Machine Learning and Land Values

[Black Boxes all the way down]

Erik Johnson
University of Alabama
erikbjohn@gmail.com

Overview

- Use images to predict land values
  - Deep learning based approach
    - Transfer Learning
    - Multiple images for each land sale
    - Increasing sophisticated mixes of photos
    - Most similar in spirit to kriging (spatially correlated land prices) / observable matching
- Compare Results to Linear Regression, Basic Neural Nets applied to traditional data
- Train using black box masking to enable prediction on properties with structures [decouple land and structures]
- Examine the role of region, neighborhood and block in price
- All fit results from land sales that were not used in the training data set

Data

- Land Sales Data – Maricopa County AZ
  - Restrict to 2017- 2018 for API based image access
  - 2377 ‘training’, 595 ‘testing’
  - Normalized Log price/sqft [0, 1]
- 5 pictures of each parcel
  - Streetview ‘own’
  - Streetview ‘across’
  - Satellite Zoom 13 [Region]
  - Satellite Zoom 15 [Neighborhood]
  - Satellite Zoom 18 [Block]
- Full Sales Data – Maricopa County [All normalized to 0,1]

Image Data

(a) Zoom Param 13

(b) Zoom Param 15

(c) Zoom Param 18

Image Data

(a) Own Image

(b) Across Street Image

Models

Tabular Data Only

- Traditional Neural Net
- Linear Regression

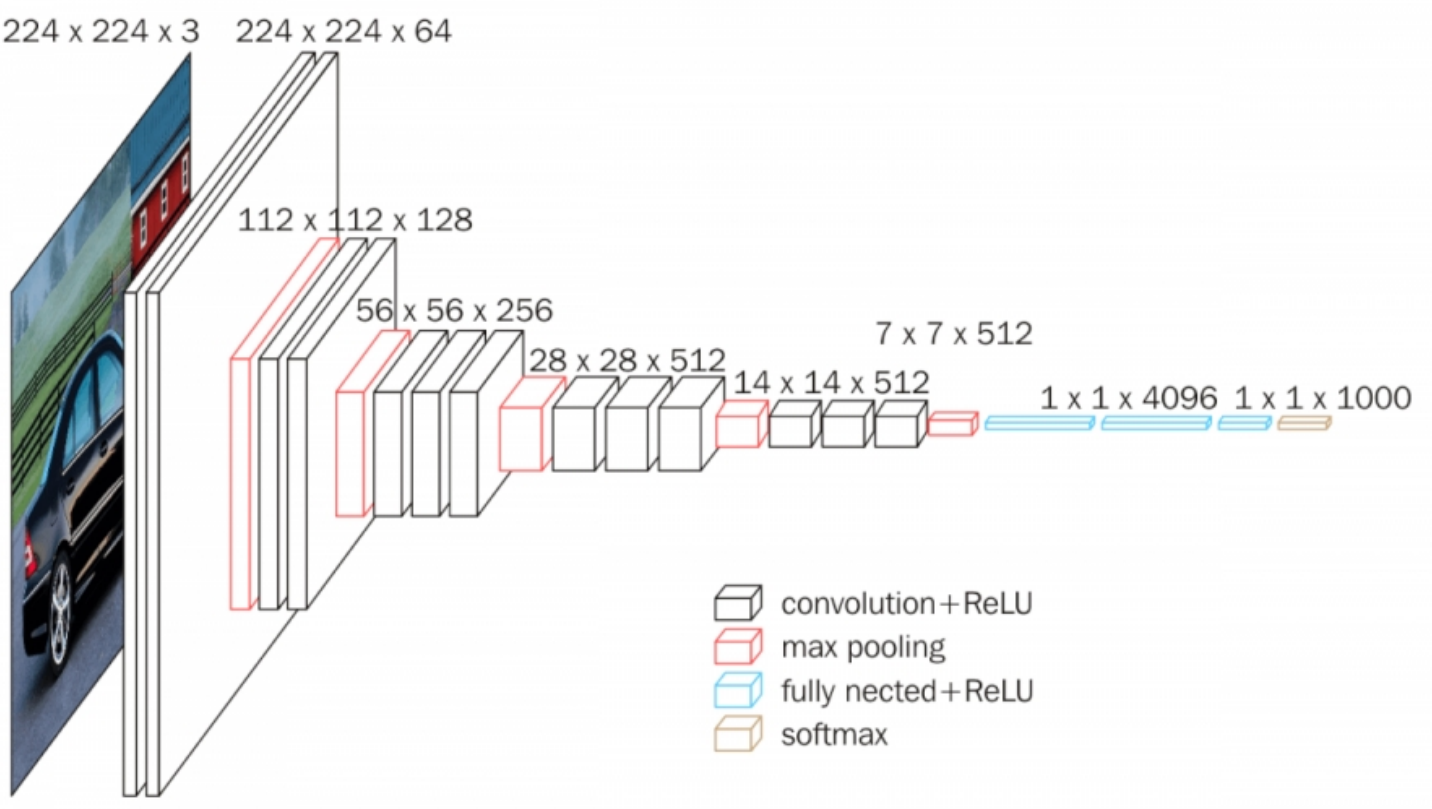
Image Data Only

- Sat 18 (centered on parcel lat/lon)
- Sat 15 (centered on parcel lat/lon)
- Sat 13 (centered on parcel lat/lon)
- Street Own
- Street Across
- Combined

