

Predicting Unoccupied Addresses for the 2030 Census

Andrew Keller
U.S. Census Bureau
Decennial Statistical Studies Division
2026 Joint Statistical Meetings
August 3, 2026

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. 7528546, Sub-Project No. 7530653, Disclosure Review Board (DRB) approval number: 2026-0066).

Outline

- Background
- Research Objective and Methods
- Retrospective Simulation
- Summary and Discussion

Background

2020 Census:

- Used administrative records (AR) to inform Nonresponse Followup (NRFU) operation
- Use AR to model occupied, vacant, delete statuses of NRFU universe.
- Conduct one visit to units with high confidence of final status via models.
- Among NRFU vacants: AR data was used to enumerate 14.51 percent.
- Among NRFU deletes: AR data was used to enumerate 3.59 percent.

What about cases identified as having good AR, but not enumerated using AR in the 2020 Census?

AR Identification	2020 Final Status			
	Total	Occupied	Vacant	Delete
Vacant	6,855,219	15.3%	48.9%	35.8%
Delete	2,981,128	6.1%	9.0%	84.8%

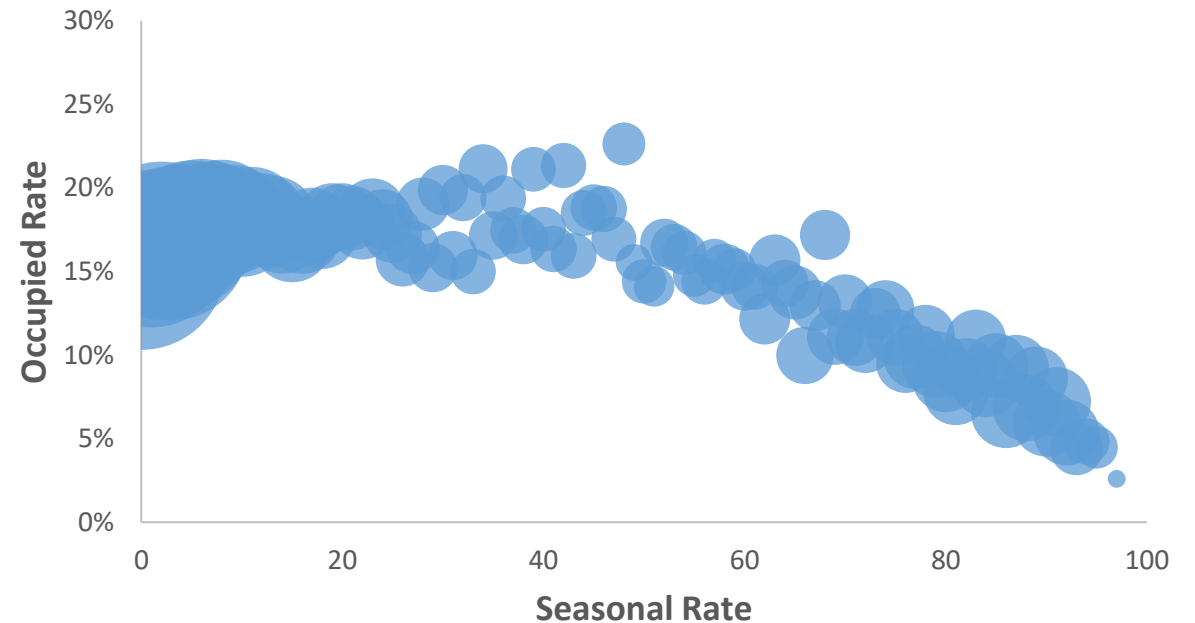
2030 Census

Research Goal:

How do we improve model predictions?

