Predicting Vacant Housing Units in the American Community Survey

Andrew Keller and Tom Mule
U.S. Census Bureau
Decennial Statistical Studies Division
2023 FCSM Research and Policy Conference
October 24, 2023

United States® Census Bureau

Outline

Background

Research Objective and Methods

Simulation

Discussion

United States® Census Bureau

Background

- 2020 Census:
  - Can we use administrative records (AR) to inform Nonresponse Followup (NRFU) operation?
  - Use AR to model occupied, vacant, delete statuses of NRFU universe.
  - Modify contact strategy where we have high confidence in address status.
  - Conduct one visit to units with high confidence of final status via models.

- Apply same concept to predict units with high confidence of vacancy in American Community Survey (ACS) while accounting for a conceptual difference:
  - We used census-related mailings around April 1, 2020 to determine vacancy for the same date.
  - We use ACS-related mailings to help determine vacancy two months later. For example, for the March ACS panel, the ACS mailings from March will help to determine whether the unit will be vacant when we go to interview in May.

- Use probability to inform contact or sampling strategy.

United States® Census Bureau

Model

- Random Forest Approach: Dependent Variable is vacancy outcome status.
- Use Previous Year(s) of ACS to form training data.
  - Administrative Records data of the same vintage.
  - Operational Data.
  - Address-Level information from Master Address File.
  - Block Group-level information from ACS Planning Database.
- Apply parameter estimates to current vintage of ACS data.
- Predicted probability of vacancy for every ACS unit in mailable Computer Assisted Personal Interview (CAPI) universe.

United States® Census Bureau

Data

Administrative Records data

- Federal AR data (Internal Revenue Service 1040 and Medicare Enrollment)
- Aggregated public information purchased by Census Bureau consisting of local tax, deed, and mortgage information
- Using information concerning land use, absence of owner at address, ownership rights on the unit
- Third-Party AR data providing information about persons at addresses
- National Change of Address information from United States Postal Service (USPS)

Operational Data

- Mailing operations (undeliverable as addressed from USPS)
- Indication of vacancy from internet response

Address-Level information from Master Address File

- Delivery Sequence File status (Residential, Commercial, Excluded from Delivery Statistics)
- Housing Unit Type (Multi, Single, Trailer, Other)
- Delivery Point Type

Block Group-level information from ACS Planning Database

- Poverty, Rental, Other Language rates, Hispanic

United States® Census Bureau

Simulation Setup

1) Take mailable CAPI universe cases from 2021, 2022 ACS universe.

2) Fit model on 2021 data.

3) Score model on 2022 data.

4) Sort predicted vacant probabilities from greatest to least.

5) Iterate over top percentages by picking a threshold. (i.e. – top 10% or 5% of predicted vacant probabilities).

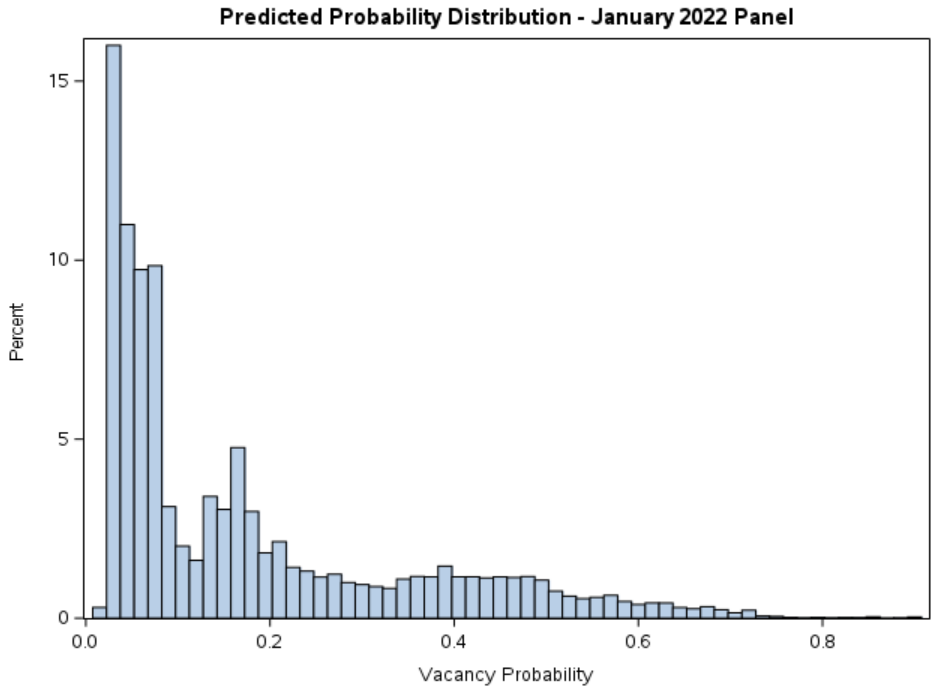
6) See how many of those were vacant in 2022 (About 25% of universe is vacant).

