CUSTOM SMALL AREA ESTIMATES

THE NOW AND THE FUTURE

Fennis Reed
Demographic Research Unit
CA Department of Finance
6/8/2021

Road Map

GOAL: Demonstrate current methods for SAE in California.
Encourage feedback and discussion.

Introduction & Study Area

Random Forest

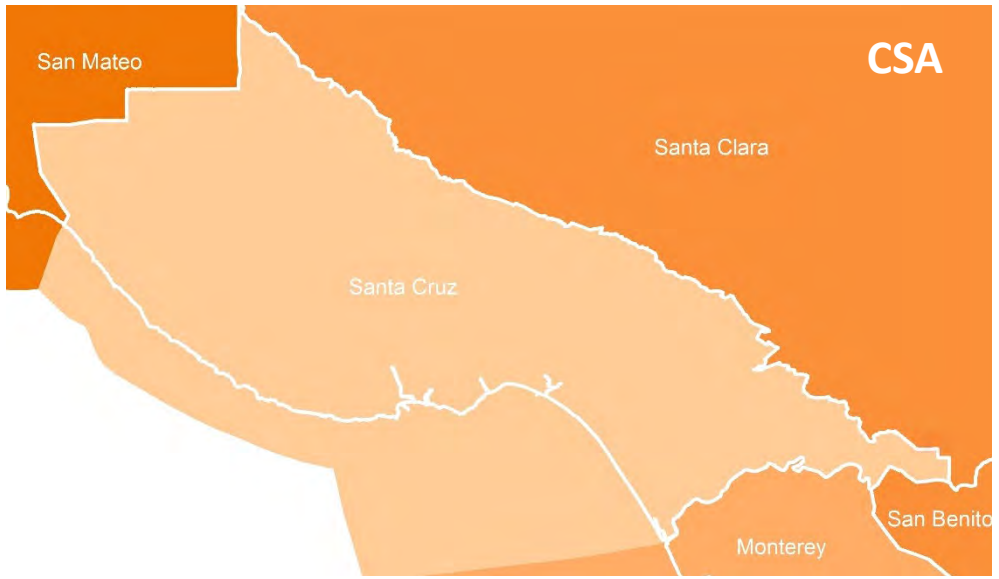
CEDS

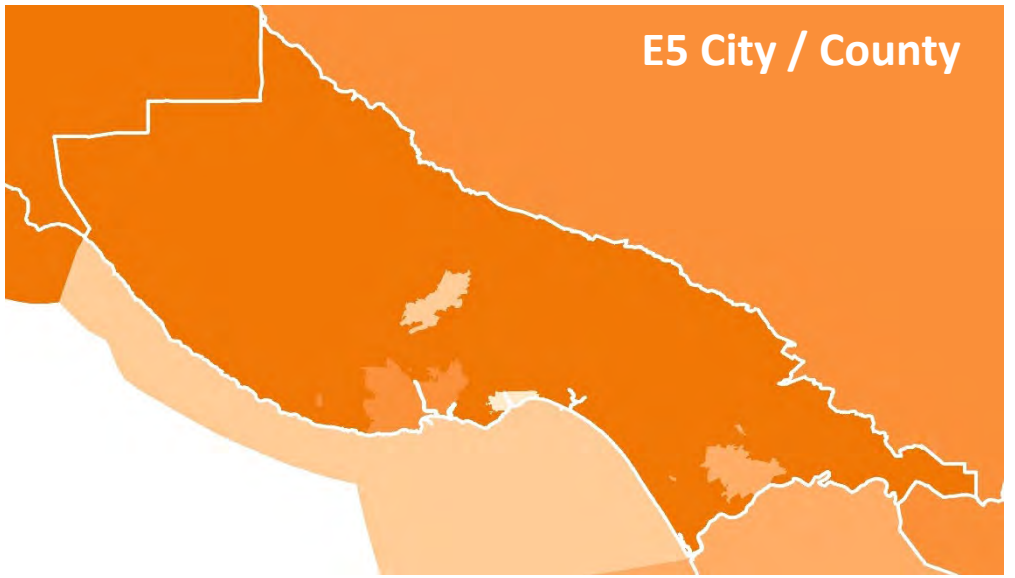
Comparison of Models

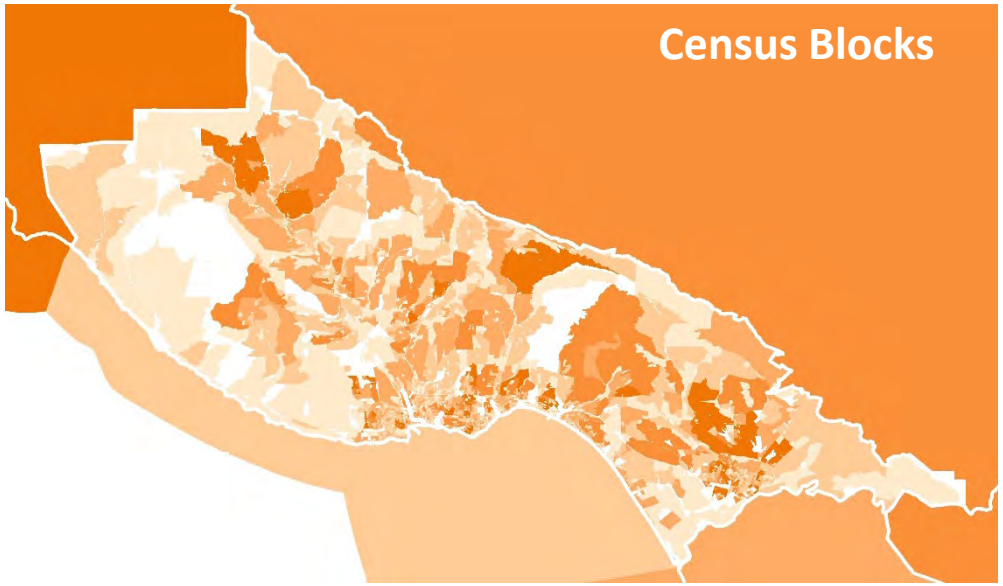
Case Studies

Future Work

Small Area Estimates







Small Area Estimates

- A series of statistical approaches concerned with estimating a parameter at a sub-survey scale.
  - Fill the gap between official statistics and requests of local data.

Bottom-up

- Uses ancillary data sources and statistical modeling to construct estimates.
- Often validated against a larger known survey unit.

Top down

- Distributes data from a larger unit according to some ancillary data.
- Often validated against the finest known unit

Small Area Estimates

- Housing Unit Method

$$Population = HU * PPH + GQ$$

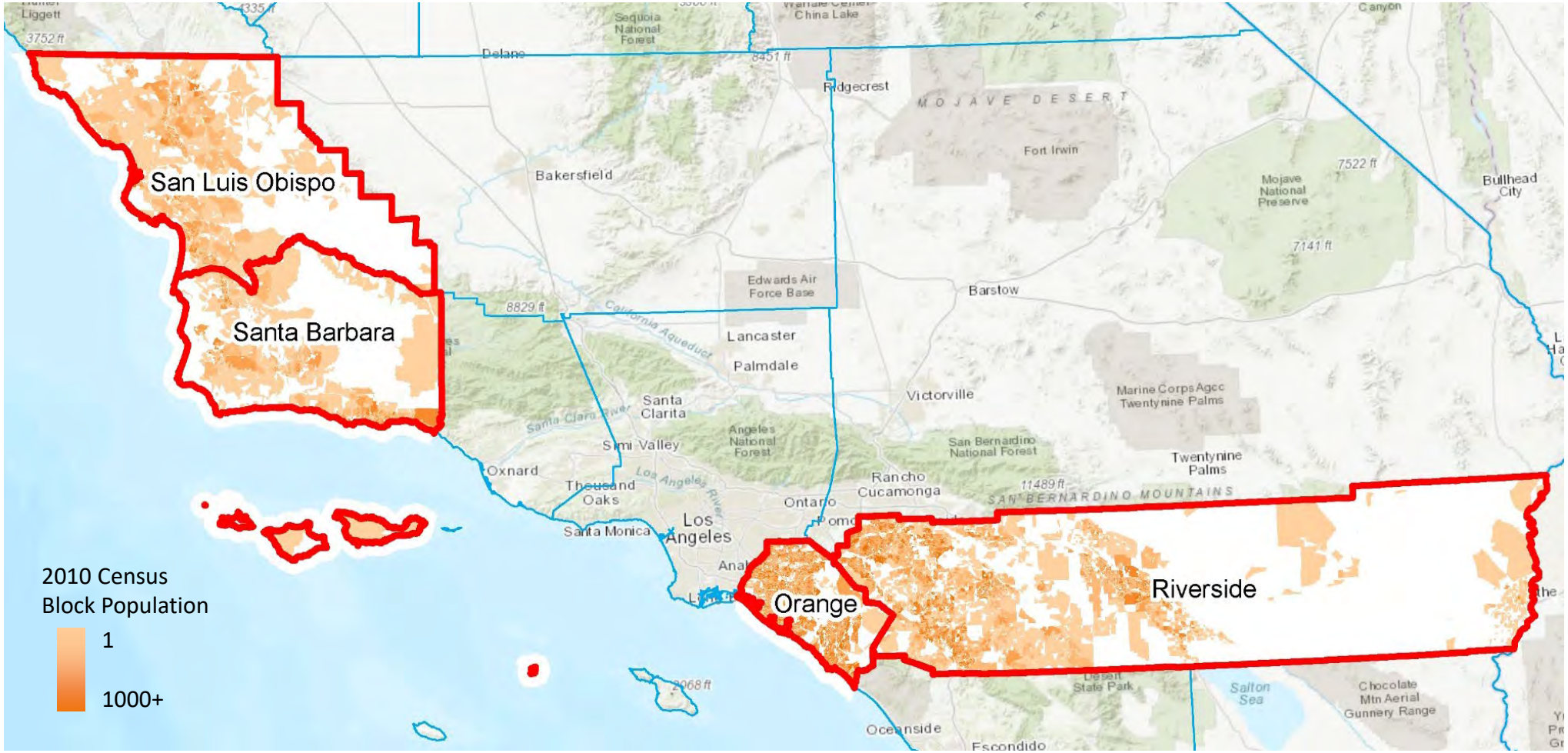
HU: Housing units
PPH: Average persons per Household
GQ: Population in group quarters

  - Requires consistent geographic coverage
  - PPH & MAUP

- Random Forest + CEDS

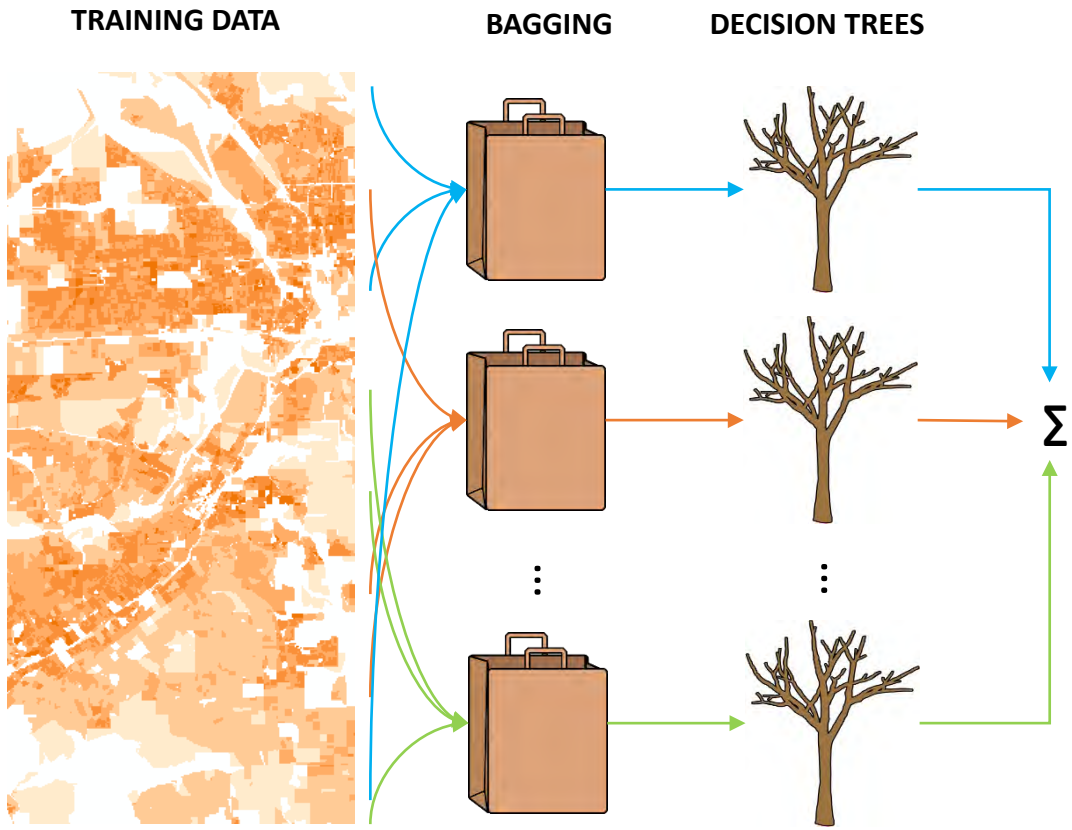
(Smith, S.K. 1986. Smith, S.K., S. Cody. 1994. Mennis, 2009)