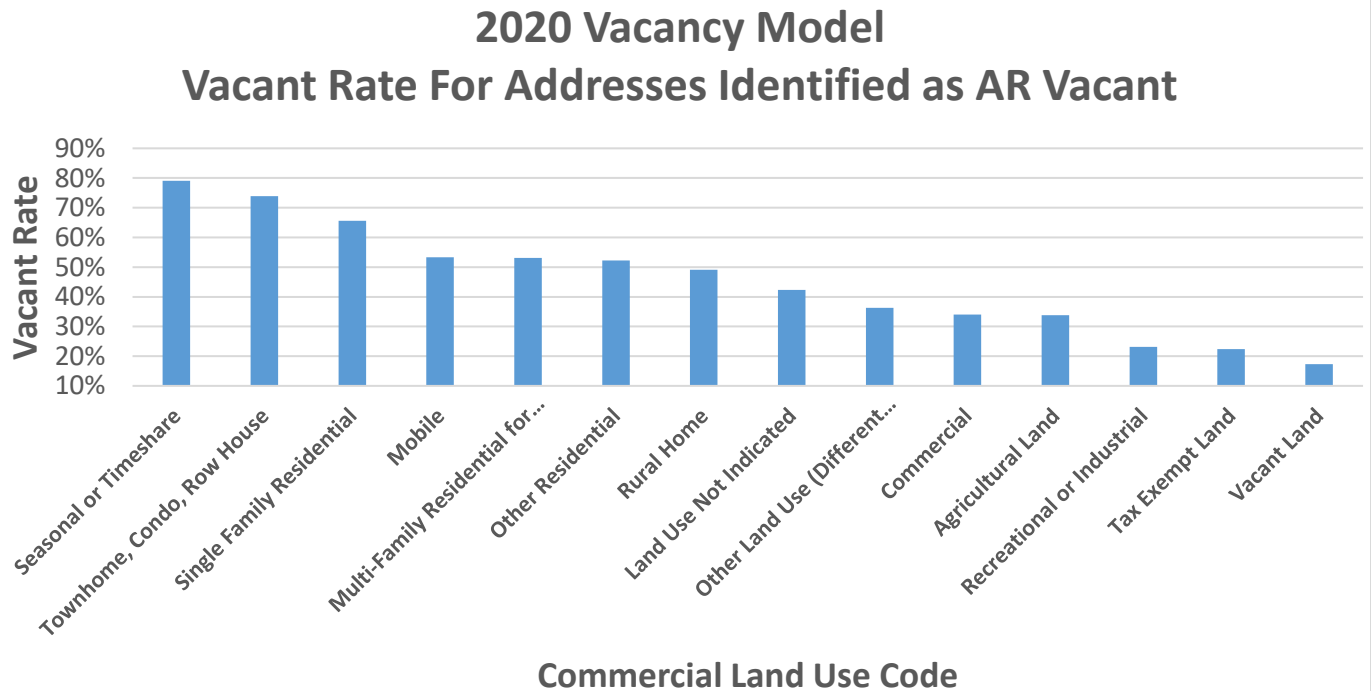
- Integrating new predictors
- Compare new model formulations to 2020 results

2020 Vacancy Model
Occupied Rate For Addresses Identified as AR
Vacant



2030 Census Research Goal: How do we improve model predictions?

- Integrating new predictors
- Compare new model formulations to 2020 results



Model: Dependent Variable

- Fitting Data: 2020 Census NRFU addresses
- Dependent variable: 2020 Census status
 - Occupied
 - Vacant
 - Non-existent/Delete
- Output: Each address will have a probability of occupied, vacant, and delete

Model: Independent Variables

Administrative Records data

- Federal AR data (Internal Revenue Service 1040 and Medicare Enrollment)
- Third-Party AR data providing information about persons at addresses
- Demographic characteristics of AR persons at address
- Duplication of AR persons across all addresses

Operational Data

- Mailing operations - Undeliverable As Addressed (UAA) from United States Postal Service (USPS)
- Neighboring addresses in NRFU universe

Address-Level information from Master Address File

- Delivery Sequence File status (Residential, Commercial, Excluded from Delivery Statistics)
- Housing Unit Type (Multi, Single, Trailer, Other)
- Description of multi-unit
- How American Community Survey classifies the address