United States® Census Bureau

Simulation Results: Predictions

1) Can we design a model to reasonably predict vacant cases?

2) What threshold should be used to perform treatment?



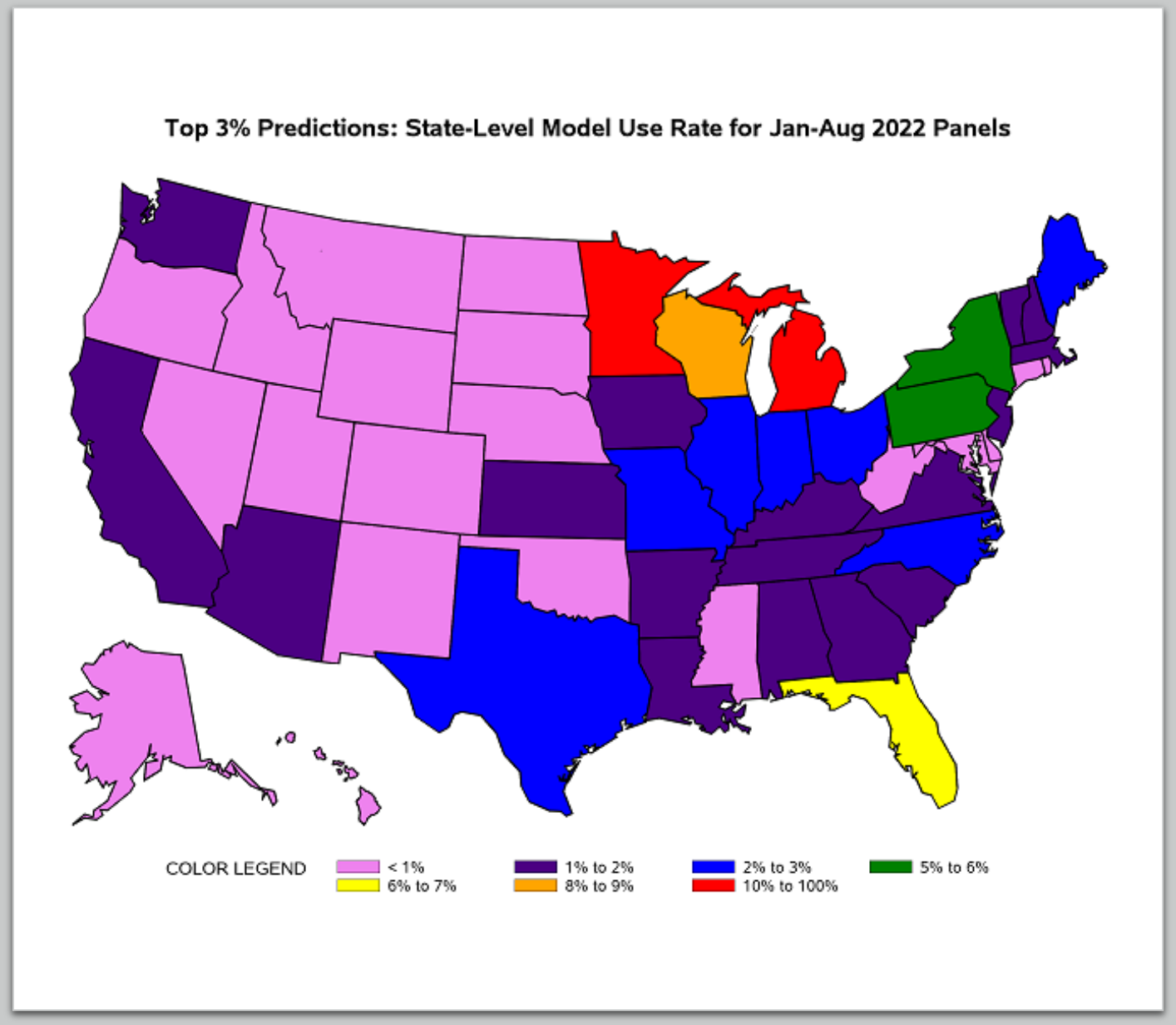
United States® Census Bureau

Simulation Results: State-Level Model Use Rates

How often do the top 3% of vacant predicted probabilities fall into the respective states?

- UT, WY, MT, SD < 1%
- CA < 2%, TX < 3%
- MI, MN > 10%

The top 3% of vacant predicted probabilities in each state is not proportional to the state population.



