Synthesizing Housing Units for the American Community Survey

Rolando A. Rodríguez
Michael H. Freiman
Jerome P. Reiter
Amy D. Lauger

CDAC: 2017 Workshop on New Advances in Disclosure Limitation

September 27, 2017

Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

What to take away from this talk

The Census Bureau must maintain data confidentiality / privacy in public output from its censuses and sample surveys.

The Census Bureau is researching new disclosure avoidance methods for the American Community Survey (ACS).

Researchers have generated fully-synthetic data for ACS housing units at the state level for a single state in a single year.

How to apply formal privacy methods to the problem is an open question.

The American Community Survey (ACS)

The ACS is the Census Bureau’s largest demographic survey.

A single year of ACS collection results in ~ 2.3 million housing-unit responses.

1-year and 5-year products released annually since 2005.

5-year data products consist of a ~ 2/3 microdata sample and over 1000 tables given for every block group.

ACS is the basis for the distribution of ~ $670 billion in federal funds annually.

Title 13 demands data release without identification

“Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, […] may […] make any publication whereby the data furnished by any particular establishment or individual under this title can be identified”—Title 13, U.S. Code, §9

We cannot permit even the disclosure of participation in the ACS.

Direct identifiers like name and address must obviously never appear in releases.

Every data release we make provides additional information about the respondents.

How do we meet the demands?

The Bureau has used a variety of methods to reduce the risk of identification.

Internal ACS data are treated with methods such as swapping.

Released data have additional controls such as top-coding and table suppression.

How do we define global disclosure risk for ad hoc methods?

- Matching external data to released ACS data
- Synthesizing “identifiers” or “quasi-identifiers”
- Chance of reproducing original records

Ideally we would make formal privacy guarantees

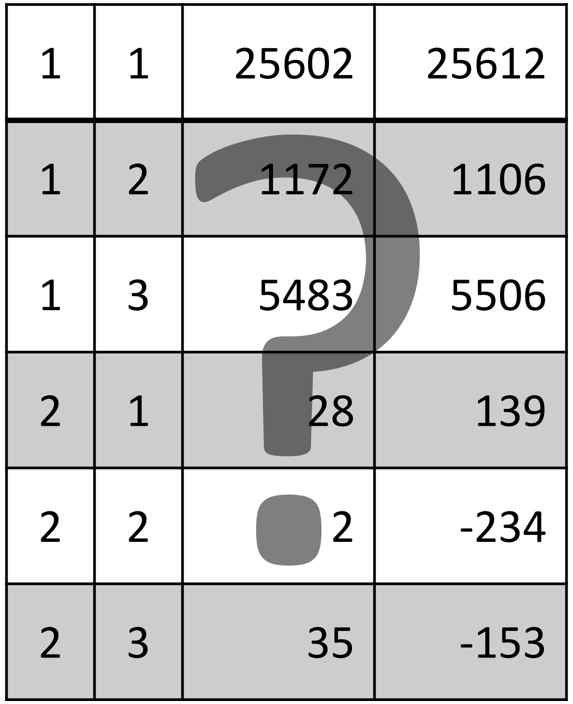
Formal privacy methods hold themselves to quantitative definitions of risk.

A serious effort is underway to make the next census formally private.

The ACS complicates the task:

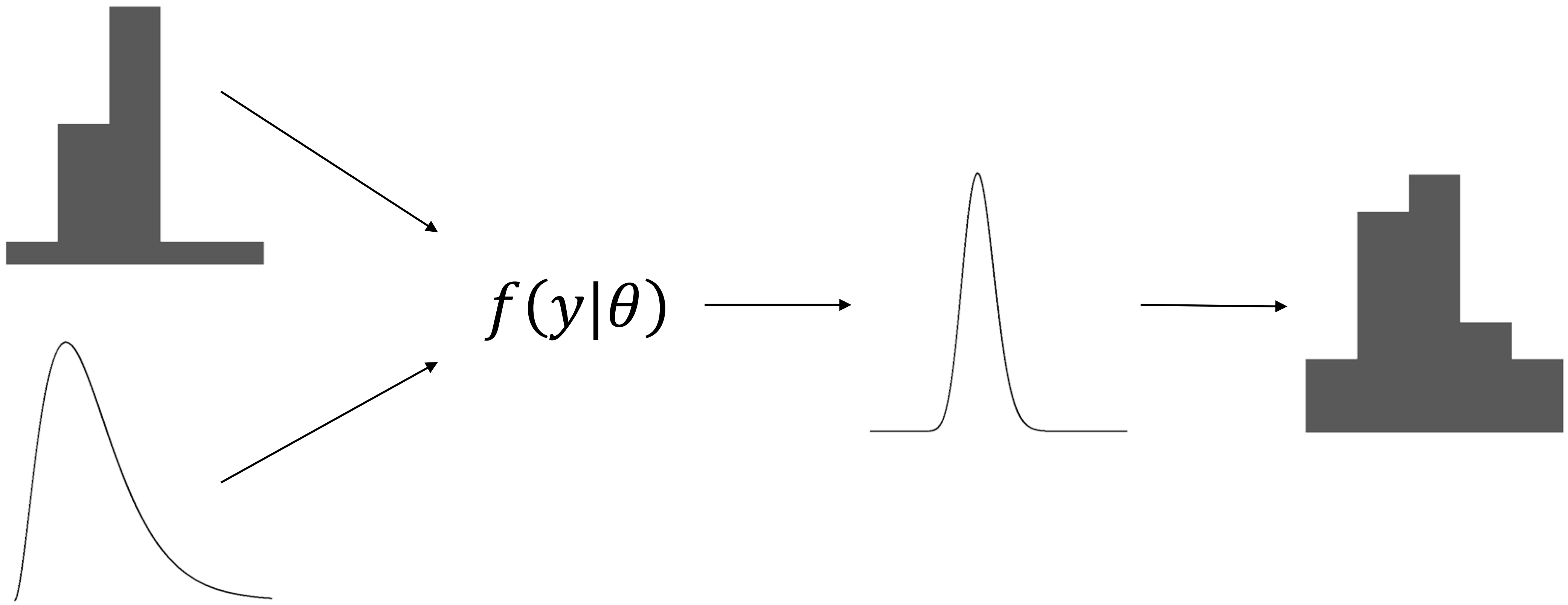
- ACS has more characteristics for housing units and people.
- ACS has complex survey weights.