Image Data Models

- All Models use Transfer Learning
  - Very useful for ‘small’ datasets
  - VGG16 pretrained classification algo – dimension reduction
  - Strip off the last layers, add new hidden and dropout layers and convert to predict continuous surface
  - Train on all 5 photos simultaneously with and without tabular data

VGG16 Structure



Does this work at all?

- Land Sales Data – Maricopa County AZ
  - Restrict to 2017- 2018 for API based image access
  - 2377 ‘training’, 595 ‘testing’
  - Normalized Log price/sqft [0, 1]
- 5 pictures of each parcel
  - Streetview ‘own’
  - Streetview ‘across’
  - Satellite Zoom 13 [Region]
  - Satellite Zoom 15 [Neighborhood]
  - Satellite Zoom 18 [Block]
- Full Sales Data – Maricopa County [All normalized to 0,1]

First Results – High Value Sat 13

(a) 17141002: Score=0.759

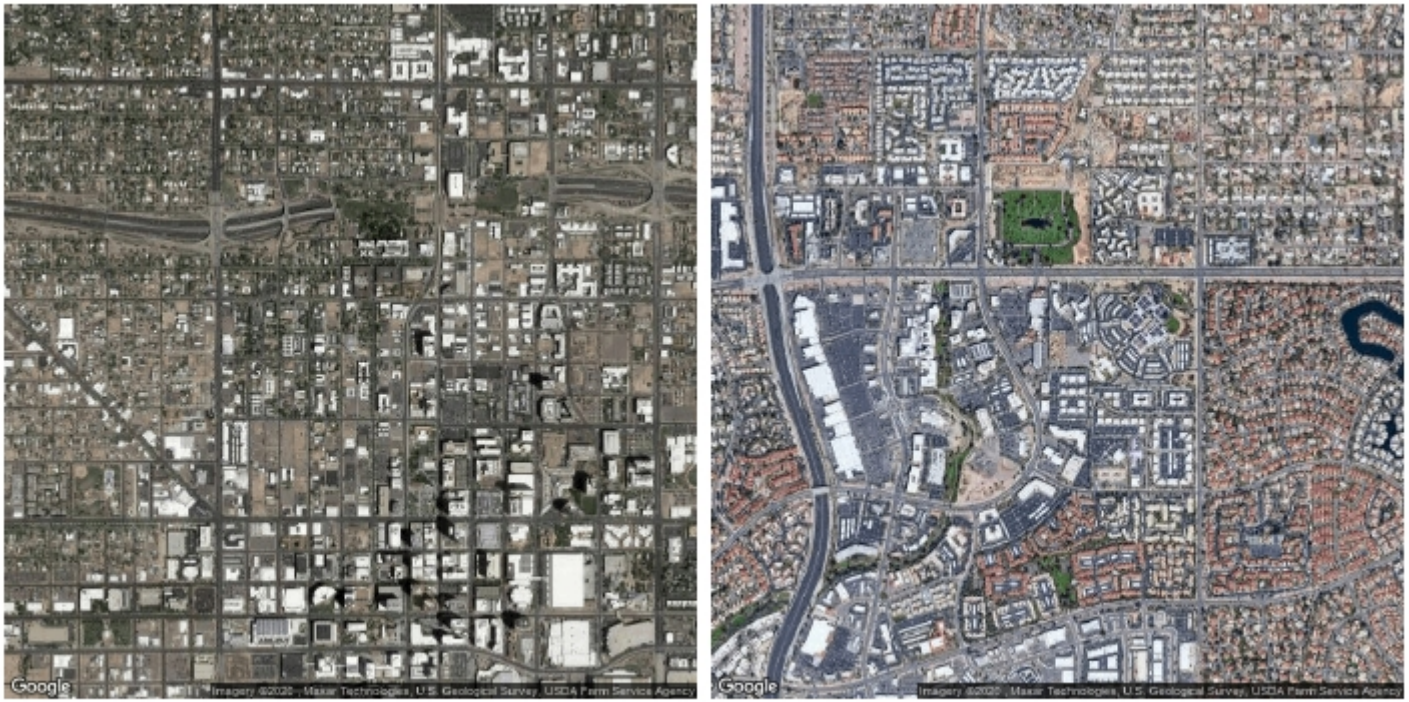
(b) 21543331D: Score=0.752

First Results – Low Value Sat 13

(e) 50415582: Score=0.267

(f) 50645255: Score=0.236

First Results – High Value Sat 15



(a) 11140091: Score=0.886

(b) 21736989B: Score=0.822

First Results – Low Value Sat 15

(e) 50645255: Score=0.226

(f) 50381039: Score=0.224

First Results – High Value Sat 18

(a) 11146141: Score=0.826

(b) 21543331D: Score=0.826

First Results – Low Value Sat 18

(e) 40051012M: Score=0.248

(f) 50641273: Score=0.240

First Results – High Value Own

(a) 11702011: Score=0.708

(b) 11905074: Score=0.706

First Results – Low Value Own

(e) 40076123A: Score=0.249