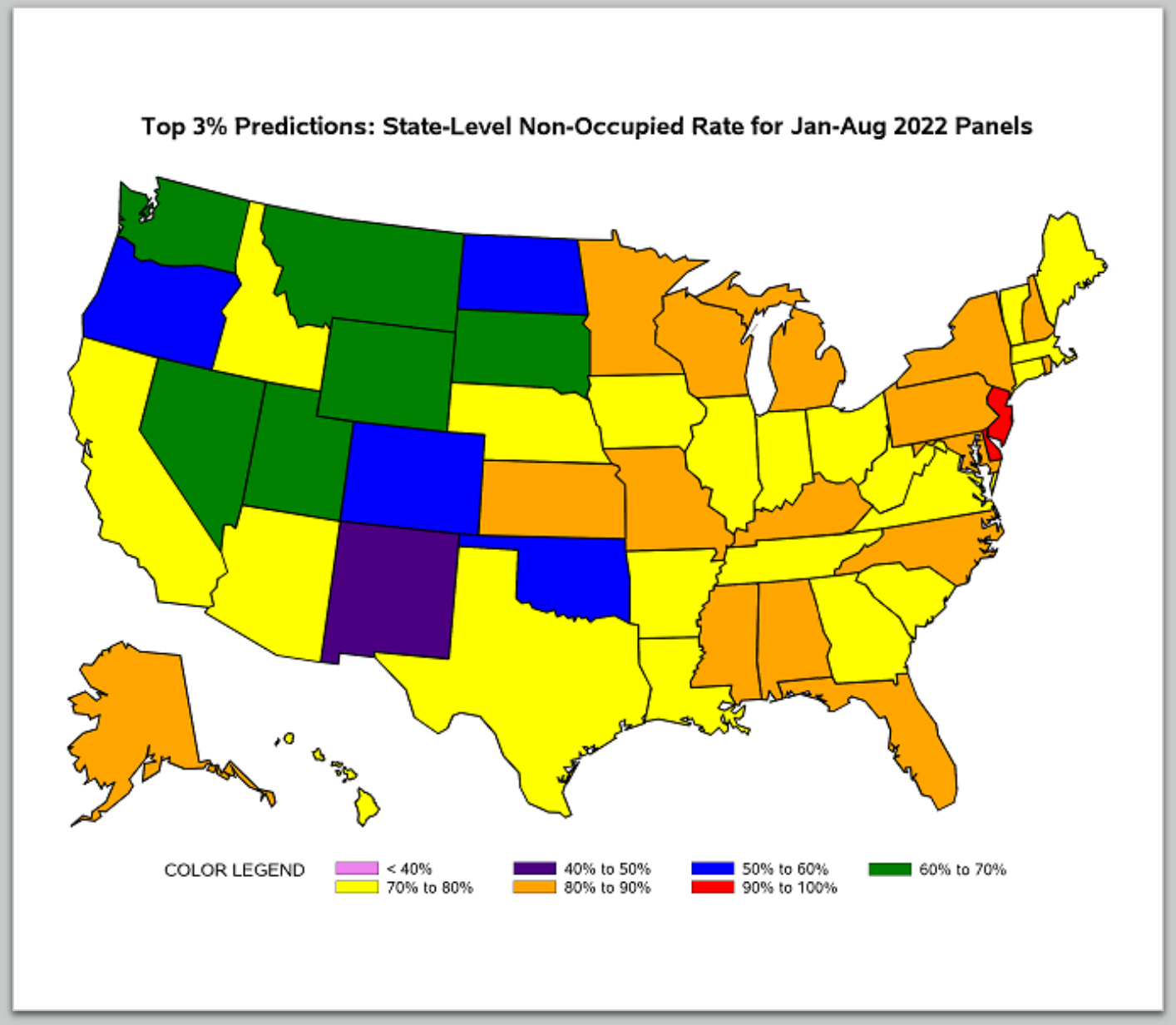
United States® Census Bureau

Simulation Results: State-Level Predictions

For Top 3% of vacant predicted probabilities, what is the state-level rate at which the vacancy prediction is non-occupied?

- UT, WY, MT, SD < 70%
- CA, TX < 80%
- MI, MN < 90%

The states with HUs that fall into the top 3% more often have higher true positive rates.

