Using Big Data to Improve Sample Efficiency

BigSurv18

Barcelona, Spain

27 October 2018

Jamie Ridenhour, Joe McMichael, Karol Krotki, Howard Speizer

RTI International is a registered trademark and a trade name of Research Triangle Institute.

Where We're Going

- General framework: incomplete list for target population
  - Target population is somewhat rare
- Append target population flags to "big data"
  - Address frame in the United States
  - Large sample from incomplete list
- Use auxiliary data to build model for identifying target pop
  - Census data
  - Consumer/marketing data
- Apply model to "big data" -> identify more likely members of target pop not on list
- Sample!

Motivation

- National Recreation Boating Survey
  - conducted on behalf of the United States Coast Guard
- Measure recreational boating activity to develop water safety exposure estimates
  - All 50 States and District of Columbia
- Target population: boat-owning households

What is a boat?


PANTHER


CT 1435 BJ 17

Motivation

- National Recreation Boating Survey
  - conducted on behalf of the United States Coast Guard
- Measure recreational boating activity to develop water safety exposure estimates
  - All 50 States and District of Columbia
- Target population: boat-owning households
- Inclusive of
  - Registered boats, *varies by state* (power boats, sail boats, pontoon boats)
  - Unregistered boats (canoes, kayaks, paddleboards)
- Survey modes: push-to-web and paper (contacts are by mail)

Frames Suitable for the Target Population

- Option 1: States maintain lists of boat registrations
  - Limitations on how data can be used

Frames Suitable for the Target Population

- Option 1: States maintain lists of boat registrations
  - Limitations on how data can be used
  - Missing some states

- Option 2: Address Based Sampling (ABS)
  - The Computerized Delivery Sequence (CDS) file provides complete coverage of all postal delivery points in the United States.
  - Prevalence of boat-ownership unknown.
  - Industry statistic: 11.9M *registered* boats in U.S. (2016) [1]

[1] https://www.growboating.org/toolkit/facts-and-figures.aspx

Frames Suitable for the Target Population

- Option 1: States maintain lists of boat registrations
  - Limitations on how data can be used
  - Missing some states

- Option 2: Address Based Sampling (ABS)
  - The Computerized Delivery Sequence (CDS) file provides complete coverage of all postal delivery points in the United States.
  - Prevalence of boat-ownership unknown.
  - Industry statistic: 11.9M *registered* boats in U.S. (2016) [1]

- Relying on ABS alone would require a substantially larger sample to obtain the desired number of boat-owning households sufficient to compute boating safety estimates

[1] https://www.growboating.org/toolkit/facts-and-figures.aspx

Frames Suitable for the Target Population

- Option 1: States maintain lists of boat registrations
  - Limitations on how data can be used
  - Missing some states