(f) 50641279: Score=0.238

First Results – High Value Across

(a) 17242068: Score=0.818

(b) 12728090: Score=0.802

First Results – Low Value Across

(e) 50415235: Score=0.0.256

(f) 20121011BA: Score=0.249

First Results - Discussion

- Seem reasonable
- Predictions based on single zoom levels
  - Not sure if information is unique since each trained separately
  - Possible to combine all images in deeper model for better learning
  - Possible to combine all images in deeper model for better learning

First Results – Accuracy (RMSE)

perspective	land_contemp
sat18	0.0809
sat15	0.0894
sat13	0.0938
own	0.1027
across	0.1068
data_nn	0.1251
data_linear	0.1321

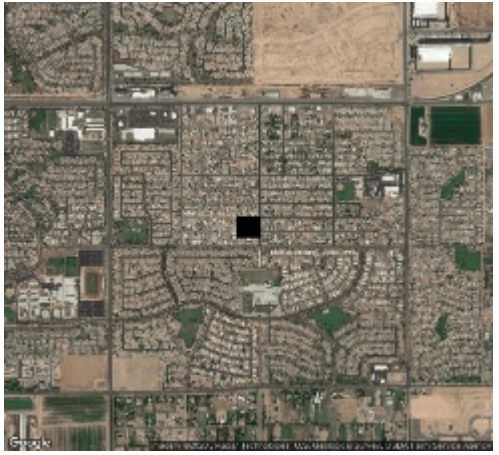
First Results – Model Correlations

	sat18	sat15	sat13	own	across	data_nn	data_linear
sat18	1.000	0.764	0.761	0.691	0.689	0.385	0.348
sat15	0.764	1.000	0.851	0.647	0.653	0.458	0.403
sat13	0.761	0.851	1.000	0.634	0.652	0.502	0.436
own	0.691	0.647	0.634	1.000	0.654	0.353	0.302
across	0.689	0.653	0.652	0.654	1.000	0.349	0.356
data_nn	0.385	0.458	0.502	0.353	0.349	1.000	0.727
data_linear	0.348	0.403	0.436	0.302	0.356	0.727	1.000

Ensemble and masking

- Combine all images in series of ‘deep’ models
  - Unmasked (all 5 pictures)
    - Best way to predict undeveloped land values
  - Masked (drop ‘own’ and cover location with black box in the sat photos)
    - No structures used in training! [mask the structure]
    - Main assumption: Surroundings drive land values

Masking Example
[Black boxes]