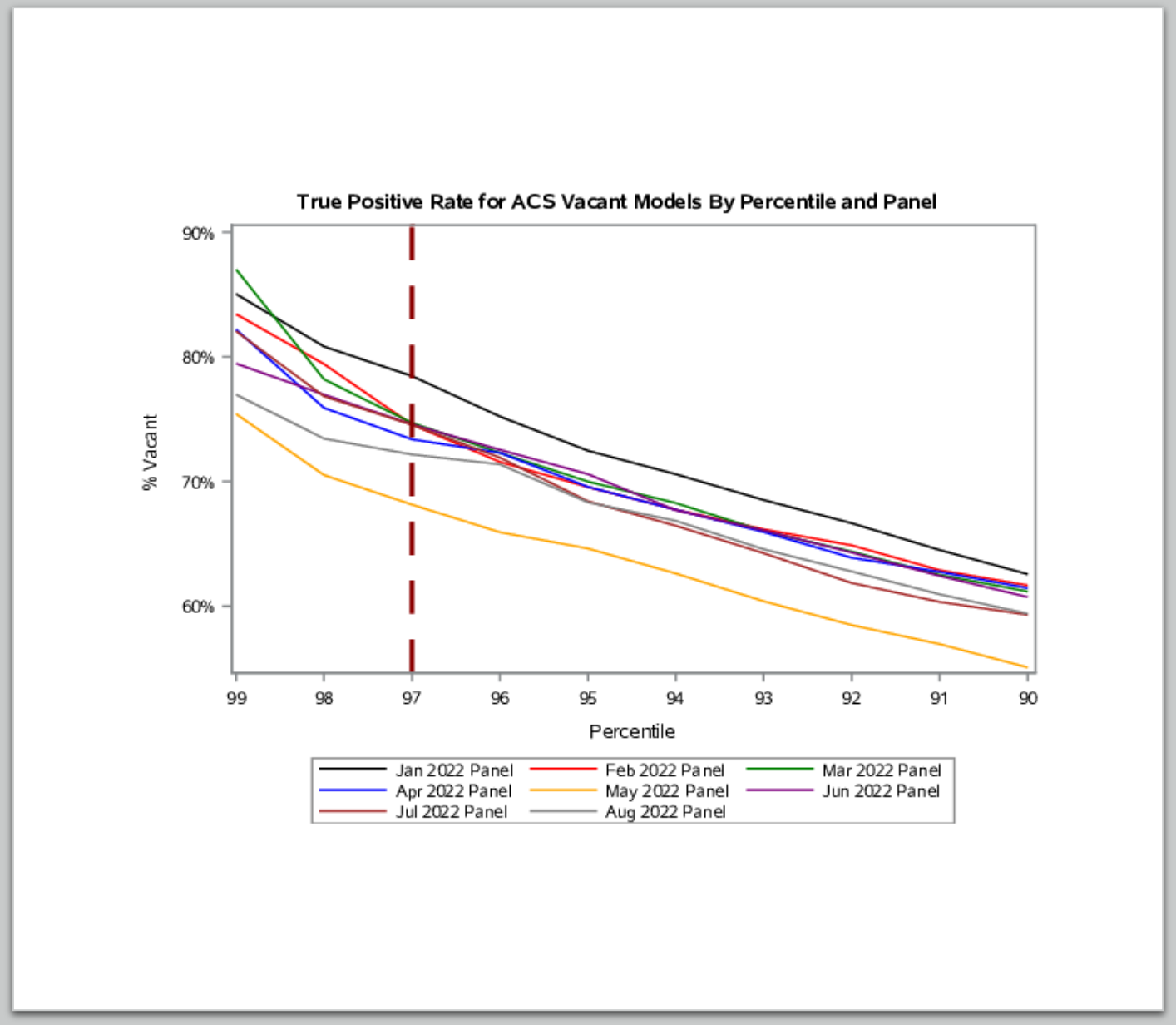


United States® Census Bureau

Simulation Results: Vacancy Status

Given we identify the top 3% of vacant predicted probabilities – how often are they vacant?