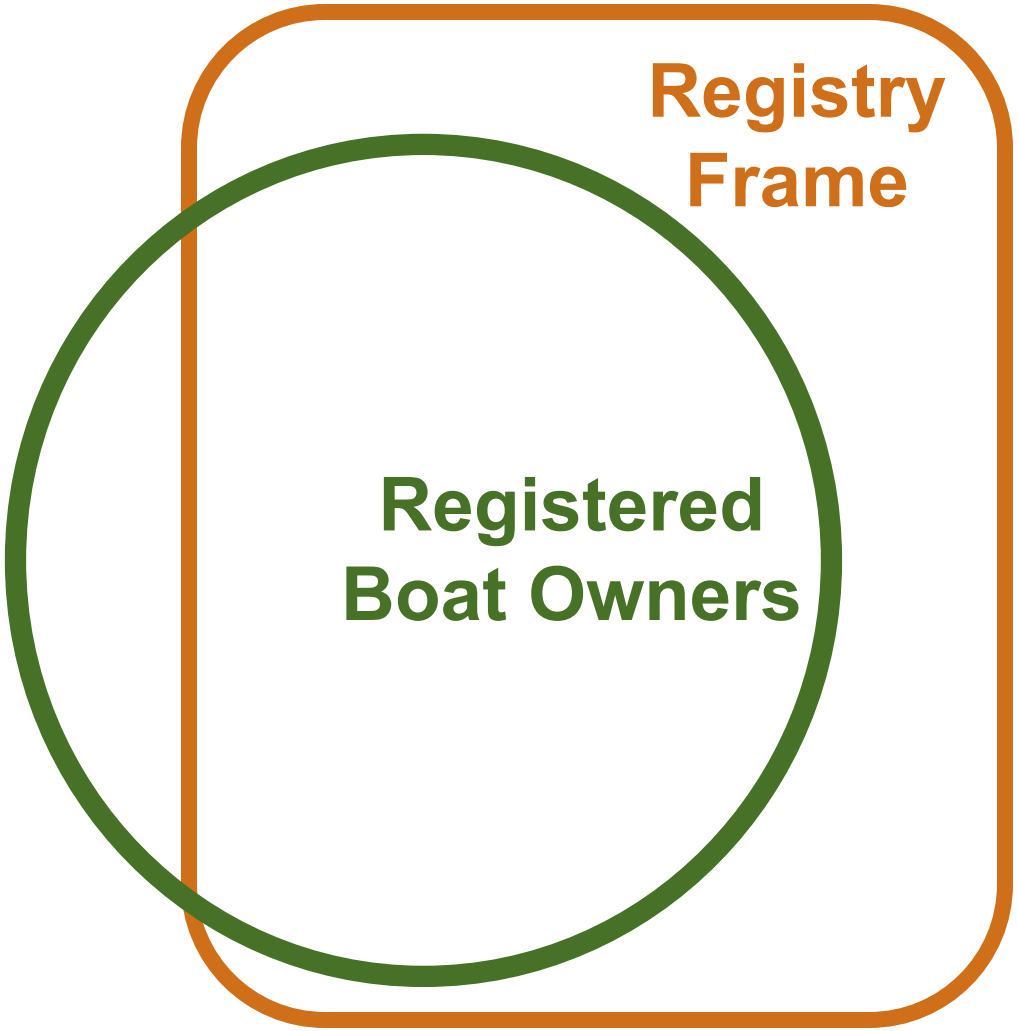
- Option 2: Address Based Sampling (ABS)
  - The Computerized Delivery Sequence (CDS) file provides complete coverage of all postal delivery points in the United States.
  - Prevalence of boat-ownership unknown.
  - Industry statistic: 11.9M *registered* boats in U.S. (2016) [1]

