New data, new approaches

Thies Lindenthal
Twitter: @thieslindenthal
The sound of machine learning that is just logistic regression
Figure 5. Example Inferences from Model 1

Emma Waterhouse: “Applying Machine Learning and Image Classification to the Built Environment: The Case of UK High Streets”

Notes: Bounding boxes are colour-coded according to the identified classification whereby Green = ‘Occupied’; Red = ‘Vacant – Not Boarded-Up’; Orange = ‘Vacant – Boarded-Up’.
We have a series of papers tapping into new data...
3D city models, linked to property values

Beauty in the Eye of the Home-Owner:
Aesthetic Zoning and Residential Property Values

Thies Lindenthal

This article empirically confirms one core motivation for architectural zoning: Shape homogeneity among neighboring homes increases the value of residential buildings. Drawing on large-scale shape and transaction data, this study first develops a data-driven measure of architectural similarity, condensing three-dimensional shapes to univariate shape distributions. These algorithm-based similarity estimates are good predictors of human perceptions of shape similarity.
(Change in) Building volumes, remotely sensed
Lindenthal (2017)

Estimating Supply Elasticities for Residential Real Estate in the United Kingdom

Fig. 6  Change in building volumes.
Notes: The left panel displays building volumes as derived from the difference between a DSM and DTM for a London neighborhood in 2009. Elevations above the ground are visualized in light colors. The right panel depicts the same area in 2015. New construction in the upper right corner is the most visible change between 2009 and 2015.

Finally, for those areas where DEM models from multiple years overlap, we can calculate changes in the building volumes across years: \( \Delta \text{Volume}_{t1,t2,xy} = \text{Volume}_{t2,xy} - \text{Volume}_{t1,xy} \). Fig. 6 visualizes this difference in volumes calculation. The left panel displays building volumes as derived from the difference between a DSM and DTM for a London neighborhood in 2009. Elevations above the ground are visualized in light colors. The right panel depicts the same area in 2015. New construction in the upper right corner is the most visible change between 2009 and 2015.

The volume data can be aggregated at any desired level. In the remainder, we will work at the lowest geographical level of the British census, the output area (OA). Census OAs are constructed by merging adjacent unit postcodes. The OAs have approximately the same population sizes and, more relevant for this study, are designed to be socially homogeneous in terms of home-ownership rates and property types (Office for National Statistics, 2016b).
Machine Learning, Building Vintage and Property Values

Thies Lindenthal (University of Cambridge) & Erik B. Johnson (University of Alabama)

07 April 2019

Abstract

This paper makes three contributions: First, it introduces an algorithm that collects pictures of individual buildings from Google Street View. Second, it trains a deep convolutional neural network (CNN) to classify residential buildings into architectural styles, taking into account spatial dependencies of building vintages. Third, it investigates whether architectural styles influence house prices. For re-sales, the architectural style is a significant determinant of transaction prices while no such effect is found for new buildings. This indicates that any premia are for vintage-related quality characteristics and not for a home’s beauty.
Asset uniqueness
(with Carolin Schmidt)

The Odd One Out:
Asset Uniqueness and Price Precision

April 2, 2019

Based on applied machine learning (ML) techniques this paper suggests that round prices are not purely random events but are linked to liquidity and the uniqueness of the asset. First, using residential transaction data from the UK, we show that the availability of information from comparable sales influences the odds of observing a sale at a round price. Second, we explore ways to play to the strengths of deep neural network and incorporate computer vision approaches and building level imagery. Adding information on a building’s vintage and the typology of its direct surroundings to the training data boosts the predictive power of the suggested ML classifiers. When a house is “the odd one out”, its value will be relatively difficult to establish which implies that sales prices suffer from a relatively high signal-to-noise ratio.
From large, unstructured data to economic analysis
Example Computer Vision: Two layers of Machine Learning

Images → Feature Vector

$\begin{bmatrix}
0.4 \\
0.2 \\
0.5 \\
...\\
0.3
\end{bmatrix}$

Classification/quantification → Price effects

ML
Collecting images is not straightforward in the UK.