- January 2022 Panel - 78% vacant
- May 2022 Panel - 68% vacant

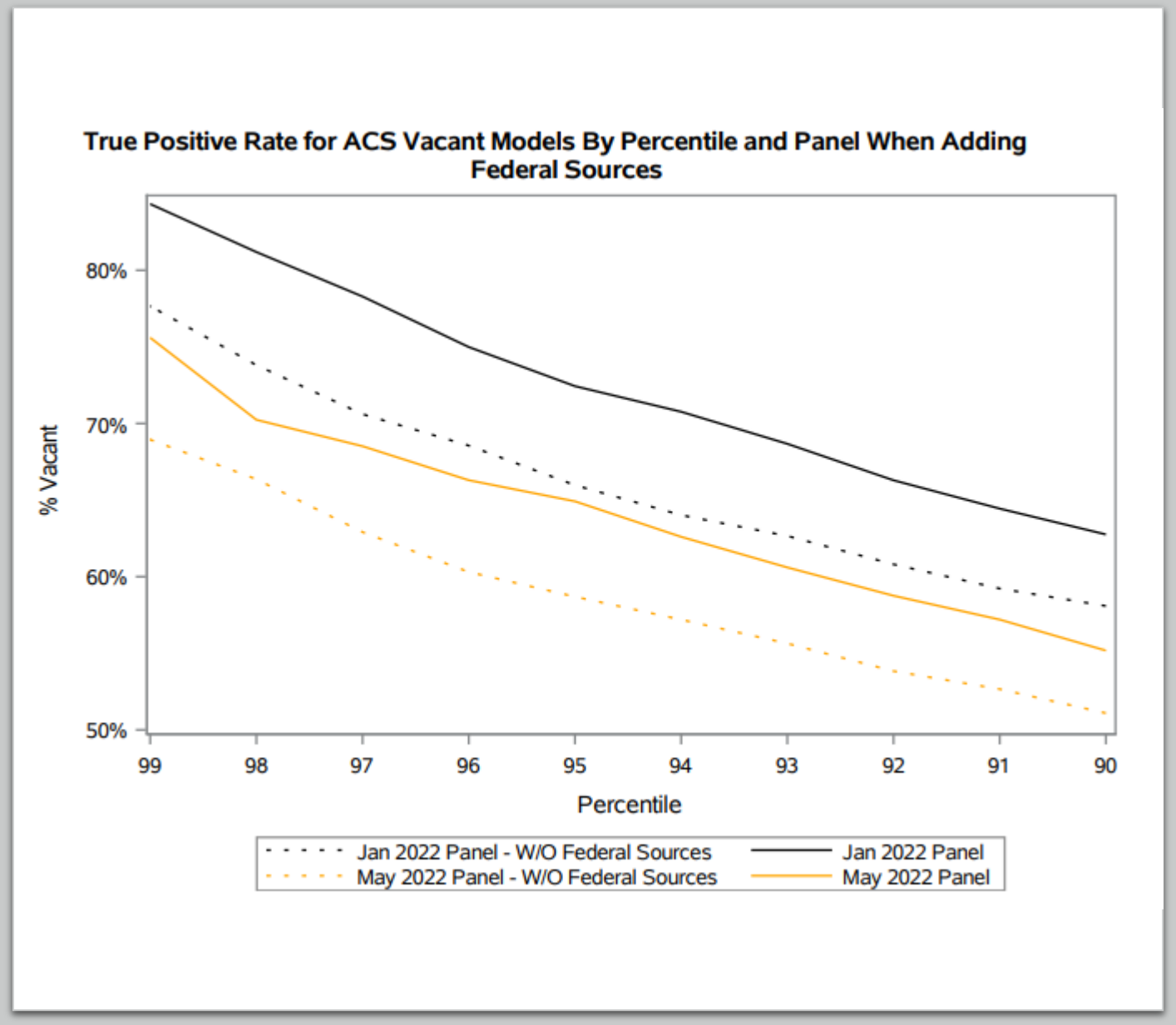


United States® Census Bureau

Simulation Results: Vacancy Status (2)

Marginal Gain of Adding Federal Sources to model

Modeling results show a 4-8% improvement in vacancy prediction across panels for various percentiles when federal sources are added into the model.