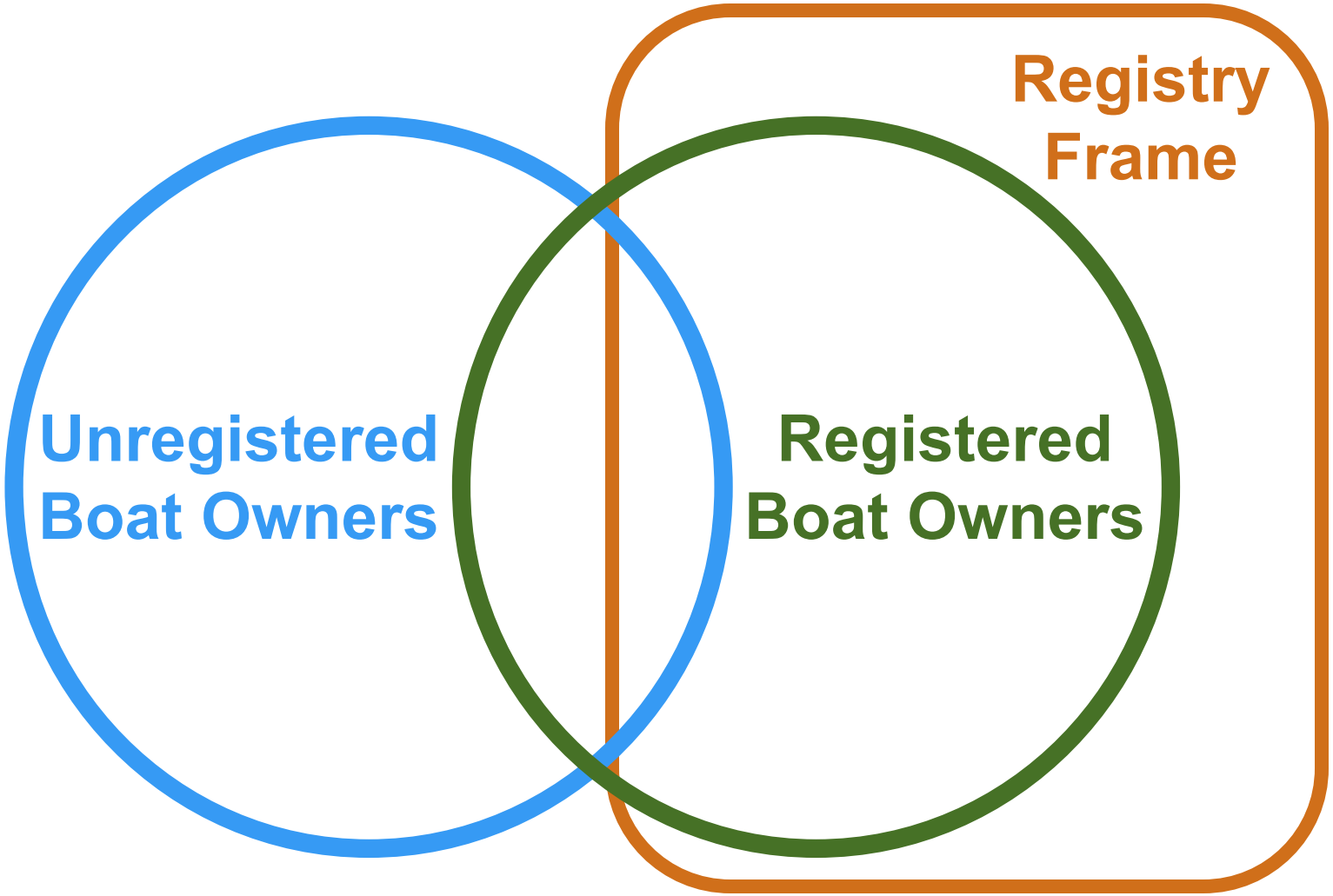
- Relying on ABS alone would require a substantially larger sample to obtain the desired number of boat-owning households