Tract Group-level information from ACS Planning Database

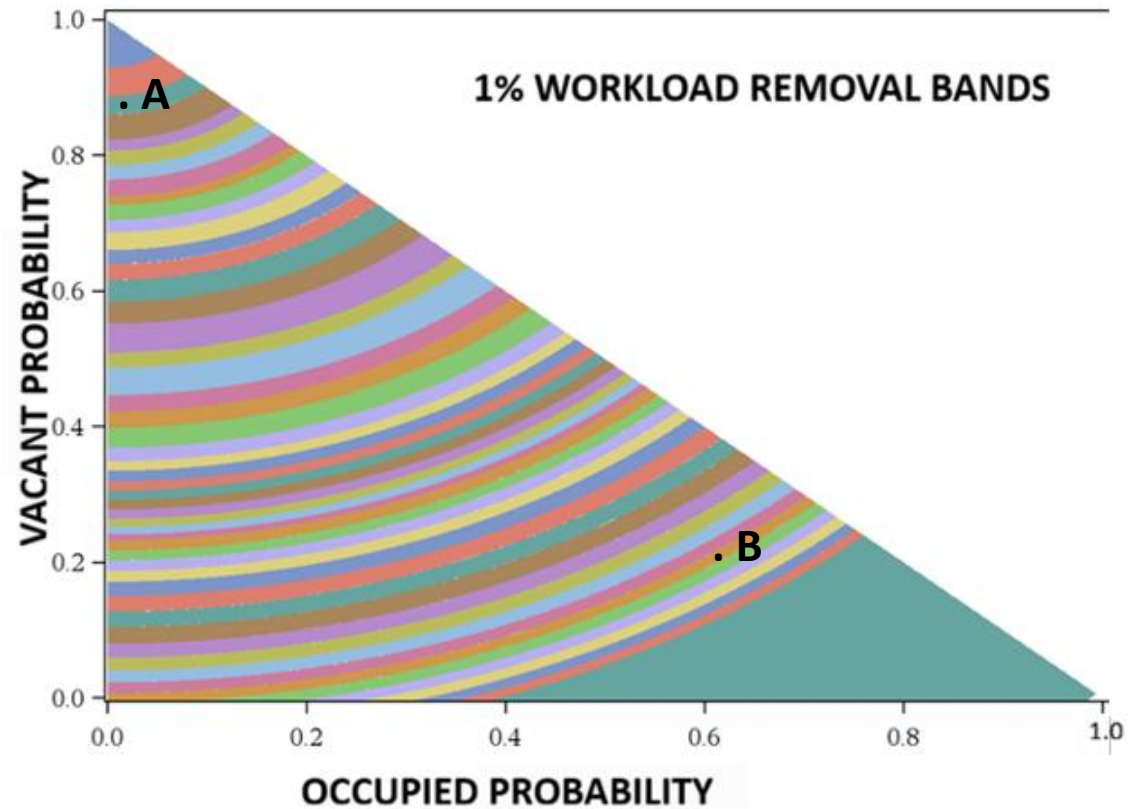
- Poverty, Rental, Other Language rates, Hispanic

Three Models

- 2020 Production (Red Model)
 - Fit on 2010 NRFU universe
- 2020 Carryover (Blue Model)
 - Fit on 2020 NRFU universe
 - Reuse variables from 2020 Production Model
- 2020 Plus (Yellow Model)
 - Fit on 2020 NRFU universe
 - Shows whether model improves with added changes