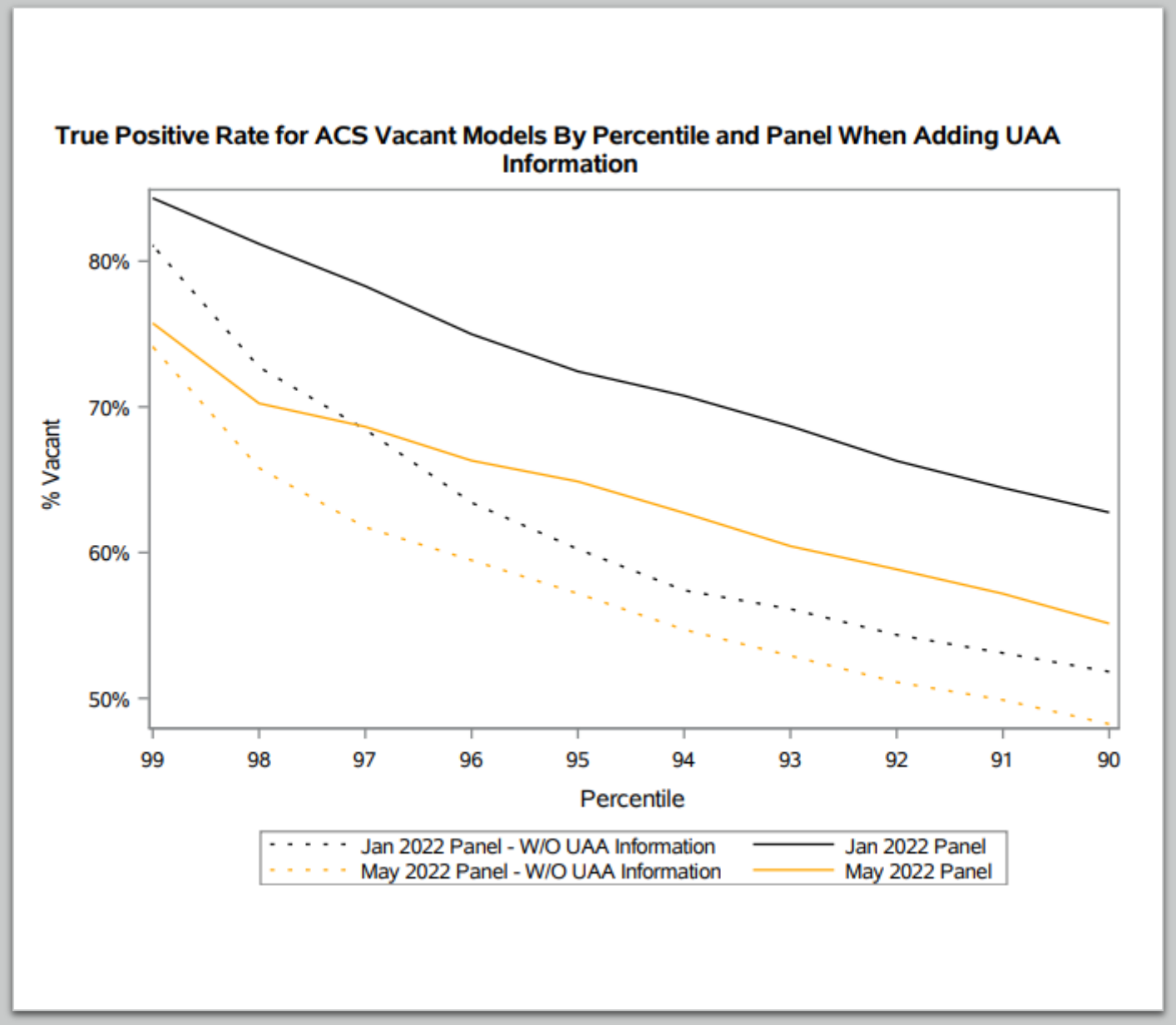


United States® Census Bureau

Simulation Results: Vacancy Status (3)

Marginal Gain of Adding UAA Information to model

Modeling results show a 4-12% improvement in vacancy prediction across panels for various percentiles when federal sources are added into the model.



United States® Census Bureau

Analysis of Case Identification and Errors

- Study top vacant predicted probabilities with non-vacant outcome.
- Develop understanding where we might be more sensitive to calling it vacant.

Example:

- Take top 10% of predicted probabilities of January 2022 Panel.
- Indication of an apartment on address file:
  - 24% of CAPI universe are apartments.
  - 12% of top 10% of predicted probability universe are apartments.
  - 15% of false positive universe is an apartment.
  - 10% of true positive universe is an apartment.
- Apartments are less likely to be modeled in the top 10% of predicted probability that their CAPI distribution. (24% vs. 12%).
- Apartments comprise distributionally more of the false positives and fewer of the true positives. (15% vs. 10%)

United States® Census Bureau

Conclusions and Generalizing

- Modeling vacant units in the ACS universe can be completed using a combination of address-level, ACS operational, geographic, and administrative records information.

- Generalizing 2020 Census methodology to account for survey design.
  - Challenge: Enhance ACS model to account for two-month delay between ACS mailing and vacant identification.

- Analyzing cost-benefit tradeoffs will help determine the threshold for using the best predictions.

- Observing differential case identification and true positive rates across geographies.

United States® Census Bureau

Conclusions and Generalizing (2)

- Universe does not have to be mailable CAPI cases – can include all CAPI cases.

- We can use contact history to update model.
  - Example – feed results from first contact into model, update probabilities with that information.

- We can use predicted probabilities to alter contact strategy.
  - This is how the information was applied for the 2020 Census.

- We can use predicted probabilities to change sampling rates.
  - Risk: Changing sampling rate for high probability vacant cases that are occupied inflates variances.

United States® Census Bureau

Any questions?

Andrew.D.Keller@census.gov

United States® Census Bureau