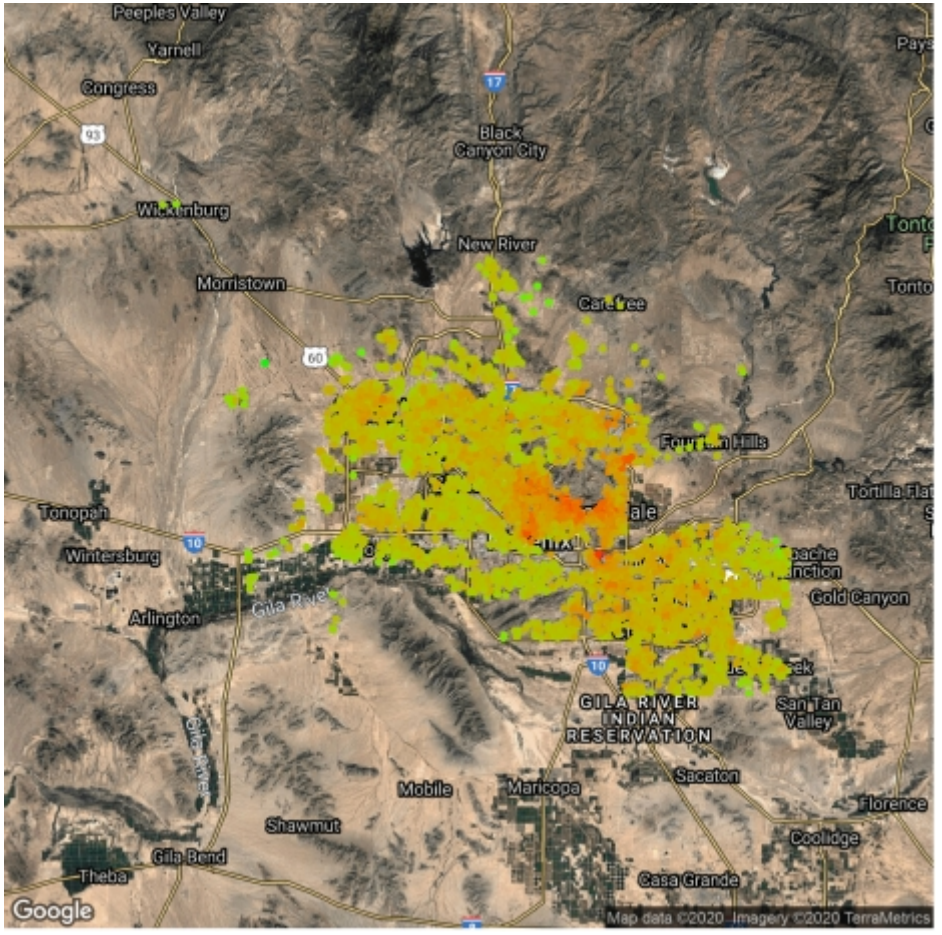
Deep model Results [masked and unmasked]

model	RMSE
sat 13 15 18 own across	0.0779
sat 13 15 18	0.0815
sat 13 15 18 across with mask	0.0789
sat 13 15 18 with mask	0.0819

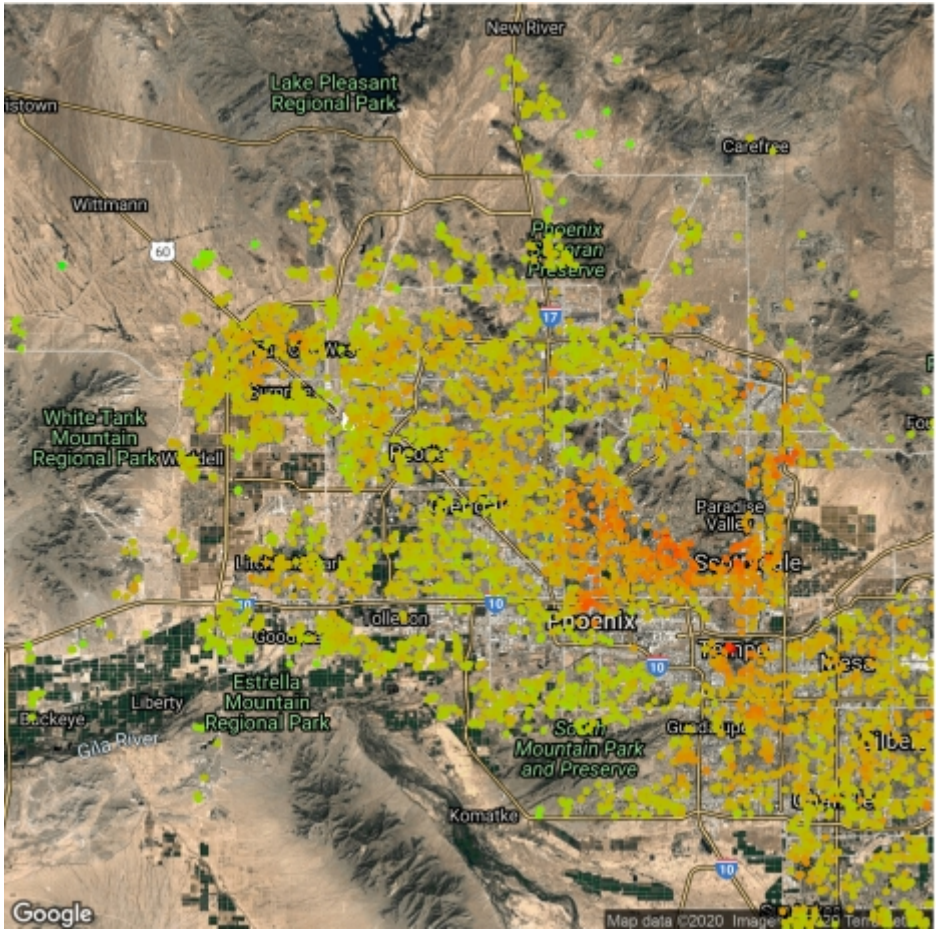
Out – of -sample predictions (full sales data)