Study Area


San Luis Obispo
Santa Barbara
Orange
Riverside
2010 Census
Block Population
1
1000+
Bakersfield
Ridgecrest
Fort Irwin
Barstow
Lancaster
Palmdale
Victorville
Edwards Air
Force Base
Mojave
National
Preserve
Bullhead
City
Marine Corps Agcc
Twentynine Palms
Twentynine
Palms
San Bernardino
National Forest
Angeles
National
Forest
Sequoia
National
Forest
Simi Valley
Oxnard
Thousand
Oaks
Santa
Clarita
Los
Angeles
Santa Monica
Ontario
Rancho
Cucamonga
Oceanside
Escondido
Salton
Sea
Chocolate
Mtn Aerial
Gunnery Range
Desert
State Park
MOJAVE DESERT
SAN BERNARDINO MOUNTAINS
California Aqueduct
Santa Clara River
Los Angeles River
Delano
China Lake
Liggett
Canyon
3752 ft
8451 ft
8829 ft
7522 ft
7141 ft
11489 ft
2068 ft

Random Forest: What is it?

- Bundle of regression trees
- Feature Bagging
- Back-transformed over the mean RF trees
- Applied in prediction



Random Forest: Parameters

Response Variable:

- 2010 Census block log density > 0

Forest Parameters:

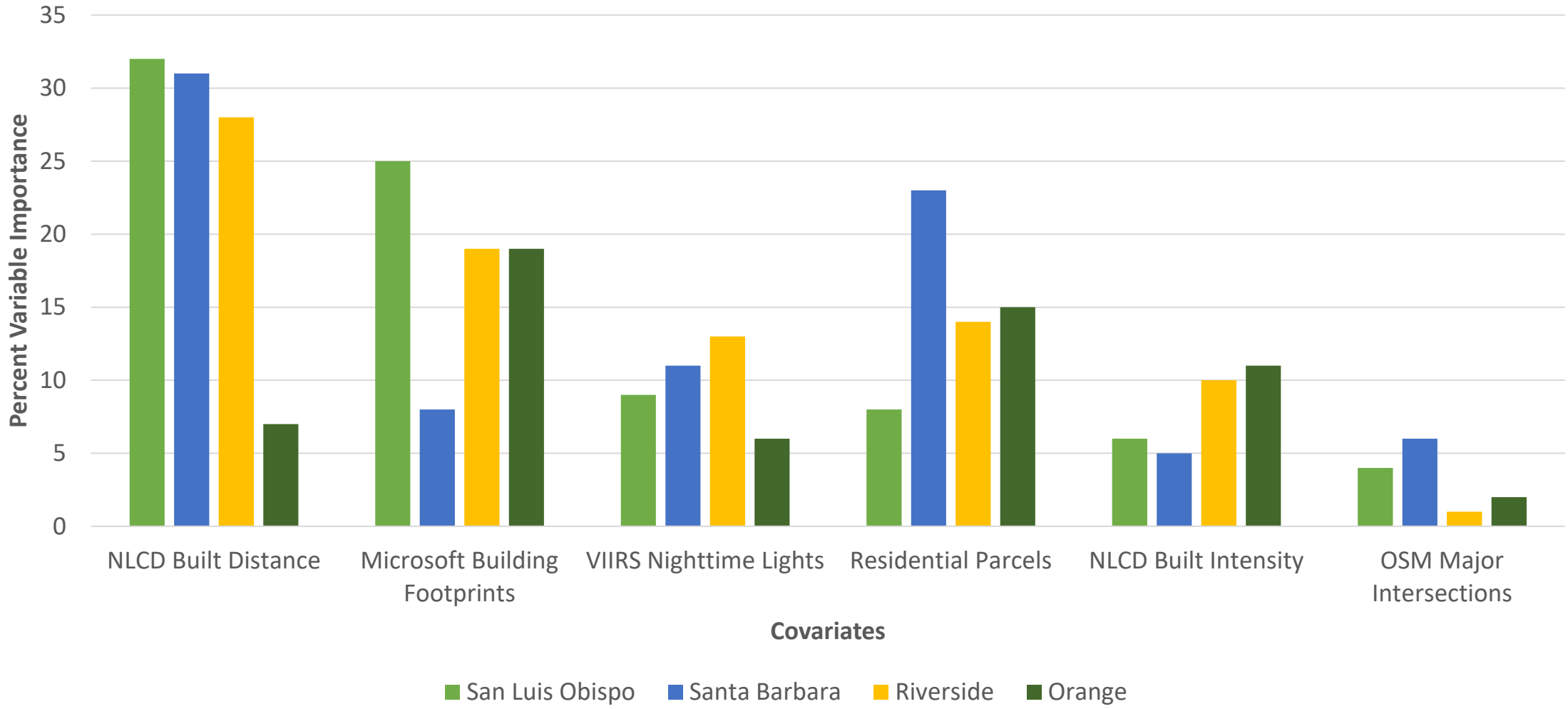
- 500 individual regression trees
- Terminal node observations = 1 or admin units / 1000
- 10% training data excluded for validation

Explanatory Covariates:

	Description	Data Source, Year	Nominal Resolution
Categorical	Water / Snow	NLCD, 2016	1" (30m)
	Urban Area		
	Bare Area		
	Tree Cover		
	Shrubland		
	Herbaceous Cover		
	Cropland		
	Floodland / Wetland		
	Urban Intensity		
Continuous Raster	Nighttime Lights	Suomi VIIRS Monthly Composite, 2019	15" (450m)
	Elevation	SRTM, 2000	3" (80m)
	Slope	SRTM, 2000	3" (80m)
	Mean Temperature	WorldClim, 1950-2000	30" (900m)
	Mean Precipitation	WorldClim, 1950-2000	30" (900m)
Converted Vector	Building Footprints	Microsoft Building Footprints, 2018	
	Protected Area	IUCN, 2020	
	Coastlines	Census TIGER Data, 2020	
	Distance to Major Highways	Census TIGER Data, 2020	
	Distance to Major Intersections	OSM, 2020	
	Waterways	OSM, 2020	
	Residential Parcels	ParcelQuest, 2021	
	Fixed Broadband	CA Broadband Mapping Program, 2019	
	Mobile Broadband	CA Broadband Mapping Program, 2019	

(Lloyd et al. 2019, Stevens et al. 2015, Liaw and Wiener 2002)

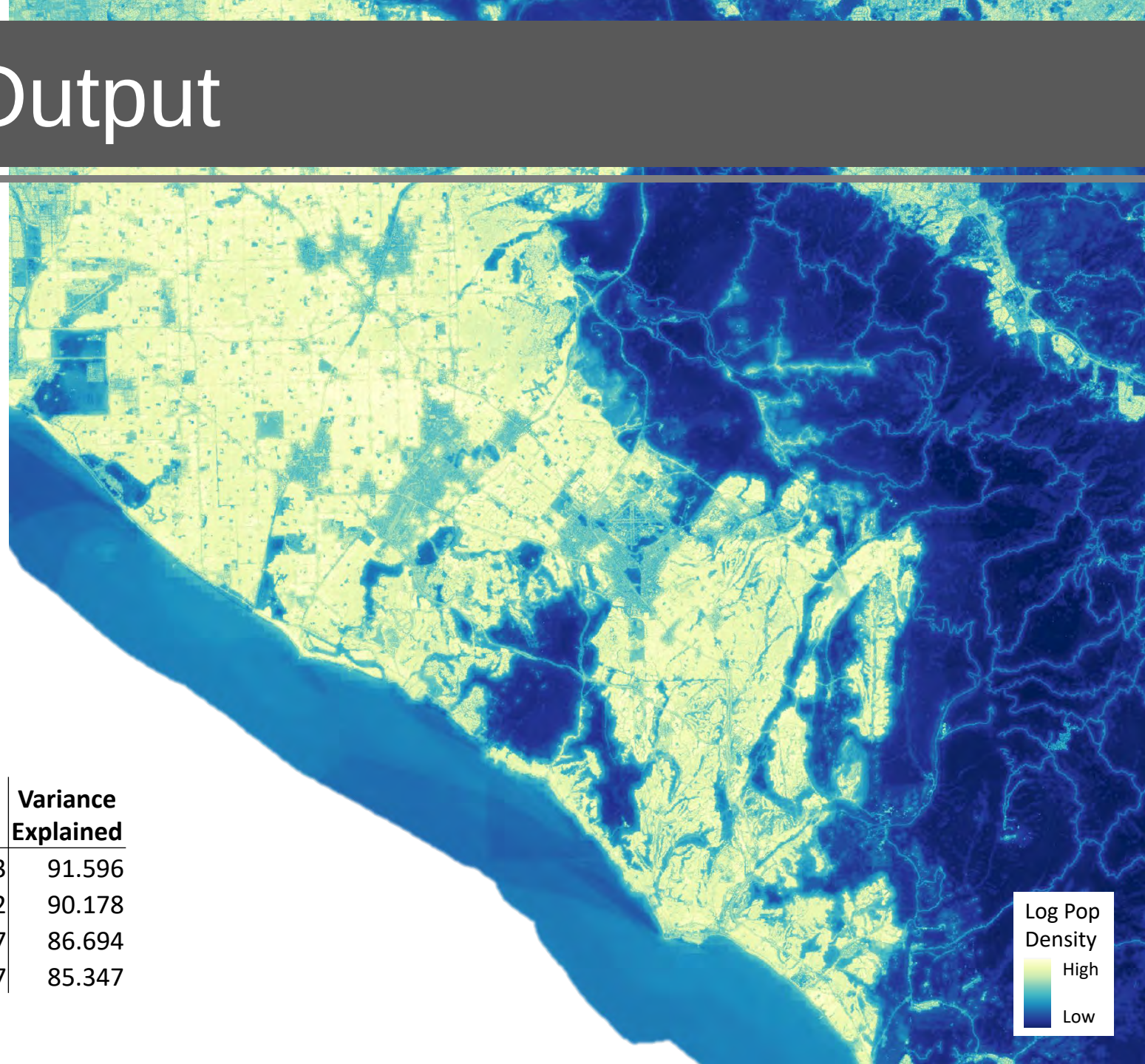
Variable Importance



(Liaw and Wiener 2002, Breiman 1996)

Random Forest Output

- 30m PPP raster
- 7-24hrs to complete



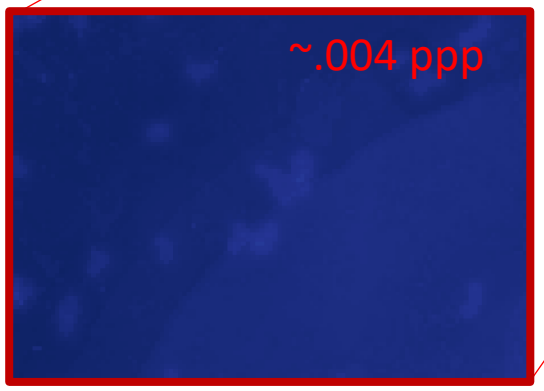
County	Training R^2	Training SE	Validation R^2	Validation SE	Variance Explained
San Luis Obispo	0.986	0.002	0.909	0.013	91.596
Santa Barbara	0.984	0.002	0.904	0.012	90.178
Riverside	0.951	0.001	0.86	0.007	86.694
Orange	0.946	0.001	0.855	0.007	85.347

