0.09</td>
</tr>
<tr>
<td>Non-Residential</td>
<td>0.67</td>
<td>0.13</td>
</tr>
</tbody>
</table>
How do we do?

Predicting DC, Montgomery, and Arlington County \textit{in-sample}?
How do we do?

Predicting Montgomery County out-of-sample? (Training using DC)

<table>
<thead>
<tr>
<th>Number of residential properties</th>
<th>Most common land use type</th>
<th>Weighted RMSE</th>
<th>Weighted Relative MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>Residential</td>
<td>0.42</td>
<td>0.26</td>
</tr>
<tr>
<td>20000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Residential</td>
<td></td>
<td>1.54</td>
<td>2.23</td>
</tr>
</tbody>
</table>
How do we do?

Predicting Montgomery County out-of-sample? (Training using DC)
How do we do?

Predicting Arlington County out-of-sample?
(Training using DC & Montgomery County)

<table>
<thead>
<tr>
<th>Number of residential properties</th>
<th>Most common land use type</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>Residential</td>
</tr>
<tr>
<td>10000</td>
<td>Residential</td>
</tr>
<tr>
<td>15000</td>
<td>Other</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Weighted RMSE</th>
<th>Weighted Relative MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>0.80</td>
<td>0.31</td>
</tr>
<tr>
<td>Non-Residential</td>
<td>1.65</td>
<td>1.35</td>
</tr>
</tbody>
</table>
How do we do?

Predicting Arlington County out-of-sample?  
(Training using DC & Montgomery County)
What comes next?
What comes next?

1. How can we improve the model?
   More features and more jurisdictions!
What comes next?

1. How can we improve the model?
   More features and more jurisdictions!

2. Can this generalize?
   To other regions?
   To other built environments?
   To other zoning characteristics?
Can this generalize to other regions?

In some cases – yes!
In other cases – not with ZTRAX alone
Can this generalize?
Not with ZTRAX alone – we need open data!
“What are we zoned for and what have we built — what is the delta? It’s nearly impossible to know.”

– Ruby Bolaria, Chan Zuckerberg Initiative

We can use data science to unlock zoning data.
Links

Blog: https://greaterdc.urban.org/blog/we-need-better-zoning-data-data-science-can-help

Technical Appendix: https://www.urban.org/sites/default/files/2019/10/15/predicting_zoned_density_using_property_records_next_steps.pdf
Appendix Slides
Feature importance

Washington, DC in-sample

- Low-density neighbors
- Share of neighbors (residential)
- Medium-density neighbors
- Mean build year
- Minimum build year
- Share low-density homes
- Living space
- Commercial office zone
- Number of residential neighbors
- Vacant zone
- Residential zone
- Minimum remodel year
- Share medium-density homes
- Number of residential properties
- Governmental zone
- Multifamily residential zone
Feature importance

DC, Montgomery County, Arlington County in-sample

- Low-density neighbors
- Medium-density neighbors
- Share of neighbors (residential)
- Minimum remodel year
- Share medium-density homes
- Number of residential neighbors
- Vacant zone
- Commercial office zone
- Maximum build year

Importance
Feature importance

Montgomery County out-of-sample

- Low-density neighbors
- Medium-density neighbors
- Share of neighbors (residential)
- Minimum build year
- Mean build year
- Share low-density homes
- Living space
- Number of residential neighbors
- Commercial office zone
- Vacant zone
- Residential zone

Importance
Predicting Montgomery County

Training on DC & Arlington County

<table>
<thead>
<tr>
<th></th>
<th>Weighted RMSE</th>
<th>Weighted Relative MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>0.43</td>
<td>0.22</td>
</tr>
<tr>
<td>Non-Residential</td>
<td>1.26</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Number of residential properties

- 10000
- 20000
- 30000
- 40000
- 50000
- 60000

Most common land use type

- Residential
- Other

Relative distance between predicted and actual FAR

- 6
- 4
- 2
- 0