- Used this since full sales data has lat and lon
- Predict price using sat 13 15 18 with mask
  - Value without structures [land value]

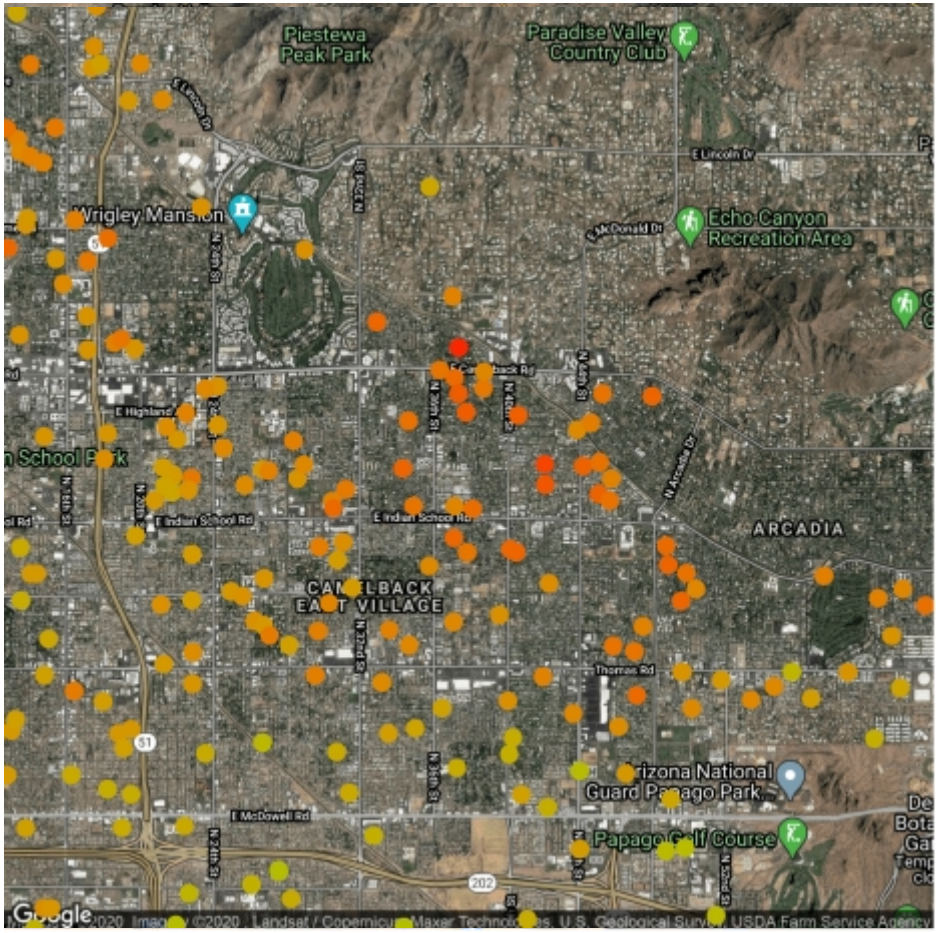
Analysis – Land Price Surface Estimation



Analysis – Land Price Surface Estimation



Analysis – Land Price Surface Estimation



Analysis – Land Price Surface Estimation

Counterfactual analysis

- What happens to price if we put a rural block in the highest value are of the city?
- What happens if we put an expensive block in a rural area?
- Lots of issues but given hugely nonlinear interactions, interesting to look at.