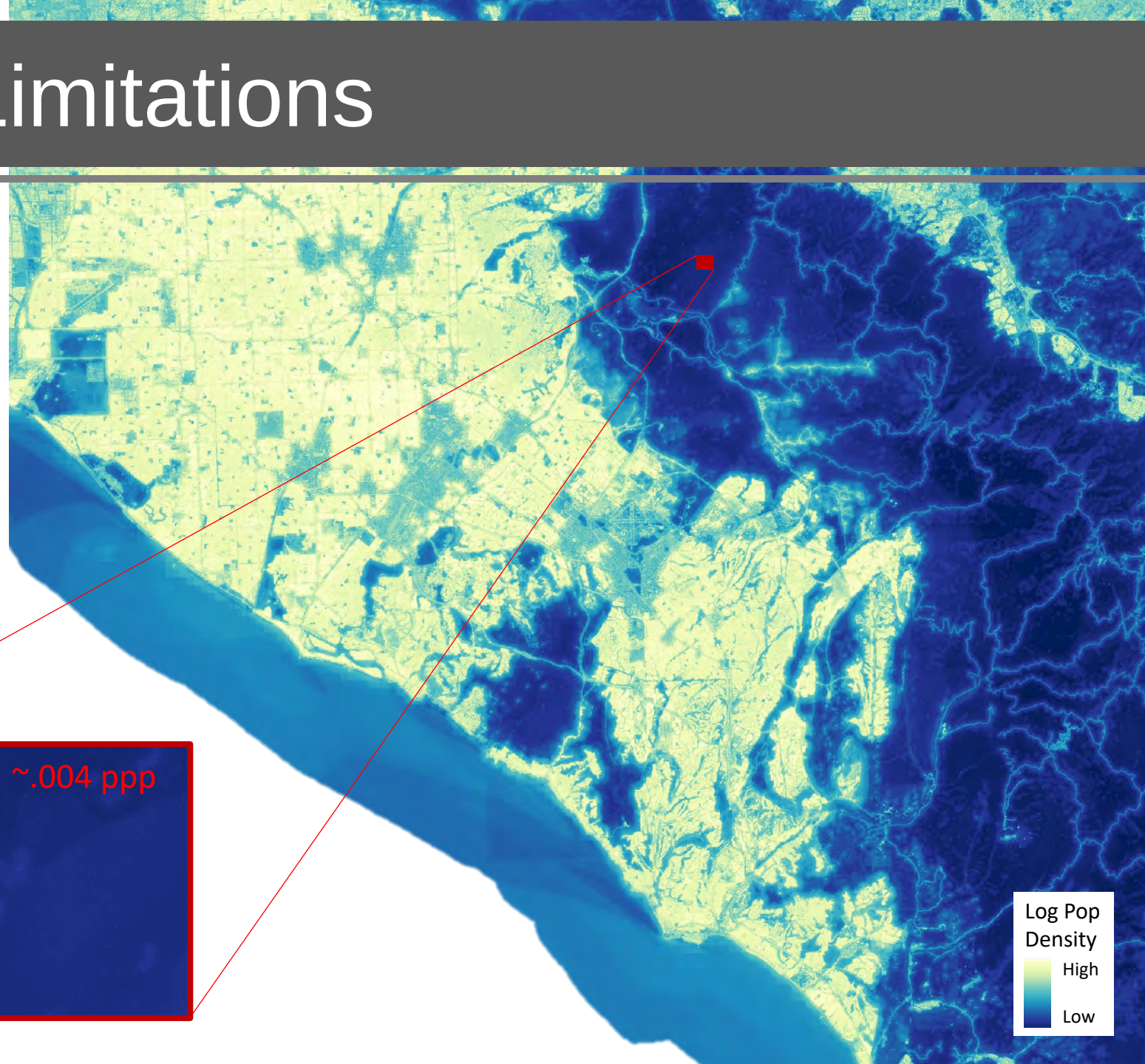
Random Forest Limitations

- Interpolation not extrapolation

- Overestimate Rural
- Underestimate Urban

- Insufficient alone





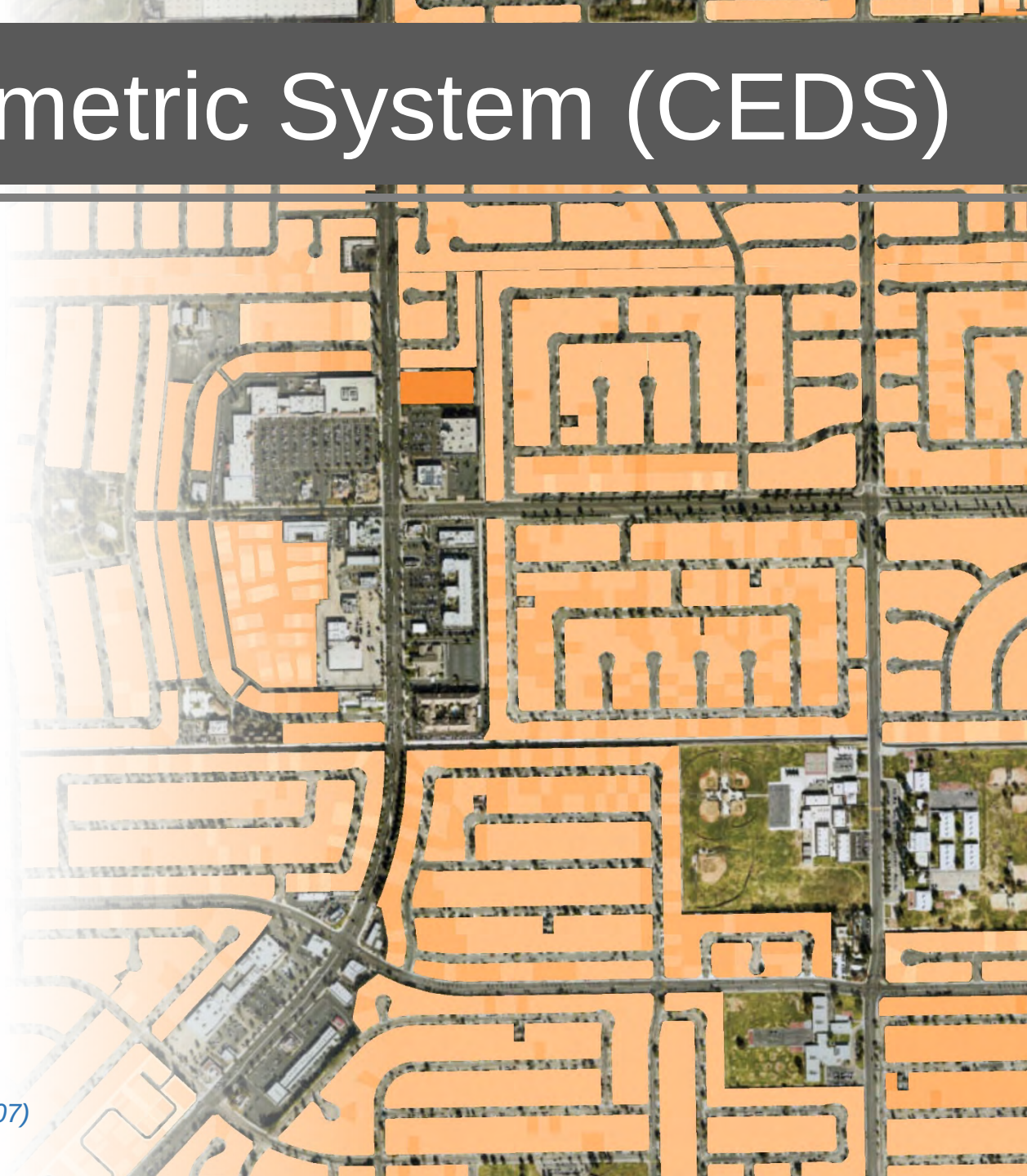
Cadastral Expert Dasymetric System (CEDS)

Original:

- Top-down population model
  - Residential area
  - Residential units
- Selects minimal difference
- Applies to larger unit

Adaptation:

- New CEDS candidates*
- Group quarters
- Building footprints
- Vacancy



**Variants from ParcelQuest, Pitney Bowes, and harmonization*

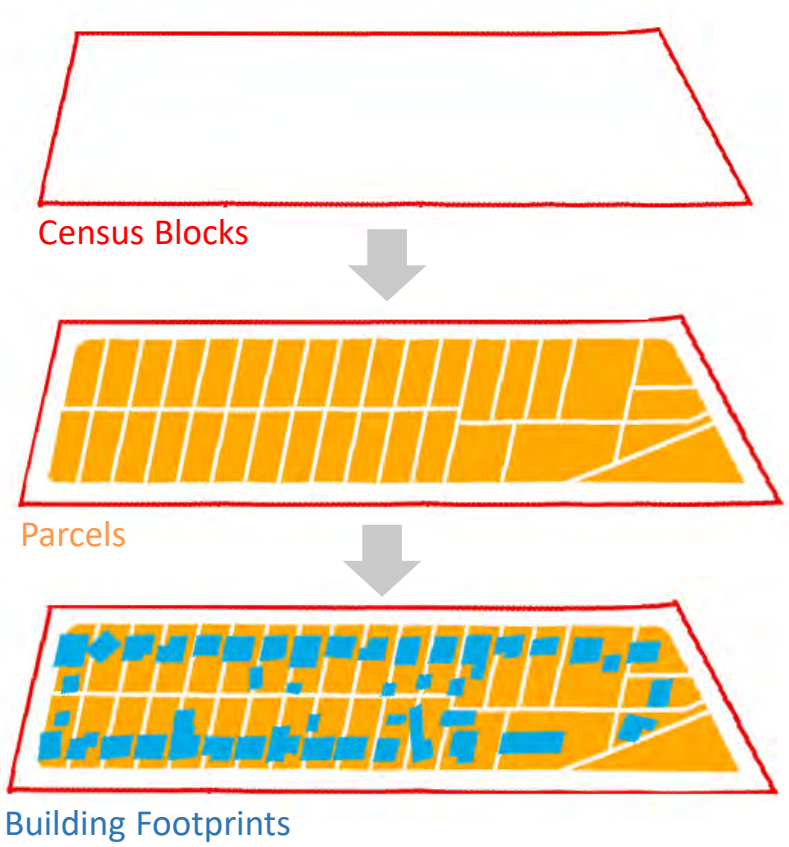
(Strode, G., V. Mesev, J. Maantay. 2018, Maantay, J.A., A.R. Marko, C. Hermann. 2007)

Parcel + Footprint Repairs

Assumption:
County -> Blocks -> Parcels -> Buildings

Problem:
Incongruent geometry

Expectation:



Reality:

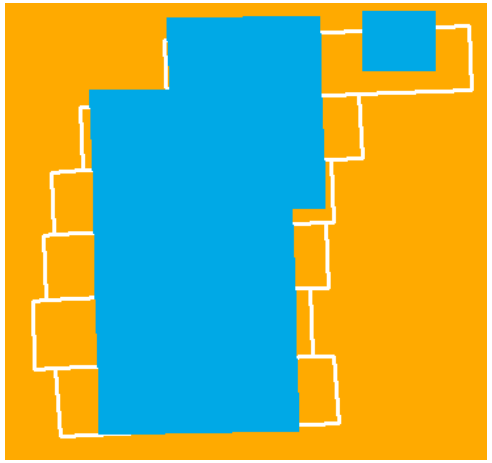
Parcel + Footprint Repairs

Assumption:
County -> Blocks -> Parcels -> Buildings

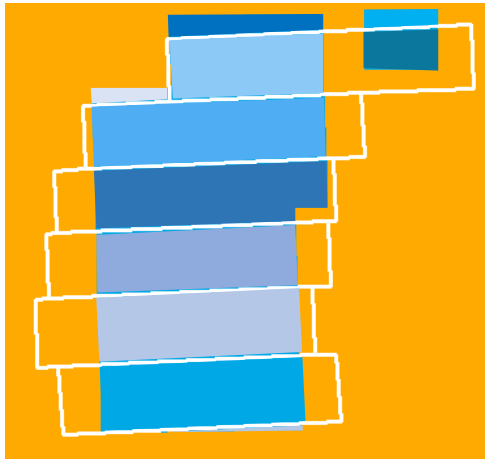
Problem:
Incongruent geometry

Solution:
- Union
- Dissolve threshold

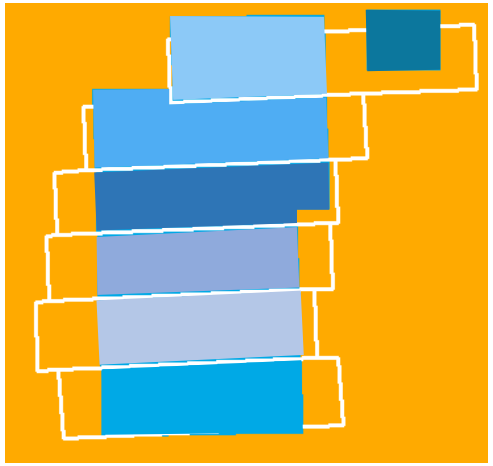
Building Footprints + Parcels



Union



Dissolve



Improvements:

Parcels

LA County Data Portal

- 2014
- Inconsistent
- Overlap
- Data gaps
- No attribution

Parcel Quest

- 2021
- Consistent
- No overlap
- No data gaps
- Abundant attribution

Improvements:

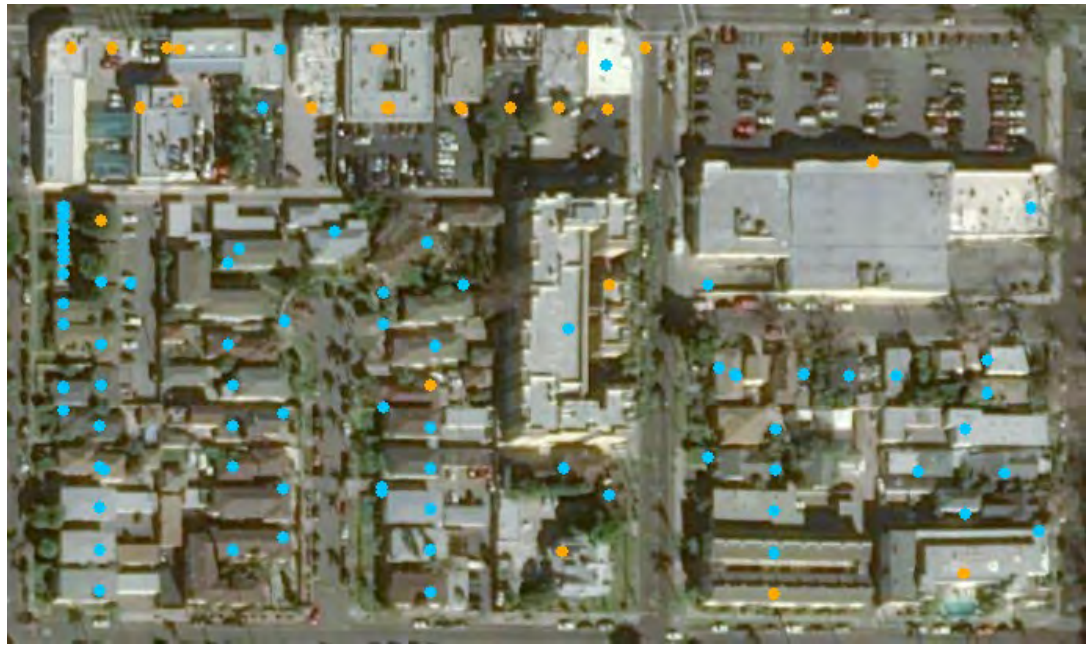
Residential Definition

Pitney Bowes

- Address points
- Location
- Duplicates

Parcel Quest

- Residential use codes
- Backfilled with PB
- Utility



Residential
Non-Residential

Improvements:

Addresses

Parcel Quest

- County variance
- Inaccurate use codes
- Backfill with PB by APN

Improvements:

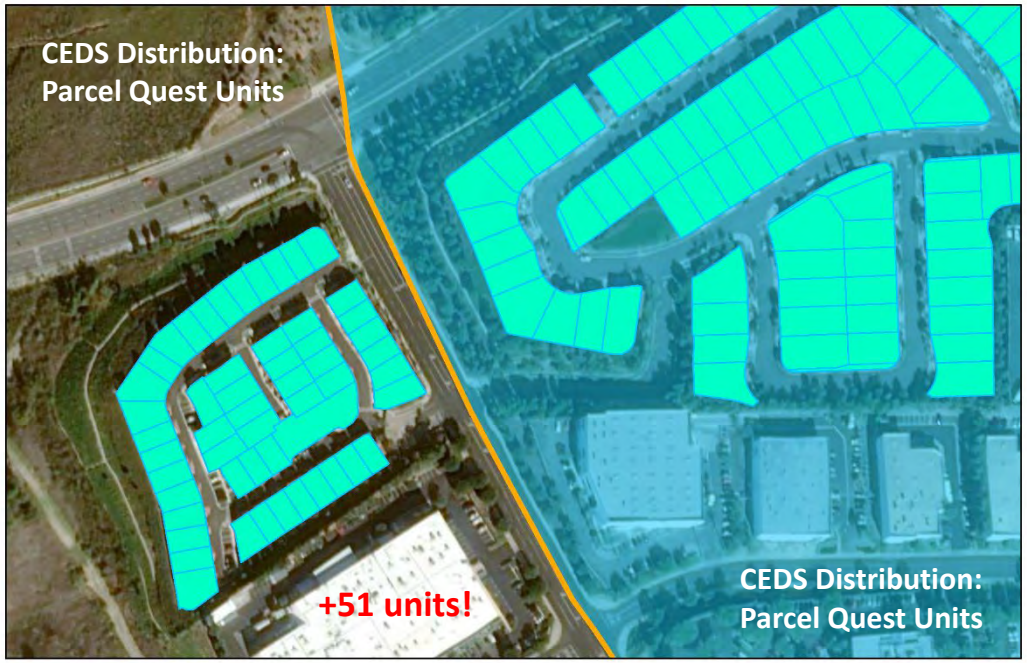
Multifamily Structures

- Trailer parks, apartments, campuses, sub divisions
- Pitney Bowes allocation:
  - Shared border, use, or APN
  - Building footprints
- ~5% have no Pitney Bowes

Improvements:

New Structures

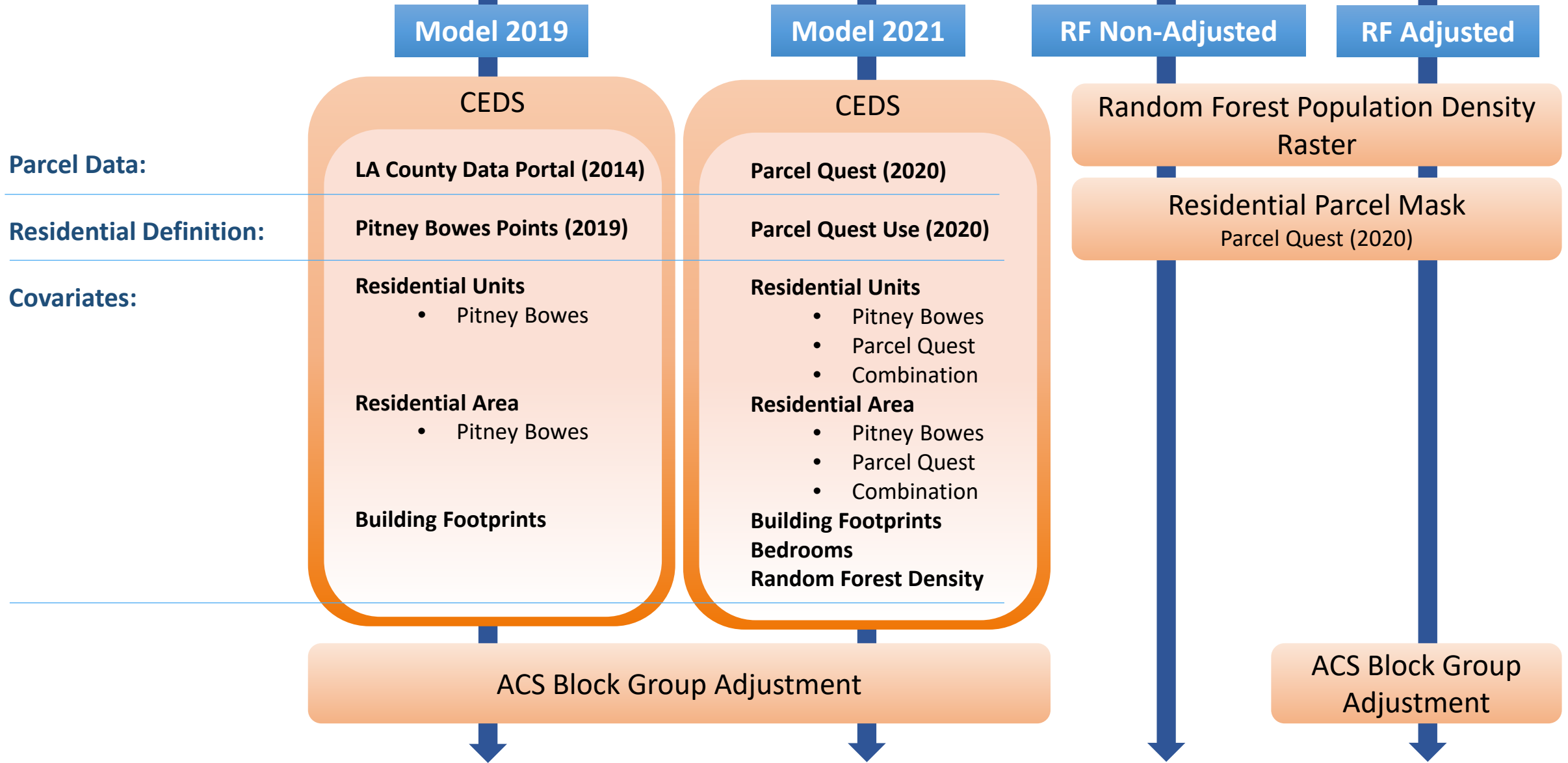
- Outside census inhabited range
- Apply neighboring distribution method