sufficient to compute boating safety estimates

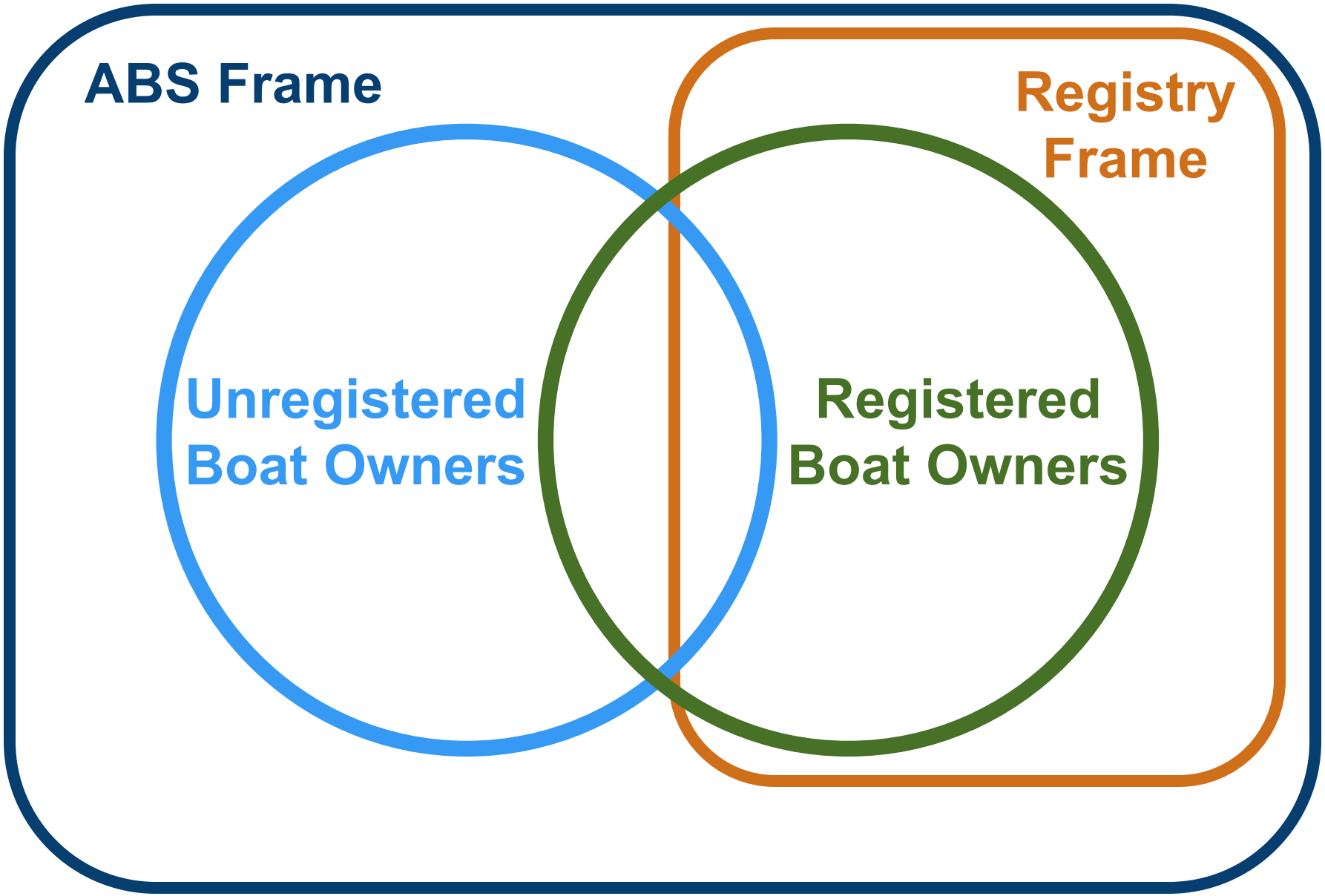
- Implemented a dual-frame survey utilizing both ABS and boat registry.