Example: 84 Vinery Road, Cambridge.
Solution: Bring a map
Using land registry maps to estimate best angle for Google Street View picture
Learning more about the housing stock, nationwide
As we speak, a picture/feature vector database is built
Building large groundtruth data set
Human expert knowledge forms basis for machine learning

- Collect mugshots of all suitable buildings in Cambridge, UK (48,000)
- Architects defined meaningful vintage/style classes
  - Vintages share design, location and quality characteristics
  - Broadly used by realtors, buyers or sellers
- Architects classified 25,000 images according to these categories
- Yes, we are sharing!
Computer vision, off the shelve
Deep convolutional neural network (CNN) to obtain feature vectors

- Pre-trained *Inception v3* model in Tensorflow API
  - Convolutional: Exceptionally suitable to detect era specific details such as window styles, ratios, brickwork, ratios
  - Freely available & frequently used
- Penultimate layer is 2048 dimensional feature vector
Meet the “Zoo”
(Yes, that’s how the universe of training data and model combinations is called)

COCO-trained models

<table>
<thead>
<tr>
<th>Model name</th>
<th>Speed (ms)</th>
<th>COCO mAP[^1]</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssd_mobilenet_v1_coco</td>
<td>30</td>
<td>21</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v1_0.75_depth_coco □</td>
<td>26</td>
<td>18</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v1_quantized_coco □</td>
<td>29</td>
<td>18</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v1_0.75_depth_quantized_coco □</td>
<td>29</td>
<td>16</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v1_ppn_coco □</td>
<td>26</td>
<td>20</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v1_fpn_coco □</td>
<td>56</td>
<td>32</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_resnet_50_fpn_coco □</td>
<td>76</td>
<td>35</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v2_coco</td>
<td>31</td>
<td>22</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_mobilenet_v2_quantized_coco</td>
<td>29</td>
<td>22</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssdlite_mobilenet_v2_coco</td>
<td>27</td>
<td>22</td>
<td>Boxes</td>
</tr>
<tr>
<td>ssd_inception_v2_coco</td>
<td>42</td>
<td>24</td>
<td>Boxes</td>
</tr>
<tr>
<td>faster_rcnn_inception_v2_coco</td>
<td>58</td>
<td>28</td>
<td>Boxes</td>
</tr>
<tr>
<td>faster_rcnn_resnet50_coco</td>
<td>89</td>
<td>30</td>
<td>Boxes</td>
</tr>
<tr>
<td>faster_rcnn_resnet50_lowproposals_coco</td>
<td>64</td>
<td></td>
<td>Boxes</td>
</tr>
</tbody>
</table>
Building a property zoo

- Re-estimate tried and tested computer vision models, from ground up, using domain relevant imagery
- No more cats and dogs

COMPUTER VISION AND REAL ESTATE:
DO LOOKS MATTER AND DO INCENTIVES DETERMINE LOOKS

Edward L. Glaeser
Michael Scott Kincaid
Nikhil Naik

Working Paper 25174
http://www.nber.org/papers/w25174
High-dimensional data, nonlinearities
Traditional econometric techniques reach their limits

• Glaeser (2018) uses PCA to reduce dimensionality, followed by regressions
• Example: Vintage detection
  • Input dimensionality (2048 or even 4096) rules out multinomial logit models
  • (somewhat deep) neural network, only two hidden layers
  • Using standardised APIs/frameworks as much as we can
How to evaluate performance?

Data science concepts: Confusion matrices, Recall, Precision, F₁-Scores, ...

Confusion matrix, “base” model (Lindenthal & Johnson)