ACS Block Group
- HU = 0
- HU > 0

Residential Parcels
- HU = 0
- HU > 0

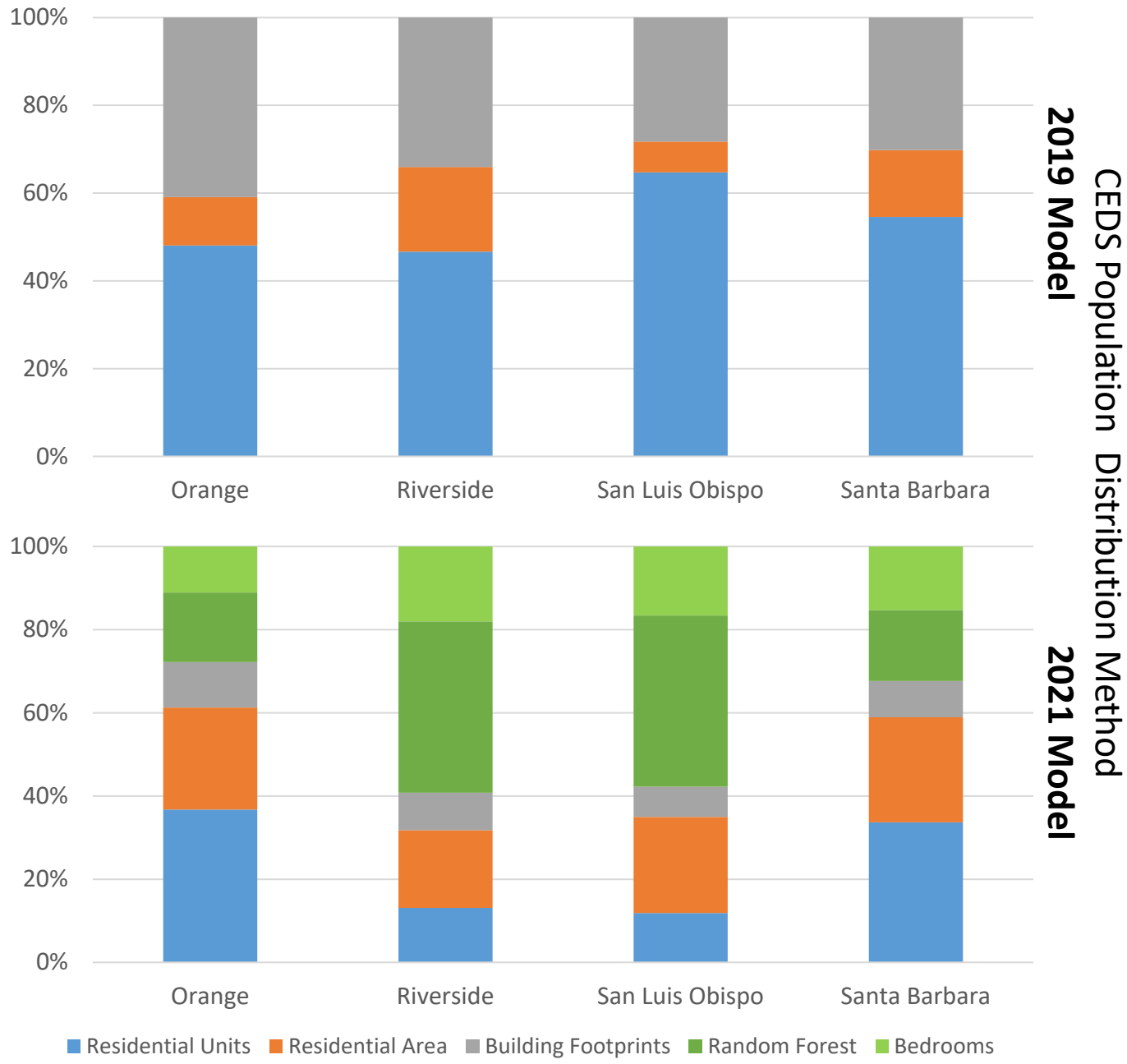
Models Compared



Results:

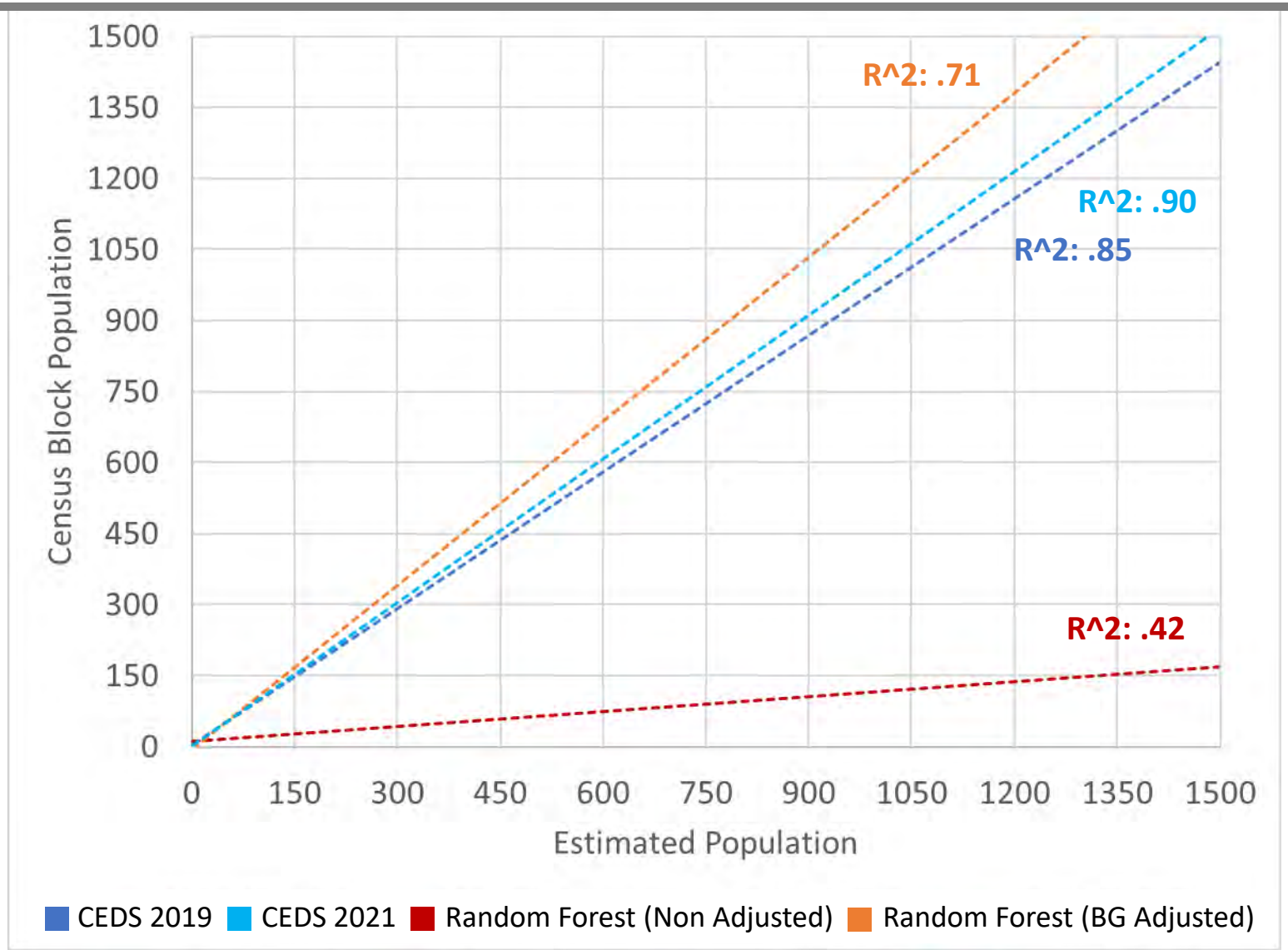
CEDS Distribution

- Residential Units favored
- Random forest used ~ 29%



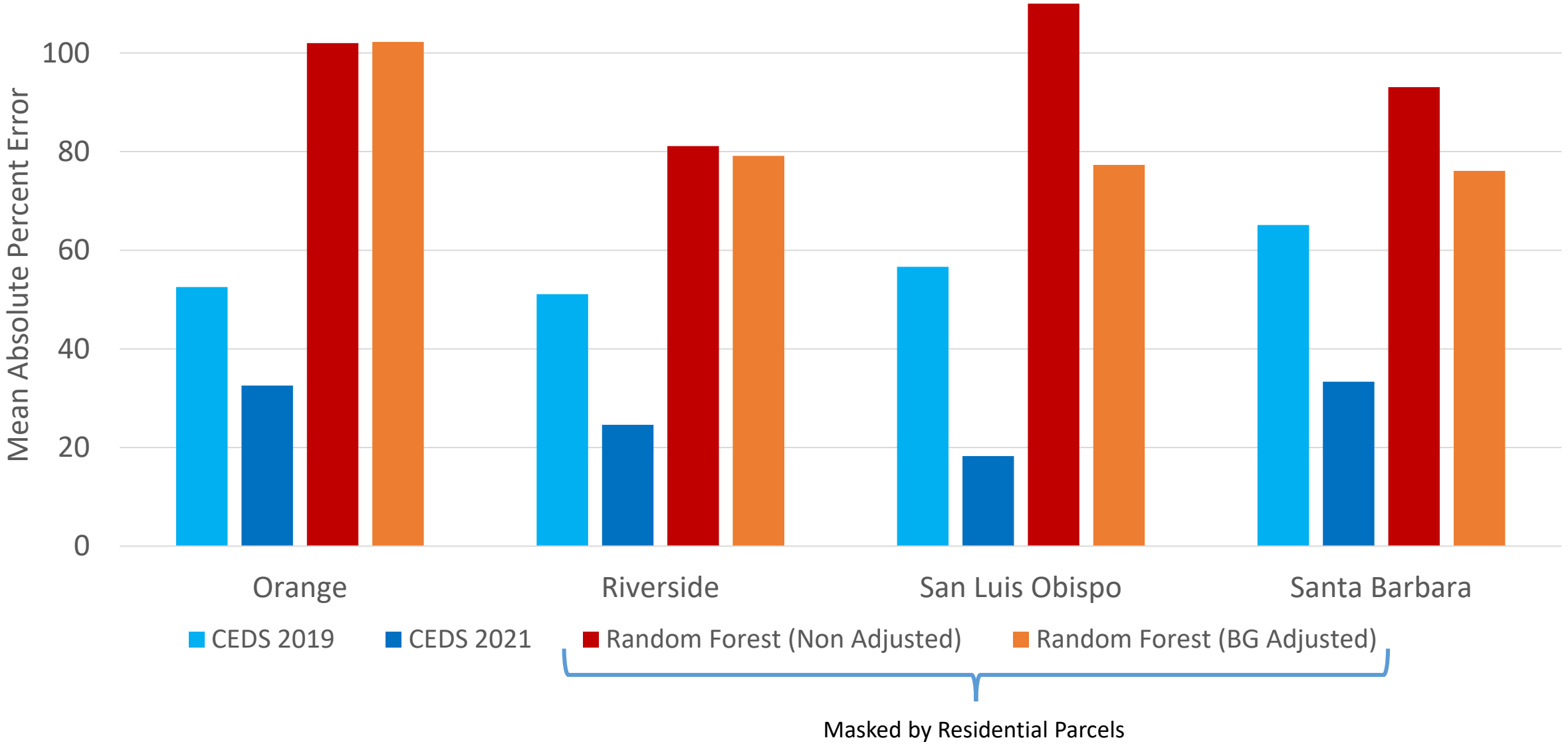
Results:

Scatterplots



Results:

MAPE

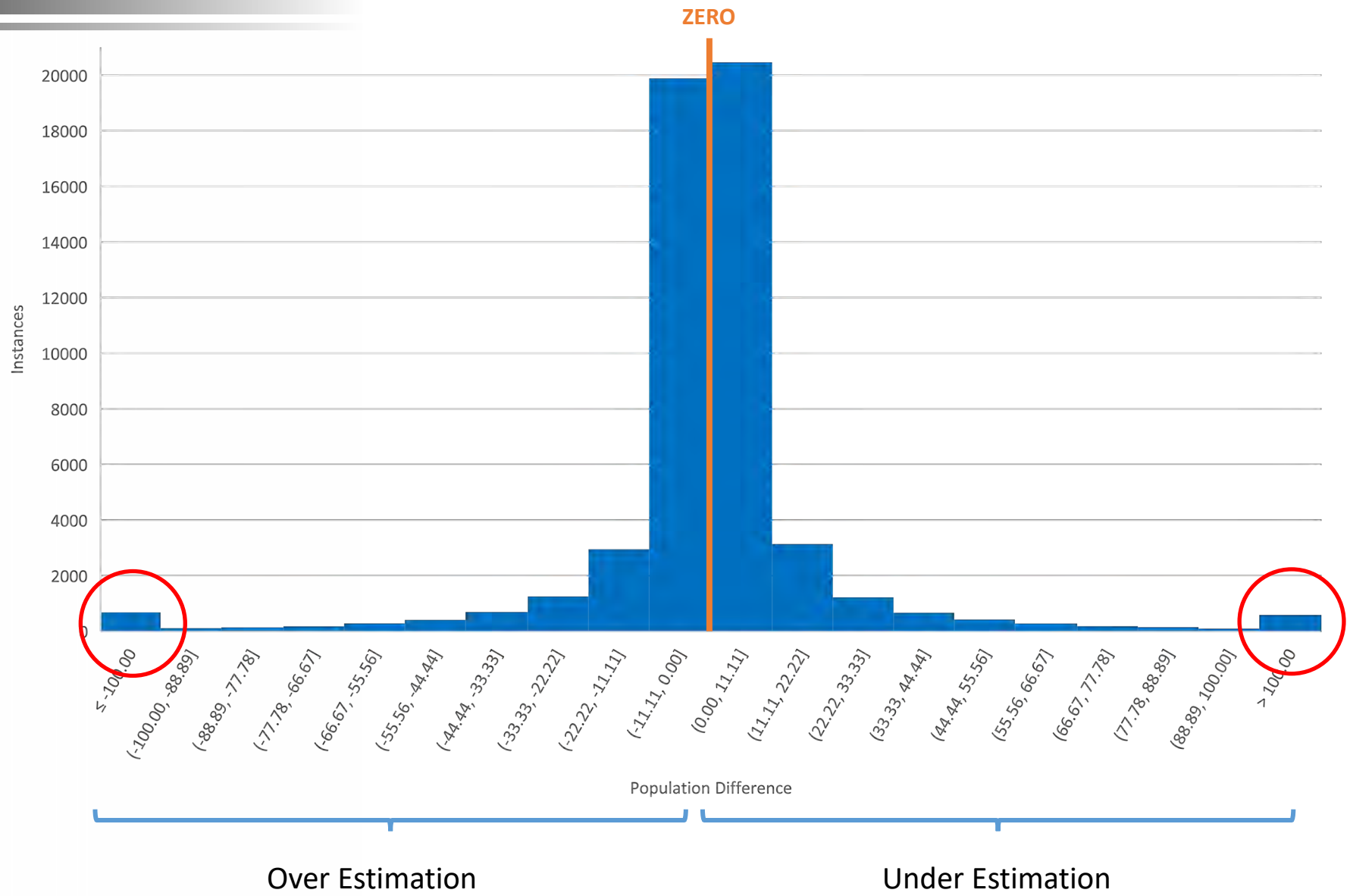


Results:

Error Histogram

Where does outlying error occur?