[1]https://www.growboating.org/toolkit/facts-and-figures.aspx


Registered
Boat Owners


Registry
Frame
Registered
Boat Owners


Registry
Frame
Unregistered
Boat Owners
Registered
Boat Owners


ABS Frame
Registry Frame
Unregistered Boat Owners
Registered Boat Owners

How to find the unregistered boat owners?

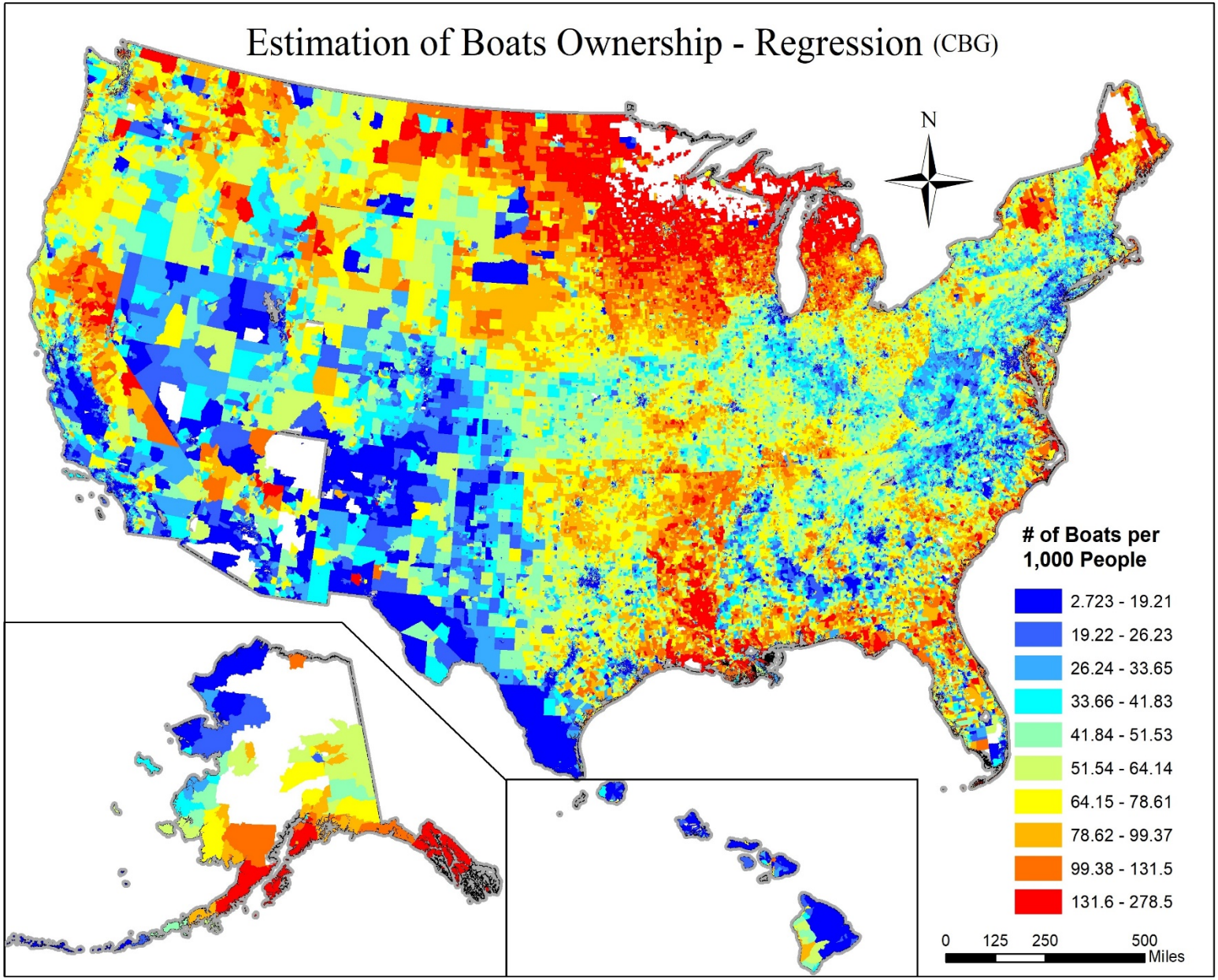
Method

- Not unusual problem: not everyone can have access to all data
- All data sources cannot be put together in entirety at one time
- Registry vendor cannot have access to ABS Frame
- Registry vendor cannot give us all addresses in registry -> take a big sample from all 50 states and DC
- Spatial data collaborator cannot have access to ABS Frame -> builds model predicting boat ownership spatially (Model 1)
- Model 1 results along with Enhanced ABS Frame variables -> build Model 2
- Model 2 predicts boat owner propensity for every address in the U.S.

Model 1 – predict number of boats in CBG

- Goal: take number of boats from ZCTA-level data and assign to Census Block (CB)
  - CB can later be aggregated to Census Block Group (CBG)
- Predictors include population density, number of owner-occupied housing units, per capita income, seven types of distance to water variables, and others.
- Linear regression model in Python ($R^2 = 0.641$)
- Model coefficients for ZCTA data applied to CBG, then normalized to total in ZCTA
- Output: estimated number of boats in CBG

Model 1 Results



Improving ABS Frame Utility

- RTI's Enhanced ABS Frame
  - CDS
  - Complete coverage of all postal delivery points in the U.S. (140M)
  - Axciom InfoBase - consumer database at person-level (over 300M)

- Boat ownership propensity model
  - Large sample of addresses matched to boat registry data from all states* (350k addresses)
  - Merged to Enhanced ABS Frame
  - Result: boat ownership propensity for all addresses

Model 2: Boat Ownership Propensity Model

- Modeling the probability of a household being eligible as a function of explanatory variables
  - Address-level model (~140M addresses in US)

- Model covariates
  - Proportion of HH in a zip code that are registered boat-owning HH
  - Estimate of boat owning household rate by Census Block Group (i.e., output from Model 1)
  - Enhanced ABS Frame variables including address-level demographic and socioeconomic indicators
  - Administrative areas public census data

Model 2: Boat Ownership Propensity Model

- Assigned 10% of data to validation dataset and developed model on remaining 90% training data
  - Stepwise-selection
  - Allowed for two-way interactions of main effects
- Prediction error measured with validation data
  - Used to terminate model-building
  - Stopping criterion: minimum average square error
- SAS PROC HPLOGISTIC

Model Selection

Statistic	Training	Validation
Area under the ROCC	0.71	0.71
Average Square Error	0.07	0.07
Misclassification Error	0.08	0.08
R-Square	0.04	0.04
Max-rescaled R-Square	0.09	0.09