Identifying AR Vacant Units Distance Function

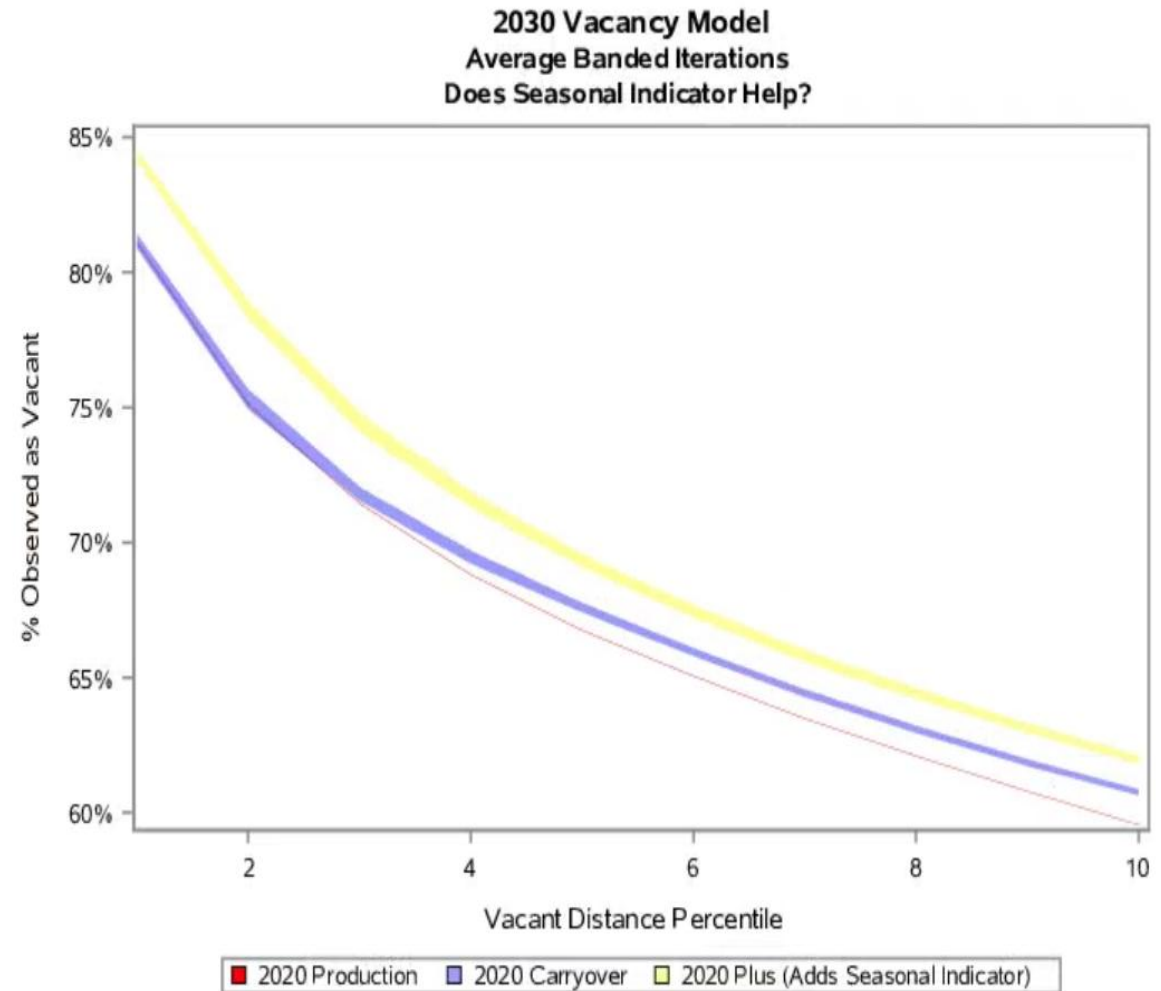
- The distance function can be visualized as successive bands of cases emanating from the point (0,1) in the top left corner
- Each successive band represents an additional 1 percent of the NRFU workload
- In this example, unit A is identified as AR Vacant while unit B is not
- Similar approach implemented for Non-existent or addresses that need to be deleted.



Adding Seasonal:

For top 10% of vacant distances, **what is the quality of prediction** when accounting for seasonal areas?

- Red: 2020 production
- Blue: Like 2020 production with changes to roster and fitting on 2020 Census data
- Yellow: Like 2020 production with inclusion of seasonal variable and Housing Unit Type
- A couple of percentage points separate yellow vs. blue/red bands.