- Misplaced PB pts
- Misallocated census units
- New Construction



Case Study:

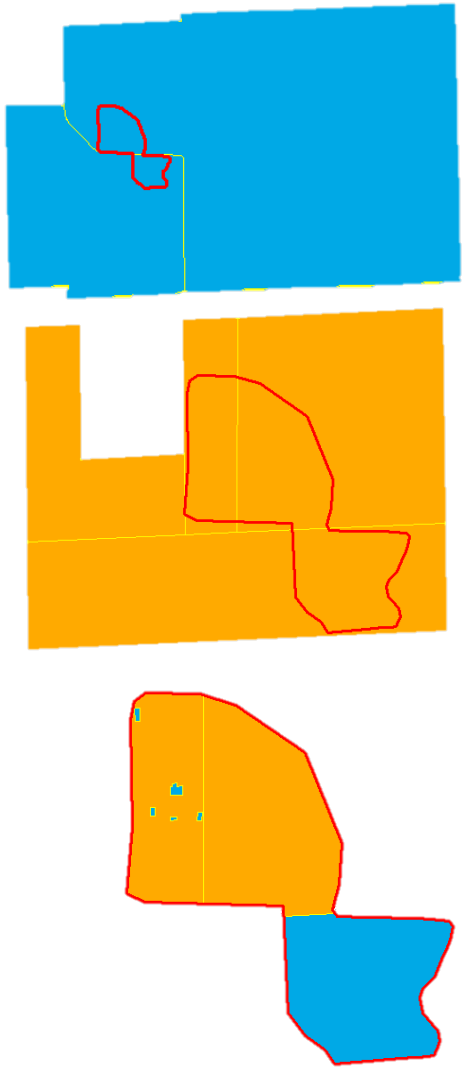
Martinez Fire, 2018

- 7 units affected

- 86 by 2010 Blocks

- County specific codes
- 32.7 by SAE parcels

- 7.15 by SAE building footprints



Populated
Un-populated

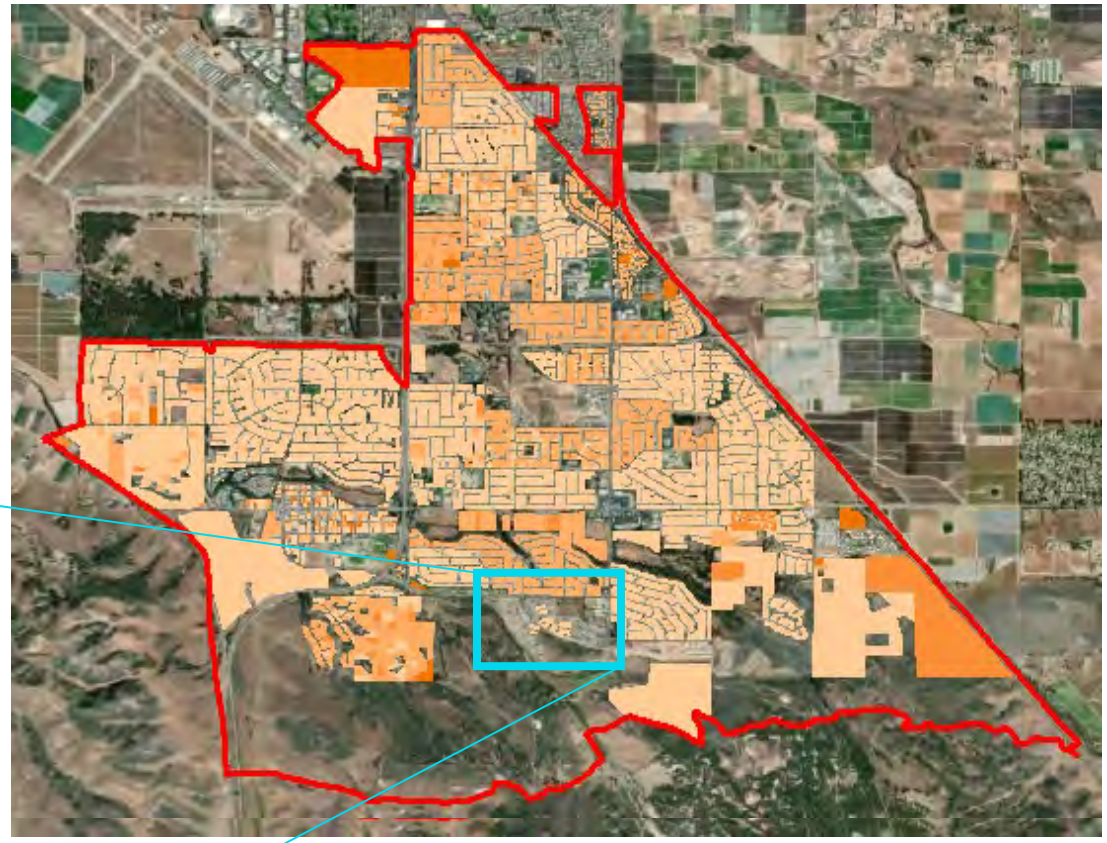
(Riverside County FD, incident CA-RRU-92674)

Case Study:

Orcutt Library District

Annual estimates for Santa Barbara libraries

- Low estimate = 30077
- Recent construction



Case Study:

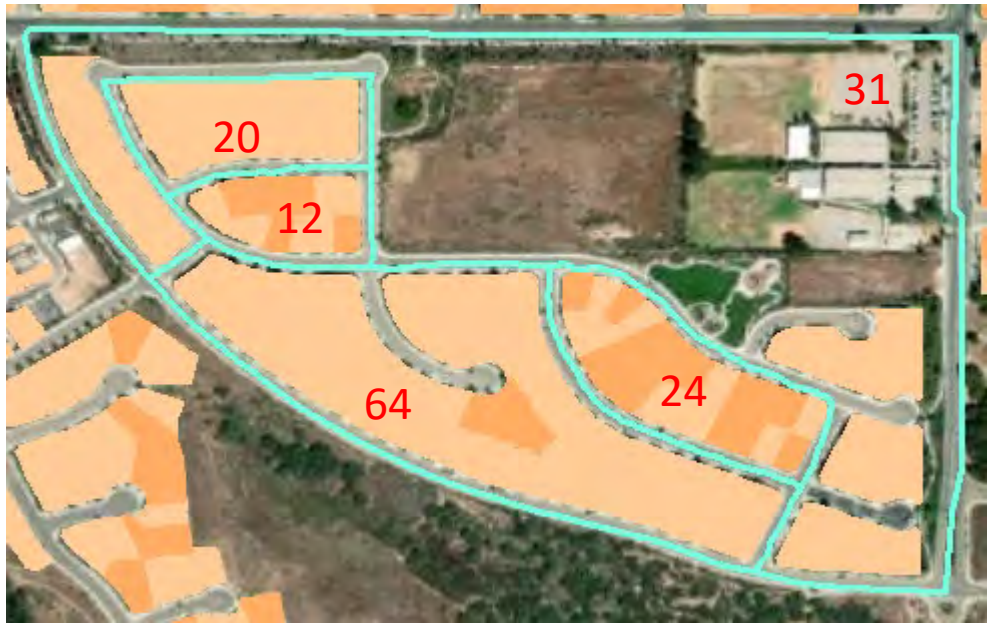
Orcutt Library District

Annual estimates for Santa Barbara libraries

- Low estimate = 30077
- Recent construction

- Revised estimate = 31001

2019 Model



2021 Model

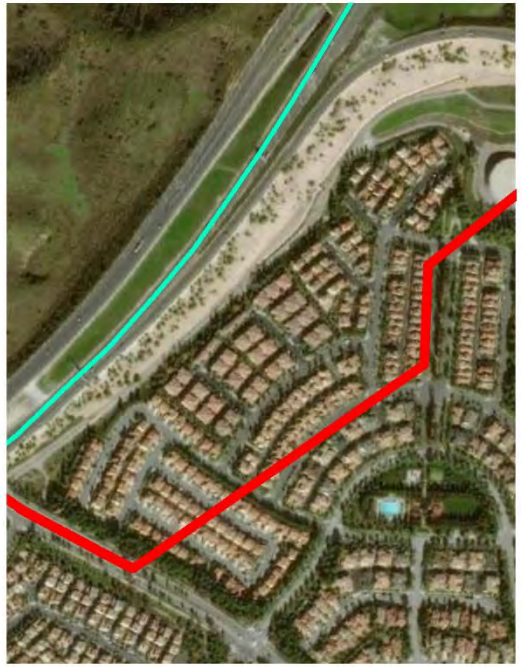
County Boundary
Block Boundary
1
Estimated housing units
8+

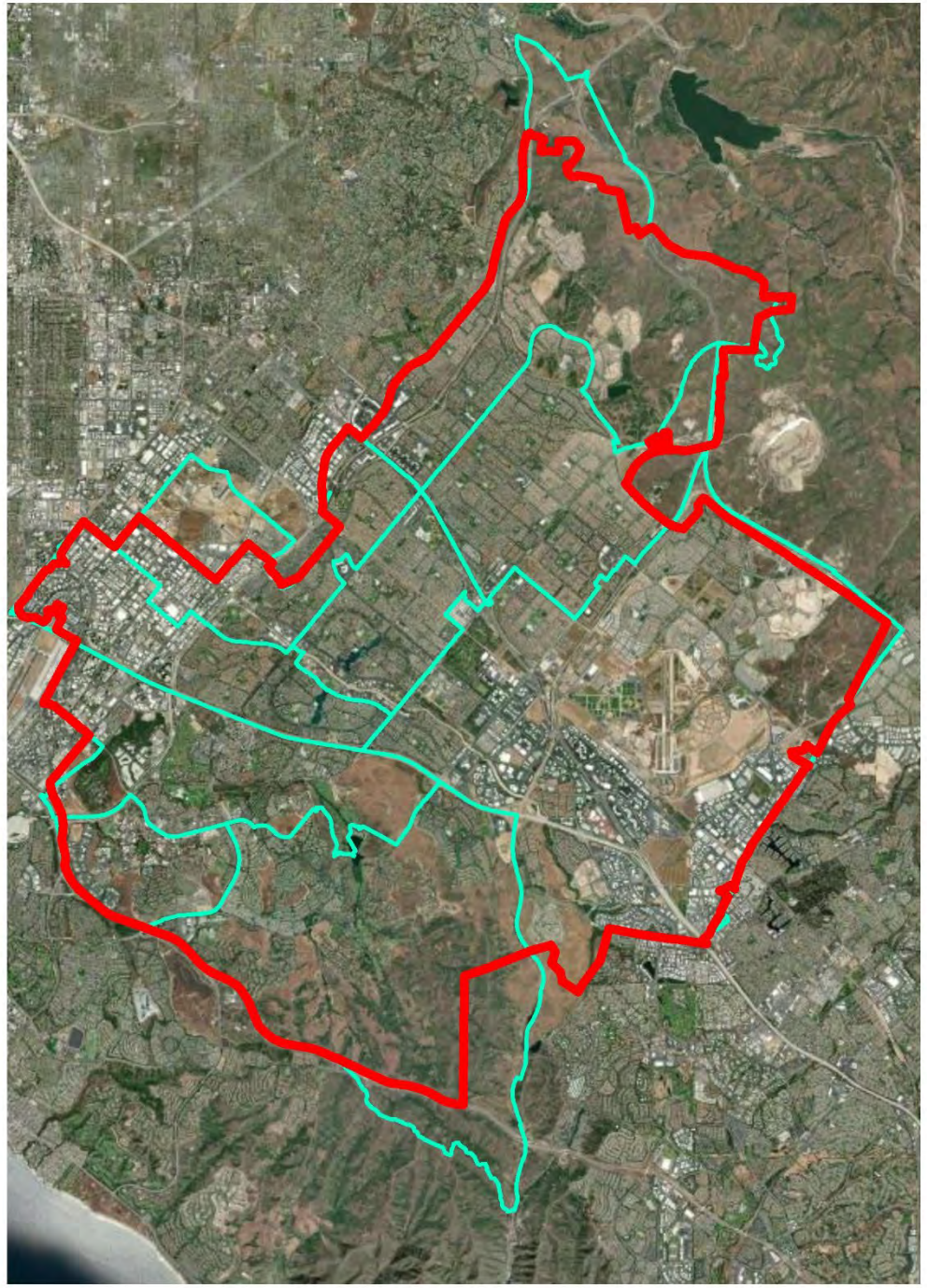
Case Study:

City of Irvine

Request for population by ZIP

- ACS ZCTA = 277,141





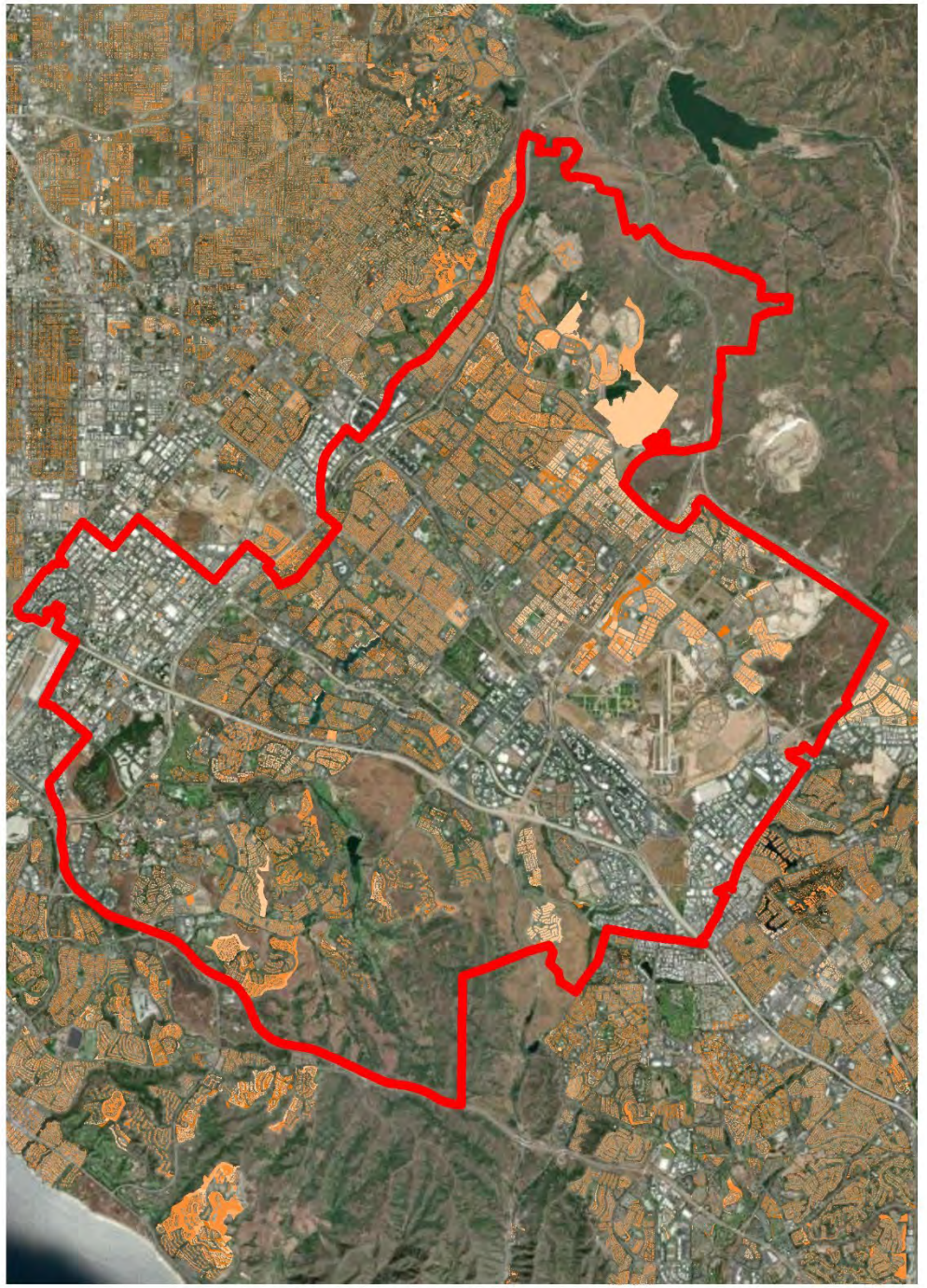
City Boundary
ZCTA Boundary

Case Study:

City of Irvine

Request for population by ZIP

- ACS ZCTA = 277,141
- Revised ACS ZCTA = 273,863



City Boundary

1

Estimated population

8+

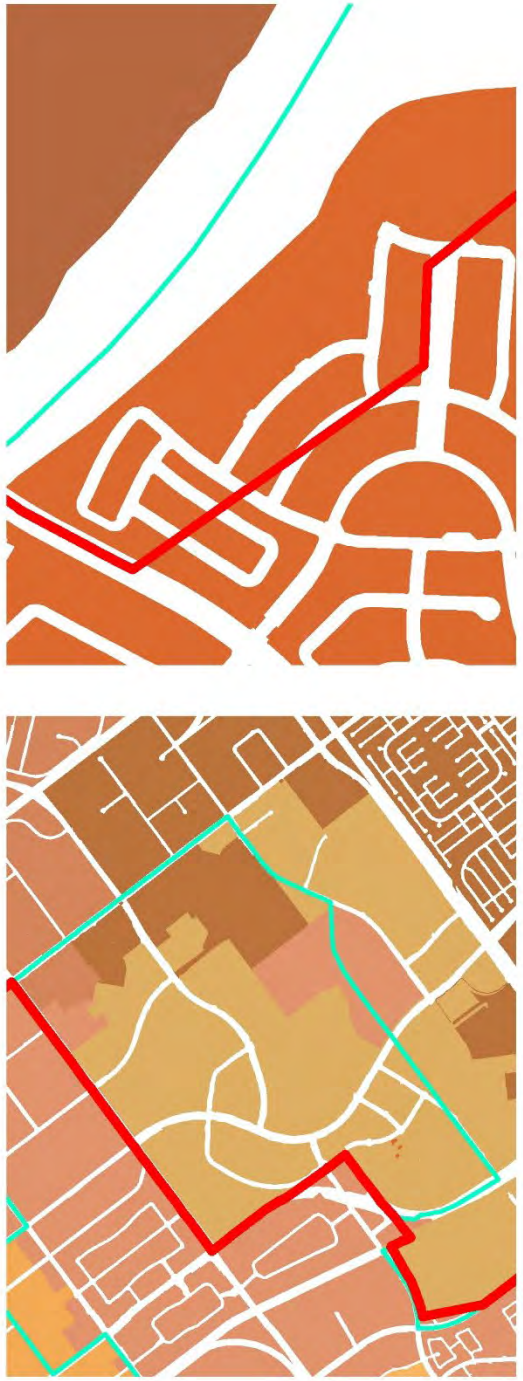
Case Study:

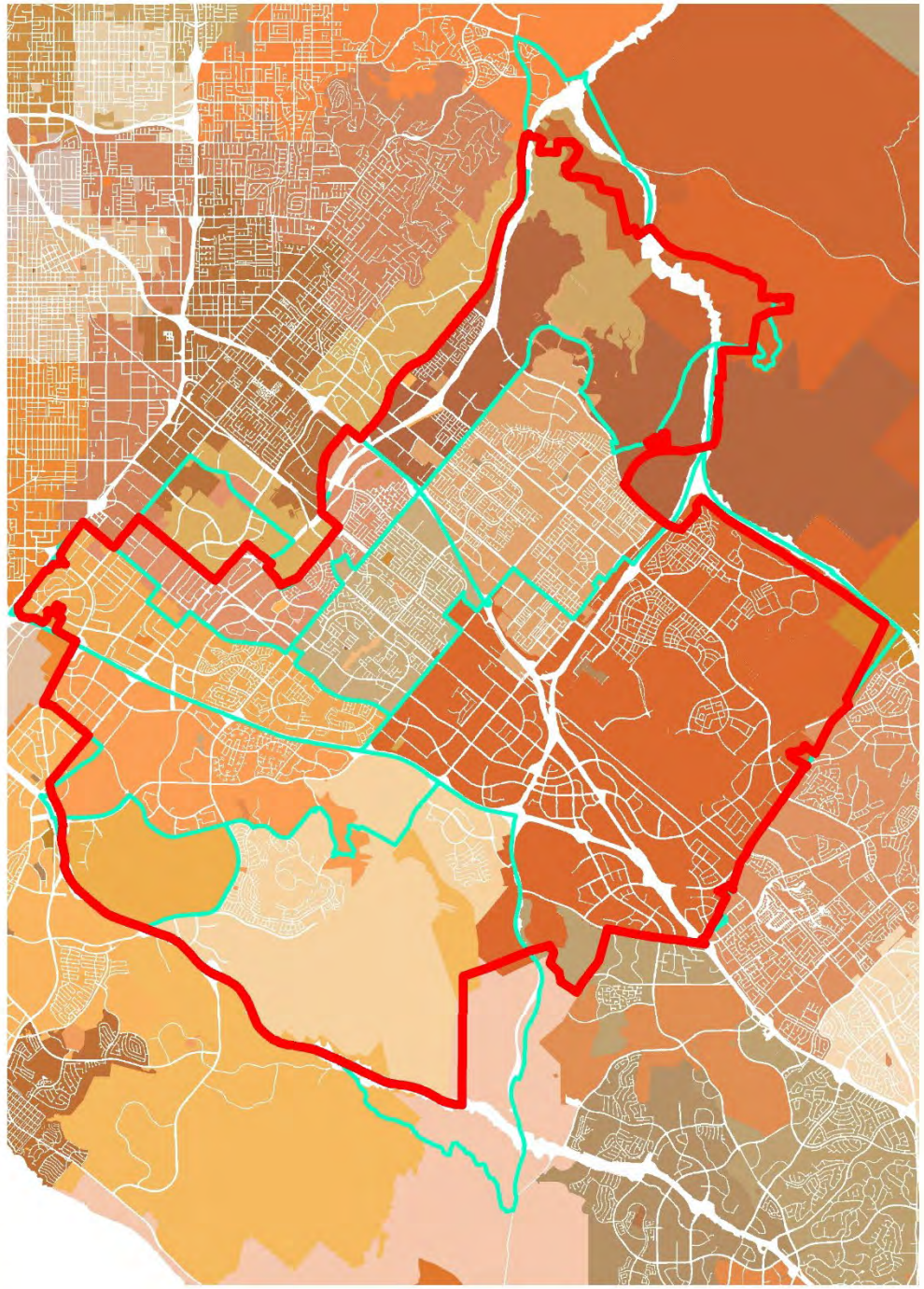
City of Irvine

Request for population by ZIP

- ACS ZCTA = 277,141
- Revised ACS ZCTA = 273,863

Sum by ZIP is possible!