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Georgian</th>
<th>Early Vict.</th>
<th>Late V./Edw.</th>
<th>Interwar</th>
<th>Postwar</th>
<th>Cont.</th>
<th>Revival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgian</td>
<td>267</td>
<td>121</td>
<td>110</td>
<td>46</td>
<td>33</td>
<td>36</td>
<td>17</td>
</tr>
<tr>
<td>Early.Vic.</td>
<td>50</td>
<td>1673</td>
<td>438</td>
<td>112</td>
<td>107</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>Late V./Edw.</td>
<td>19</td>
<td>183</td>
<td>3099</td>
<td>272</td>
<td>86</td>
<td>33</td>
<td>44</td>
</tr>
<tr>
<td>Interwar</td>
<td>12</td>
<td>71</td>
<td>335</td>
<td>5786</td>
<td>1419</td>
<td>84</td>
<td>82</td>
</tr>
<tr>
<td>Postwar</td>
<td>2</td>
<td>24</td>
<td>56</td>
<td>462</td>
<td>3364</td>
<td>75</td>
<td>38</td>
</tr>
<tr>
<td>Cont.</td>
<td>5</td>
<td>63</td>
<td>82</td>
<td>141</td>
<td>366</td>
<td>778</td>
<td>79</td>
</tr>
<tr>
<td>Revival</td>
<td>6</td>
<td>43</td>
<td>54</td>
<td>156</td>
<td>170</td>
<td>114</td>
<td>437</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Georgian</th>
<th>Early Vict.</th>
<th>Late V./Edw.</th>
<th>Interwar</th>
<th>Postwar</th>
<th>Cont.</th>
<th>Revival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgian</td>
<td>74%</td>
<td>6%</td>
<td>3%</td>
<td>1%</td>
<td>1%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>Early.Vic.</td>
<td>14%</td>
<td>77%</td>
<td>10%</td>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>7%</td>
</tr>
<tr>
<td>Late V./Edw.</td>
<td>5%</td>
<td>8%</td>
<td>74%</td>
<td>4%</td>
<td>2%</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Interwar</td>
<td>3%</td>
<td>3%</td>
<td>8%</td>
<td>83%</td>
<td>26%</td>
<td>7%</td>
<td>11%</td>
</tr>
<tr>
<td>Postwar</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>7%</td>
<td>61%</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>Cont.</td>
<td>1%</td>
<td>3%</td>
<td>2%</td>
<td>2%</td>
<td>7%</td>
<td>66%</td>
<td>11%</td>
</tr>
<tr>
<td>Revival</td>
<td>2%</td>
<td>2%</td>
<td>1%</td>
<td>2%</td>
<td>3%</td>
<td>10%</td>
<td>58%</td>
</tr>
</tbody>
</table>

Recall

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.74</td>
<td>0.77</td>
<td>0.74</td>
<td>0.83</td>
<td>0.61</td>
<td>0.66</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Precision

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.42</td>
<td>0.67</td>
<td>0.83</td>
<td>0.74</td>
<td>0.84</td>
<td>0.51</td>
<td>0.45</td>
</tr>
</tbody>
</table>

$F_1$-score

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$-score</td>
<td>0.54</td>
<td>0.72</td>
<td>0.78</td>
<td>0.78</td>
<td>0.70</td>
<td>0.58</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Give me my asterisks*** back, please!
Causal inference in ML via predictive power?

- Reverse regressions: Will the ML prediction be a significant regressor?
  - Building Vintage = Year + Location + Type + ... + ML-Prediction

<table>
<thead>
<tr>
<th></th>
<th>No ML pred.</th>
<th>Base pred.</th>
<th>Spatial pred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgian</td>
<td>1282</td>
<td>914</td>
<td>912</td>
</tr>
<tr>
<td>Early Victorian</td>
<td>10232</td>
<td>6209</td>
<td>5712</td>
</tr>
<tr>
<td>Late Vic./Edw.</td>
<td>16223</td>
<td>8863</td>
<td>8186</td>
</tr>
<tr>
<td>Interwar</td>
<td>18334</td>
<td>10009</td>
<td>8895</td>
</tr>
<tr>
<td>Postwar</td>
<td>16846</td>
<td>11228</td>
<td>9471</td>
</tr>
<tr>
<td>Contemporary</td>
<td>2885</td>
<td>1926</td>
<td>1743</td>
</tr>
<tr>
<td>Revival</td>
<td>3336</td>
<td>2255</td>
<td>2099</td>
</tr>
</tbody>
</table>
Our Toolbox
(Not very dogmatic)

- R
- Keras / Tensorflow
- PostGIS / Postgres for databases
- Ionic framework to build apps
From large, unstructured data to economic analysis
Example Computer Vision: Two layers of Machine Learning

Images

Feature Vector

Classification

Further analysis

ML

\[
\begin{bmatrix}
0.4 \\
0.2 \\
0.5 \\
... \\
0.3
\end{bmatrix}
\]

ML

Classification/quantification

Price effects
Models and People

Inference

Style propabilities:

Contemp. Victorian 27.06 %
Georgian 21.08 %
Interwar 20.74 %

Correct!
Wrong
I Don't Know
Shared challenges

• Computing power: Solved
• ML models: OK (good enough for now)
• Large training data sets: We need them all
• Pretrained models: nothing property specific
• Feedback: Expensive to collect/manage
• Methodology not common, too many rookie blunders
How to defeat nimbyism: build more beautiful houses

The government hopes that better architecture could make it easier to build