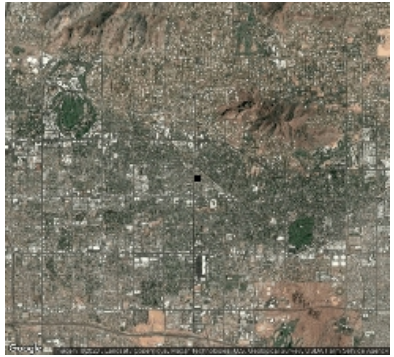
SAT 13 ACTUAL

SAT 15 ACTUAL

SAT 18 ACTUAL

Base Prediction

0.246



0.629

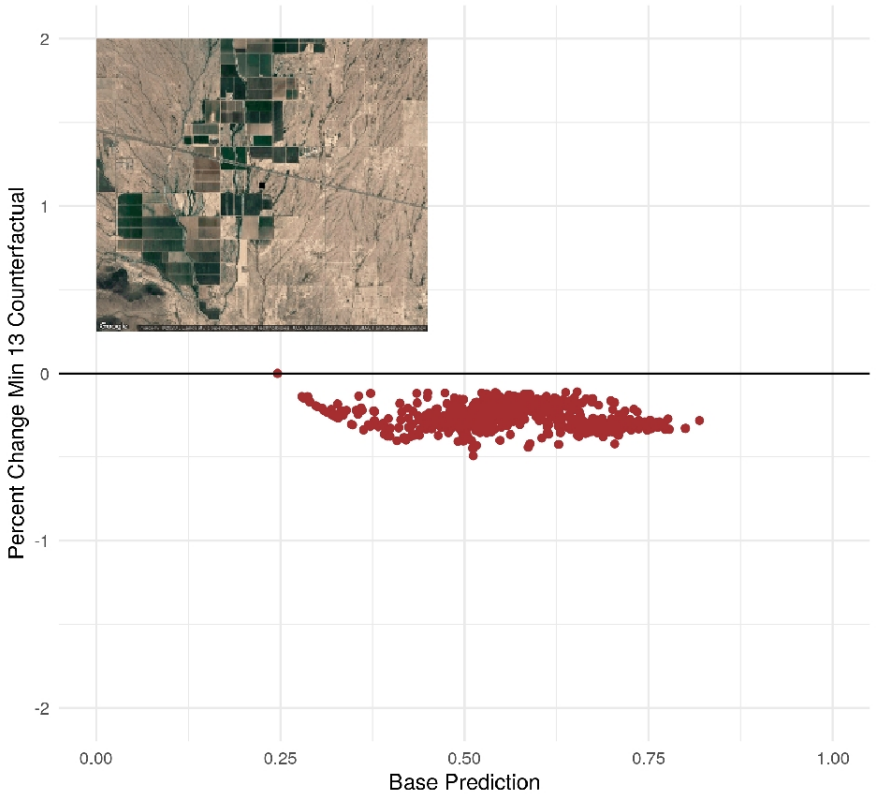
Max 13 Counterfactual

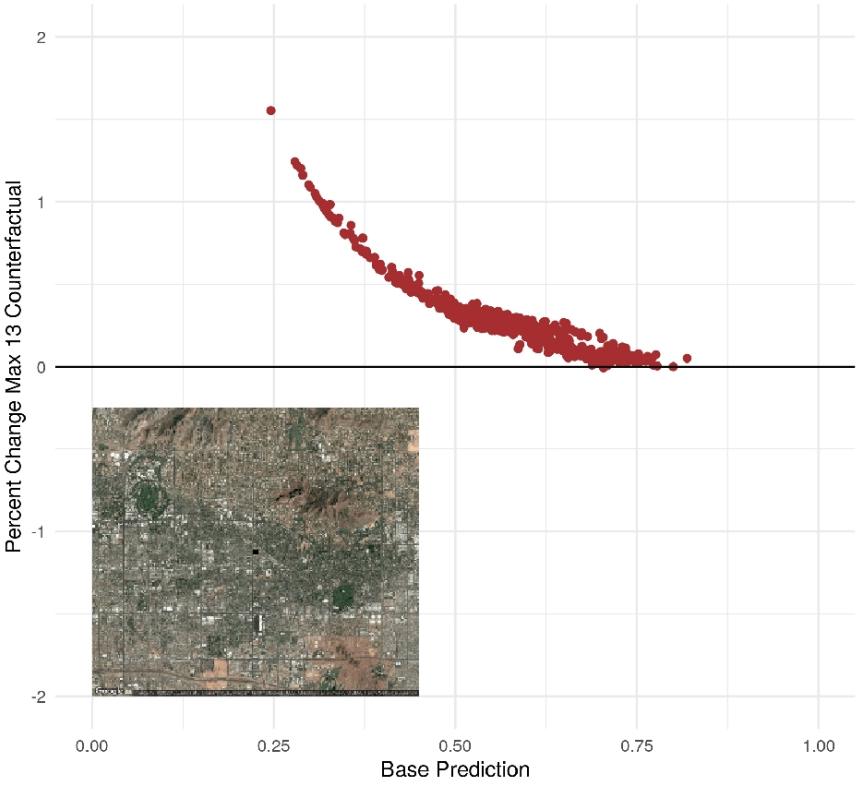
SAT 13 HIGH VALUE

SAT 15 ACTUAL

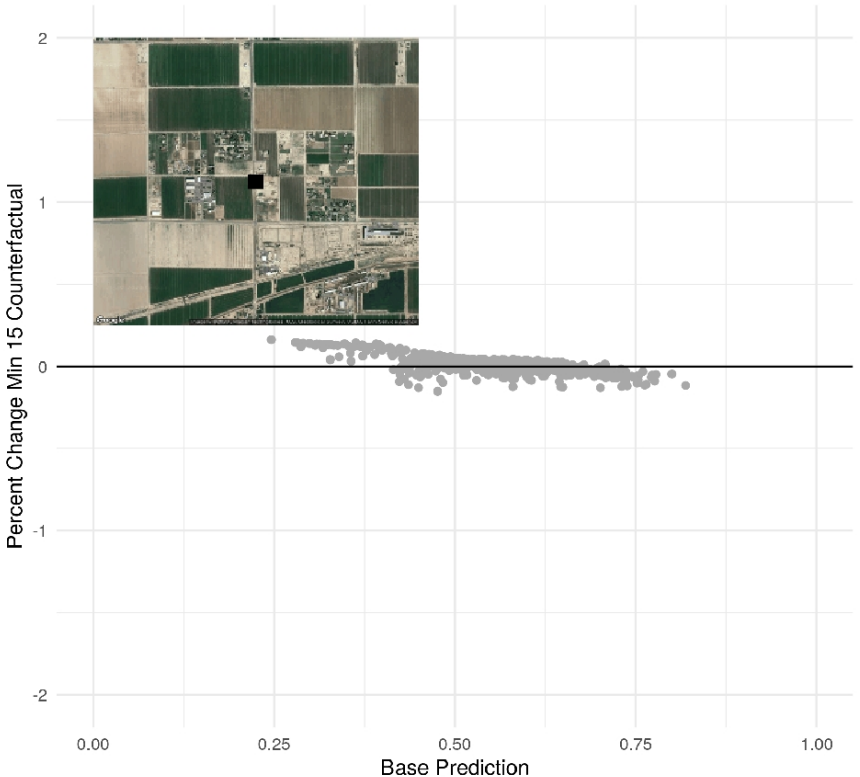
SAT 18 ACTUAL

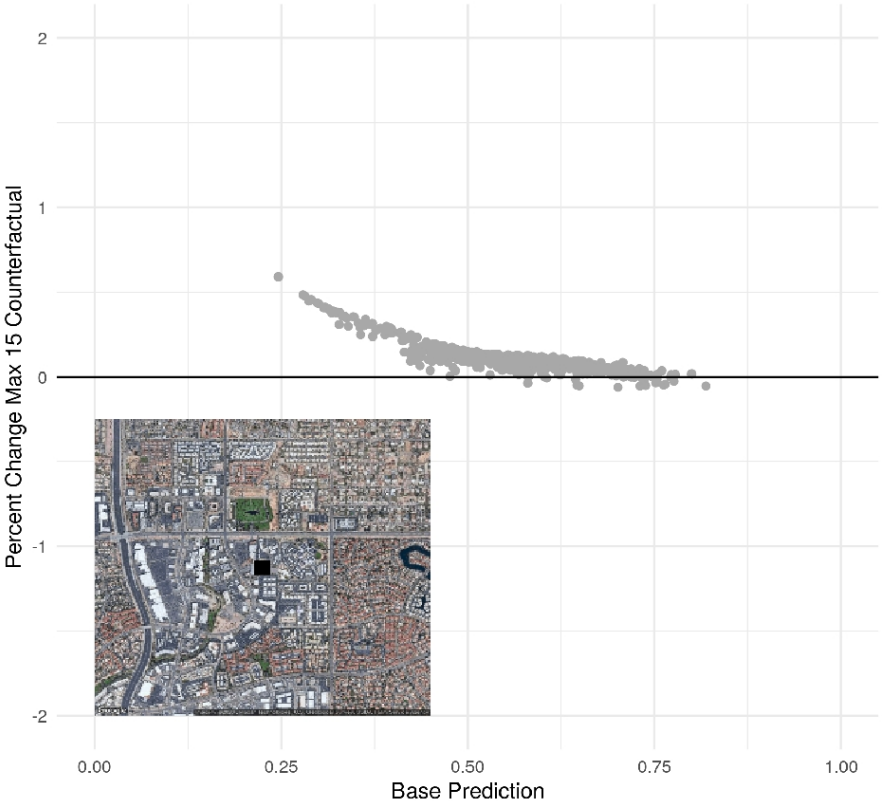
Counterfactual – Sat 13