Model 2: Boat Ownership Propensity Model

- Propensity variable for all addresses on ABS Frame
  - Grouped into 20 strata based on 5-percentile cut-points
  - Disproportionate allocation
  - ABS sample de-duplicated against registry frame

Data Collection Results*

	Registry Frame		ABS Frame		Total (both frames)	
	Total	Rate	Total	Rate	Total	Rate
Total	37,650		56,685		94,335	

*Results based on 6 of 12 completed cohorts

Data Collection Results*

	Registry Frame		ABS Frame		Total (both frames)	
	Total	Rate	Total	Rate	Total	Rate
Total	37,650		56,685		94,335	
Screened	13,288	35.3%	9,296	16.4%	22,584	23.9%

*Results based on 6 of 12 completed cohorts

Data Collection Results*

	Registry Frame		ABS Frame		Total (both frames)	
	Total	Rate	Total	Rate	Total	Rate
Total	37,650		56,685		94,335	
Screened	13,288	35.3%	9,296	16.4%	22,584	23.9%
Eligible	11,959	90.0%	3,939	42.4%	15,898	70.4%

*Results based on 6 of 12 completed cohorts

Data Collection Results*

	Registry Frame		ABS Frame		Total (both frames)	
	Total	Rate	Total	Rate	Total	Rate
Total	37,650		56,685		94,335	
Screened	13,288	35.3%	9,296	16.4%	22,584	23.9%
Eligible	11,959	90.0%	3,939	42.4%	15,898	70.4%
Yield		32.0%		7.0%		16.8%

*Results based on 6 of 12 completed cohorts

Data Collection Results*

	Registry Frame		ABS Frame		Total (both frames)	
	Total	Rate	Total	Rate	Total	Rate
Total	37,650		56,685		94,335	
Screened	13,288	35.3%	9,296	16.4%	22,584	23.9%
Eligible	11,959	90.0%	3,939	42.4%	15,898	70.4%
Yield		32.0%		7.0%		16.8%

Registry frame yields high, as expected, ABS frame completes coverage.

ABS alone would have been cost prohibitive.

*Results based on 6 of 12 completed cohorts

ABS Eligibility Propensity Strata Results

Propensity Strata	Screening Rate (%)	Eligibility Rate (%)	Yield (%)
1-5 (low propensity)	11.0	28.9	3.2
6-10	13.3	35.1	4.7
11-15	14.0	35.3	5.0
16-19	16.5	37.4	6.2
20 (high propensity)	17.7	45.4	8.0

Other Results

- Most registered boat owners originate from list frame (84%)
  - Respondent error in knowledge of boat registration

Other Results

- Most registered boat owners originate from list frame (84%)
  - Respondent error in knowledge of boat registration
- Non-registered boat owners originate from both frames
  - 55% list frame, 45% ABS

Other Results

- Households with only unregistered boats own more boats
  - Only unregistered average 3.1 boats in HH
  - Registered average 1.6 boats in HH

Other Results

- Households with only unregistered boats own more boats
  - Only unregistered average 3.1 boats in HH
  - Registered average 1.6 boats in HH

- No differences in average age of person completing the survey

Other Results

- Households with only unregistered boats own more boats
  - Only unregistered average 3.1 boats in HH
  - Registered average 1.6 boats in HH

- No differences in average age of person completing the survey

- Boat operated more than three nautical miles from shore in past 12 months
  - 10.3% of registered boat owners say yes, 8.7% of non-registered boat owners say yes ($p = 0.003$)

Conclusions

- When there exists a list frame that is incomplete and/or inaccurate supplement with larger data
  - Draw sample from list frame (in this case boat registry)
  - Merge to large data (in this case ABS)
  - Build model to predict eligibility
  - Apply to large data
  - Create propensity strata and allocate
- This approach maximizes coverage while controlling for screening costs
- Limitations
  - List frame is of registered boats only though supplemented with other data

More Information

Jamie Ridenhour
Research Statistician
1.919.541.6567
jridenhour@rti.org

Extra Info

Recreational boat: power boat, cabin boat, pontoon boat, air boat, house boat, personal watercraft, sail boat, canoe, kayak, paddleboard, row boat, or inflatable boat.