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

CSA

E5 City / County

Census Blocks

Parcels
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

\[ \text{Population} = HU \times PPH + GQ \]

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

• Requires consistent geographic coverage
• PPH & MAUP

• Random Forest + CEDS

Study Area

2010 Census Block Population

1
1000+
Random Forest: What is it?

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

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

(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

<table>
<thead>
<tr>
<th>County</th>
<th>Training R^2</th>
<th>Training SE</th>
<th>Validation R^2</th>
<th>Validation SE</th>
<th>Variance Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Luis Obispo</td>
<td>0.986</td>
<td>0.002</td>
<td>0.909</td>
<td>0.013</td>
<td>91.596</td>
</tr>
<tr>
<td>Santa Barbara</td>
<td>0.984</td>
<td>0.002</td>
<td>0.904</td>
<td>0.012</td>
<td>90.178</td>
</tr>
<tr>
<td>Riverside</td>
<td>0.951</td>
<td>0.001</td>
<td>0.86</td>
<td>0.007</td>
<td>86.694</td>
</tr>
<tr>
<td>Orange</td>
<td>0.946</td>
<td>0.001</td>
<td>0.855</td>
<td>0.007</td>
<td>85.347</td>
</tr>
</tbody>
</table>
Random Forest Limitations

- Interpolation not extrapolation
- Overestimate Rural
- Underestimate Urban
- Insufficient alone

\[ \sim 0.004 \text{ ppp} \]
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
Assumption:
County -> Blocks -> Parcels -> Buildings

Problem:
Incongruent geometry

Expectation:
Census Blocks → Parcels → Building Footprints

Reality:
Parcel + Footprint Repairs
Assumption:
County -> Blocks -> Parcels -> Buildings

Problem:
Incongruent geometry

Solution:
• Union
• Dissolve threshold
Improvements:

Parcels

LA County Data Portal
• 2014
• Inconsistent
• Overlap
• Data gaps
• No attribution

Parcel Quest
• 2021
• Consistent
• No overlap
• No data gaps
• Abundant attribution
Residential Definition

Improvements:

**Pitney Bowes**
- Address points
- Location
- Duplicates

**Parcel Quest**
- Residential use codes
- Backfilled with PB
- Utility
Improvements:

Addresses

Parcel Quest
- County variance
- Inaccurate use codes
- Backfill with PB by APN
• Trailer parks, apartments, campuses, sub divisions

• Pitney Bowes allocation:
  • Shared border, use, or APN
  • Building footprints

• ~5% have no Pitney Bowes
New Structures

Improvements:

- Outside census inhabited range
- Apply neighboring distribution method

ACS Block Group

- HU = 0
- HU > 0

Residential Parcels

- HU = 0
- HU > 0

CEDS Distribution:

- None
- Parcel Quest Units

+51 units!
Random Forest Population Density Raster

ACS Block Group Adjustment

RF Non-Adjusted

Random Forest Population Density Raster

Residential Parcel Mask Parcel Quest (2020)

ACS Block Group Adjustment

RF Adjusted

Residential Parcel Mask Parcel Quest (2020)

Parcel Data:

LA County Data Portal (2014)

Pitney Bowes Points (2019)

Residential Definition:

Residential Units
  - Pitney Bowes

Residential Area
  - Pitney Bowes

Covariates:

Building Footprints

Residential Units
  - Pitney Bowes

Residential Area
  - Pitney Bowes

Bedrooms

Random Forest Density

Model 2019

CEDS

Model 2021

CEDS

Parcels Quest (2020)

Parcels Quest Use (2020)

Residential Units
  - Pitney Bowes
  - Parcel Quest
  - Combination

Residential Area
  - Pitney Bowes
  - Parcel Quest
  - Combination

Building Footprints

Bedrooms

Random Forest Density

Models Compared
CEDS Distribution

• Residential Units favored
• Random forest used ~ 29%
Results:

Scatterplots
Results:

**MAPE**

<table>
<thead>
<tr>
<th>Location</th>
<th>CEDS 2019</th>
<th>CEDS 2021</th>
<th>Random Forest (Non Adjusted)</th>
<th>Random Forest (BG Adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orange</td>
<td>50</td>
<td>60</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>Riverside</td>
<td>40</td>
<td>50</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>San Luis Obispo</td>
<td>30</td>
<td>40</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>Santa Barbara</td>
<td>20</td>
<td>30</td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>
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

(Riverside County FD, incident CA-RRU-92674)
Case Study: Orcutt Library District

Annual estimates for Santa Barbara libraries
- Low estimate = 30077
- Recent construction
Annual estimates for Santa Barbara libraries

- Low estimate = 30077
- Recent construction

- Revised estimate = 31001
Case Study: City of Irvine

Request for population by ZIP

- ACS ZCTA = 277,141
Case Study:
City of Irvine

Request for population by ZIP

- ACS ZCTA = 277,141
- Revised ACS ZCTA = 273,863
Request for population by ZIP

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

Sum by ZIP is possible!
Santa Barbara Library Districts

Case Study:

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