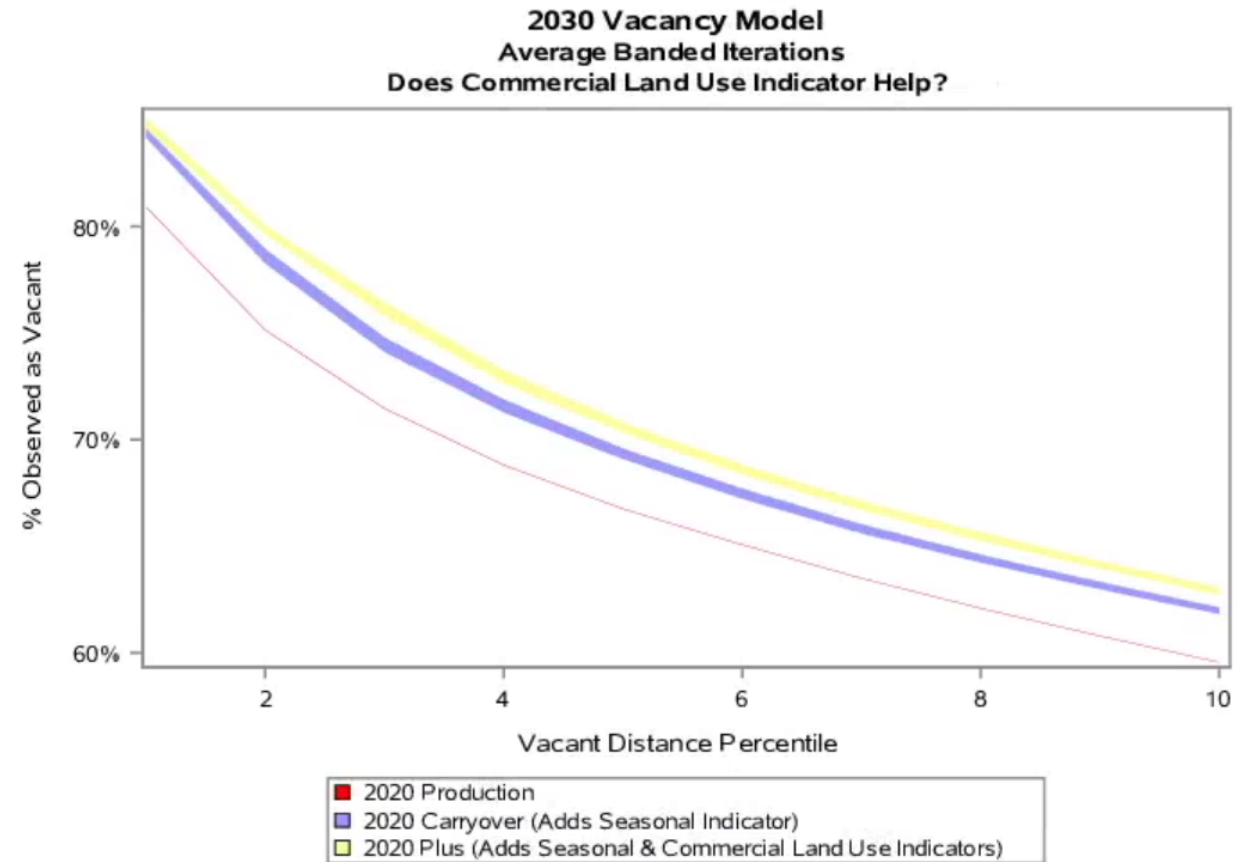
Effect on Field Work of Adding Seasonal

- Adding seasonal indicator variable identifies more cases in highly seasonal areas at the start of NRFU
- Complete fewer visits in highly seasonal areas since we would do fewer visits to the units with the best predictions of vacancy.

Adding Commercial Land Use Indicators:

For top 10% of vacant distances, **what is the quality of prediction** when accounting for land use?