City Boundary
ZCTA Boundary

Case Study:

Santa Barbara Library Districts

Used in appropriation limits

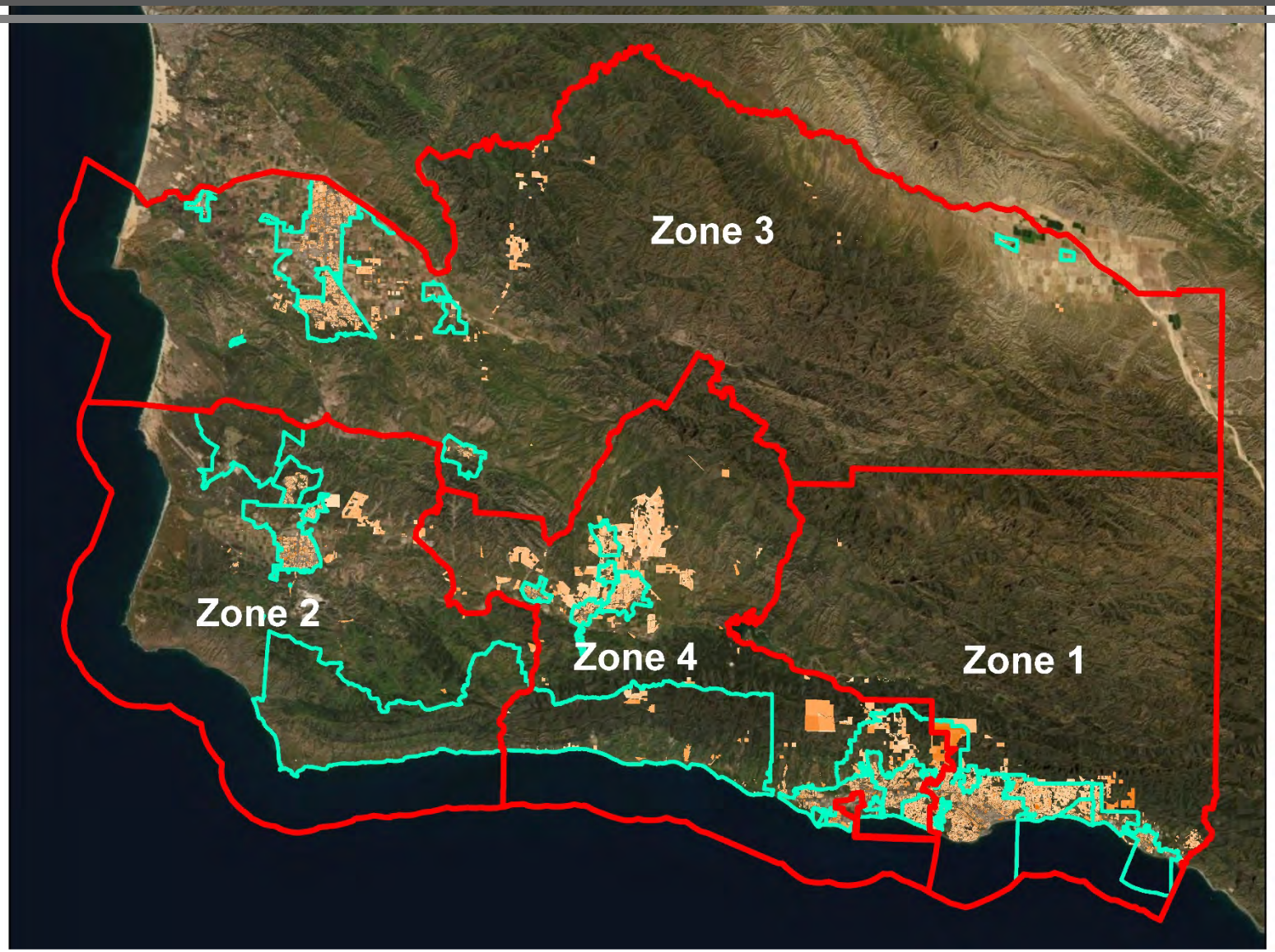
State: 4 zones

County: sub districts

E5 Estimate: 441,172

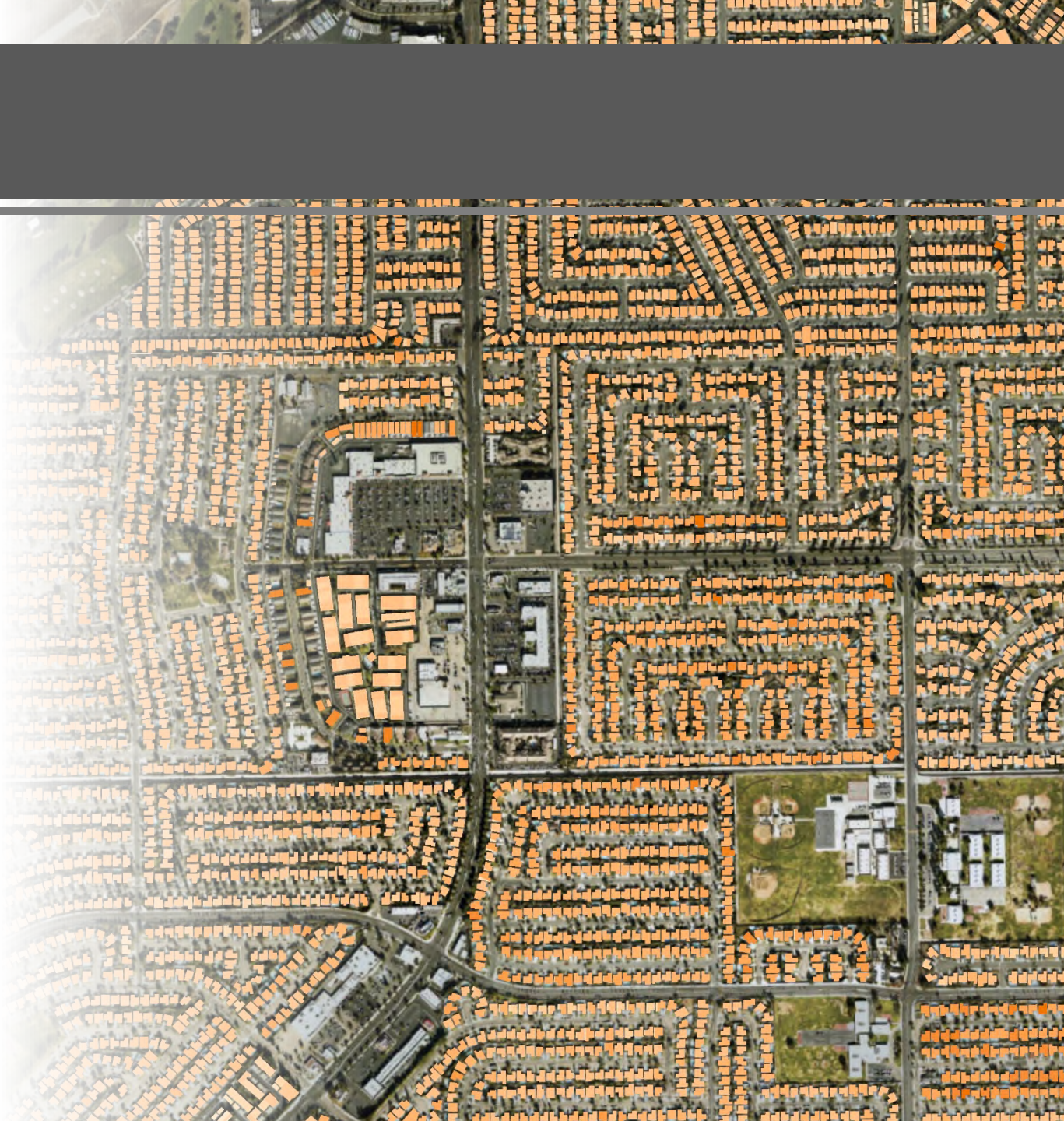
- Zone 1: 123,829
- Zone 2: 55,656
- Zone 3: 158,581
- Zone 4: 103,106

Expansive Reporting



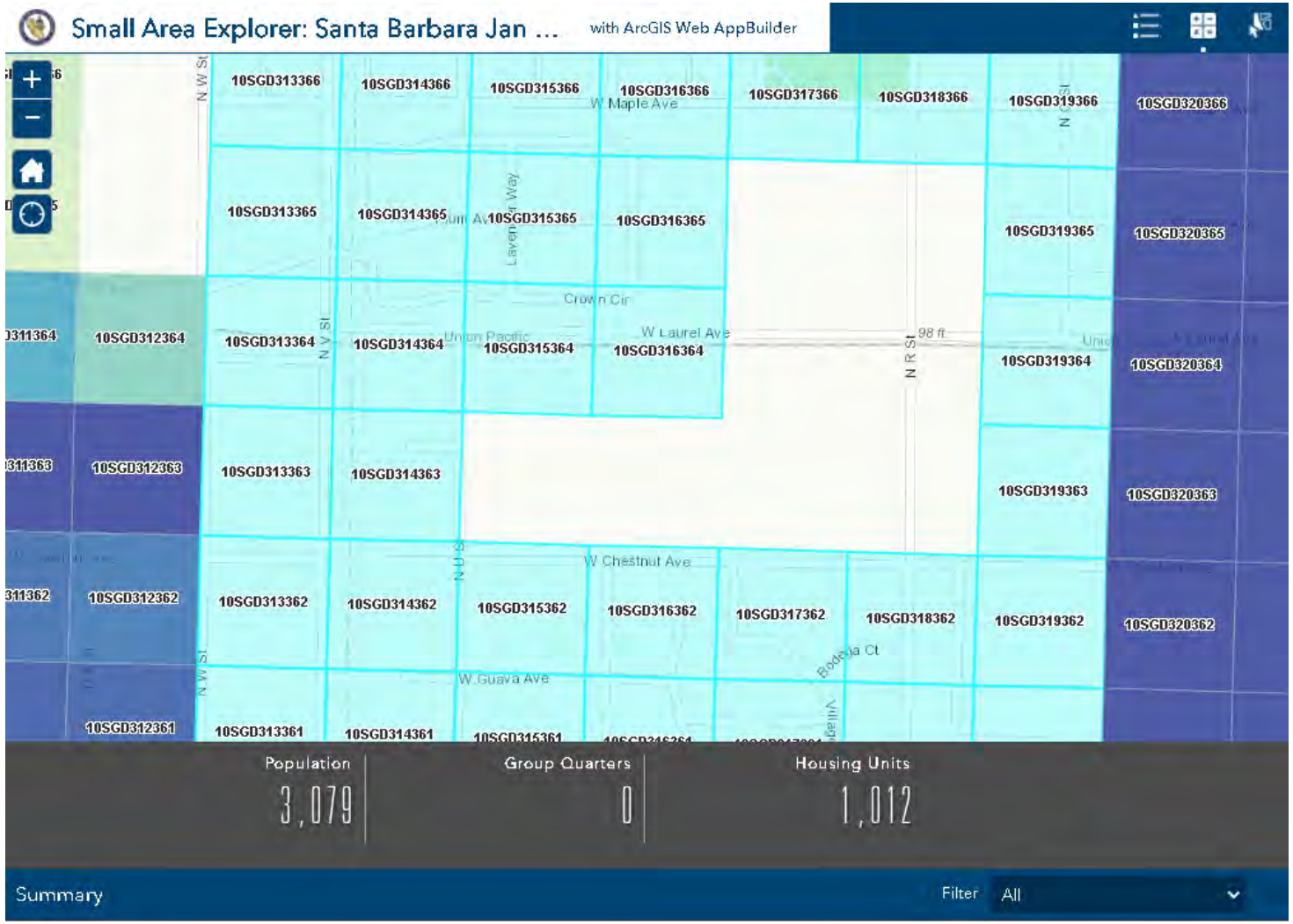
Conclusion

- RF comparable to CEDS when masked
- Best estimates combine CEDS and RF methods
- Can always improve with county specific data



Future Directions

- Web app
- CQR
- CA Neighborhoods Count



Discussion

Fennis Reed
Research Specialist
Demographic Research Unit
CA Department of Finance

CONTACT:
(916) 323-4086
fennis.reed@dof.ca.gov