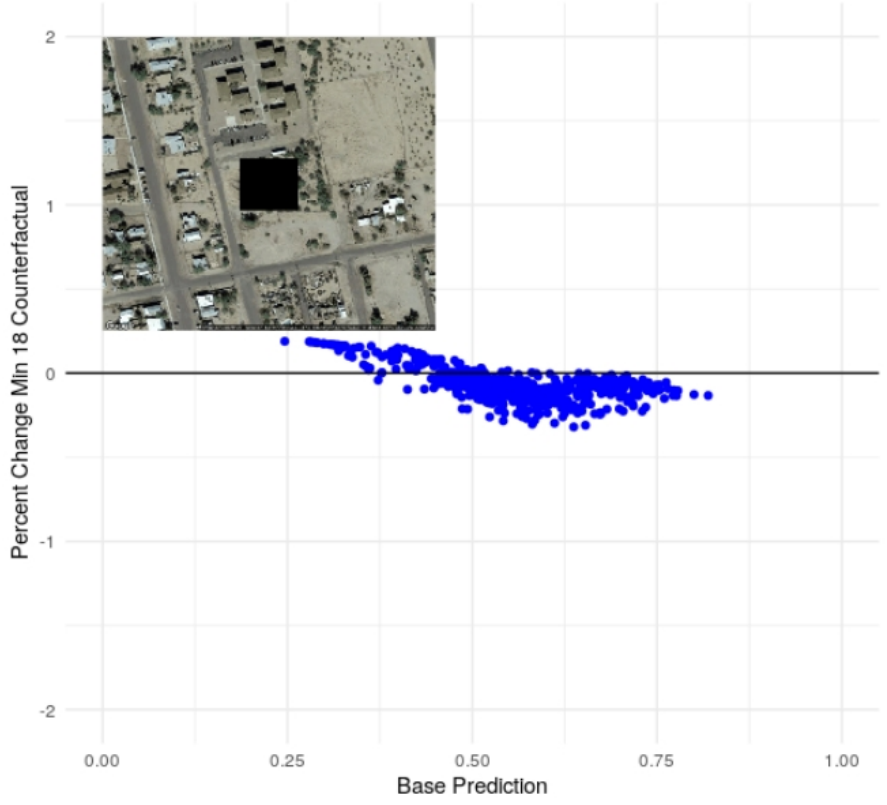


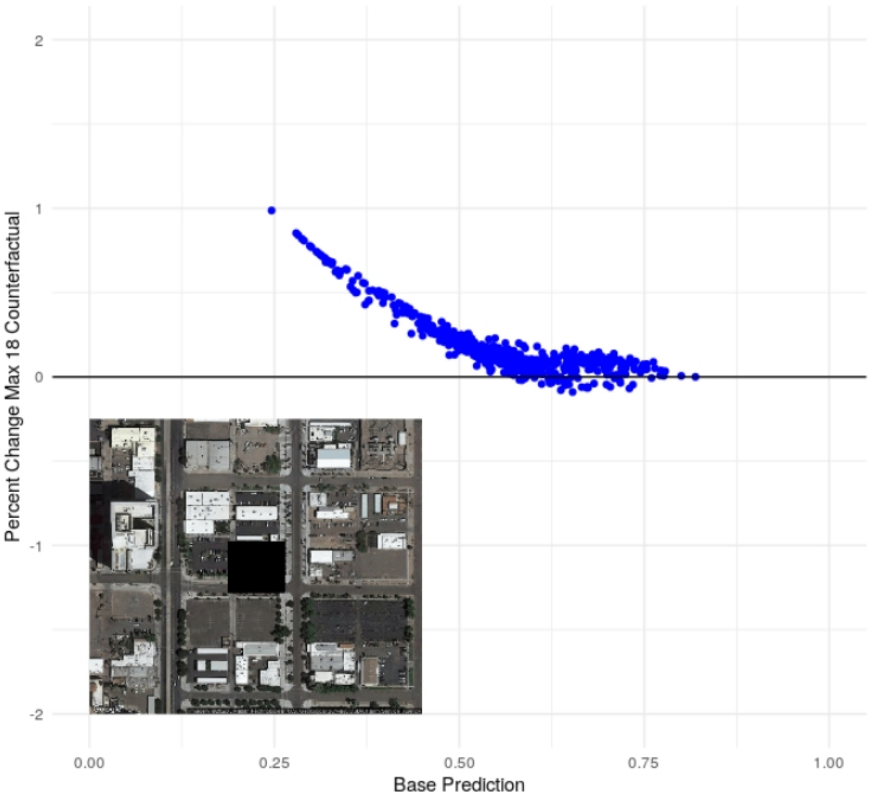
Counterfactual – Sat 15





Counterfactual – Sat 18





Conclusion

- Satellite imagery alone predicts land values well
- Masking allows for generalization
- Heterogeneous effects of ‘place’
- Contemporaneous image data matters (probably why the across and own street view data did not contribute in large part to RMSE)
- Extremely practical and models can be updated on the fly as data comes in (using mini-batch gradient descent)
- Future is in not “one perfect model” but ensemble results from multiple models.
- Easy to include year fixed effects – if relevant image data is available.
- Only 2972 observations for training this model – more data would greatly improve SE