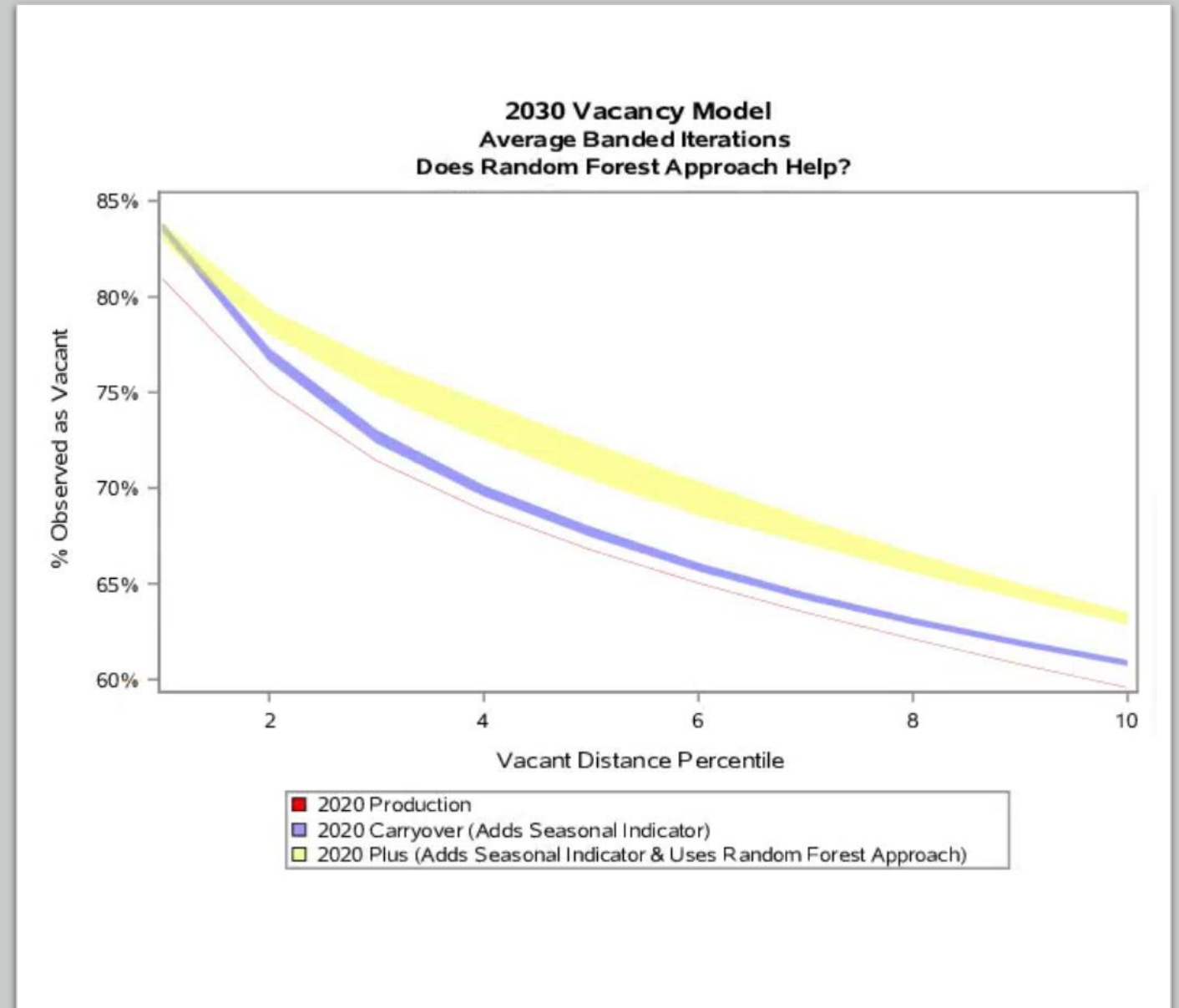
- Yellow band is always above the red line and blue band, regardless of percentile.



Changing to Random Forest:

For top 10% of vacant distances, **what is the quality of prediction** when changing the model fitting?

- The random forest approach, in which complex relationships between variables are accounted for, is an improvement over the logistic regression approach.



Research Summary

Refining unoccupied modeling methodology used for 2020 Census

- Use administrative records, operational, address-level and tract-level socioeconomic and demographic information

Incorporating changes to improve model performance

- Integrate covariates such as seasonal areas and land use indicators
- Change statistical modeling fitting framework

Goals

- Reduce misclassification
 - Modeled unoccupied vs. 2020 Census occupied
 - Modeled vacant vs. 2020 Census delete
- Understand how field work might be affected

Going Forward...

Incorporate more information

- USPS Delivery Sequence File
- How was address added onto census frame?

Understand changes to field contact strategy

- Should we do something different than 2020?

Apply new models in 2026 and 2028 Census Tests

Questions?

andrew.d.keller@census.gov