We will first try synthetic data methods.



1	1	25602	25612
1	2	1172	1106
1	3	5483	5506
2	1	28	139
2	2	2	-234
2	3	35	-153

Synthetic data are predictions from models



$f(y|\theta)$

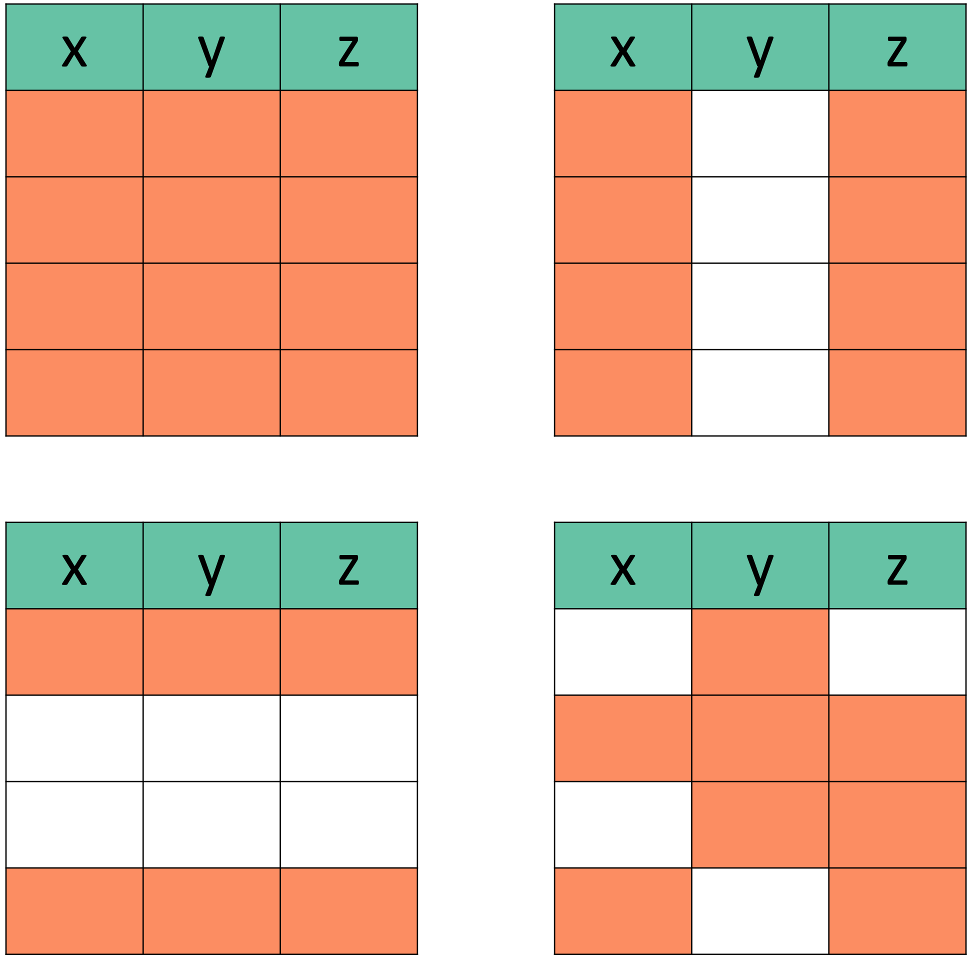
Synthetic data come in flavors

Synthesis of every variable for every record = fully synthetic data.

Anything else = partially synthetic data.

Partially synthetic data can be row (record) or column (variable) partial, or both.

Partially synthetic data currently used for disclosure avoidance in ACS group quarters.



x	y	z
🟧	🟧	🟧
🟧	🟧	🟧
🟧	🟧	🟧
🟧	🟧	🟧

x	y	z
🟧		🟧
🟧		🟧
🟧		🟧
🟧		🟧

x	y	z
🟧	🟧	🟧
🟧	🟧	🟧

x	y	z
	🟧	
🟧	🟧	🟧
	🟧	🟧
🟧		🟧

Our current plan: develop synthetic data, then make it formally private

Create fully-synthetic data for housing unit attributes at coarse geographies (state).

Once housing unit results are reasonable, synthesize persons, then geographies.

Models are fit conditionally on previous models to build up a joint distribution:

$$f_Y(y|\Theta) = f_{Y_1}(y_1|\Theta_1) f_{Y2|Y1}(y_2|y_1, \Theta_1, \Theta_2) \ldots$$

What models?

Two useful models are CART and regression

We use classification and regression trees (CART) to synthesize factors and counts.

CART does not directly fit into posterior-predictive paradigm.

We use linear regression to synthesize (rounded) continuous variables.

Regression does allow for posterior prediction, but has more assumptions.

We like trees because they grow easily

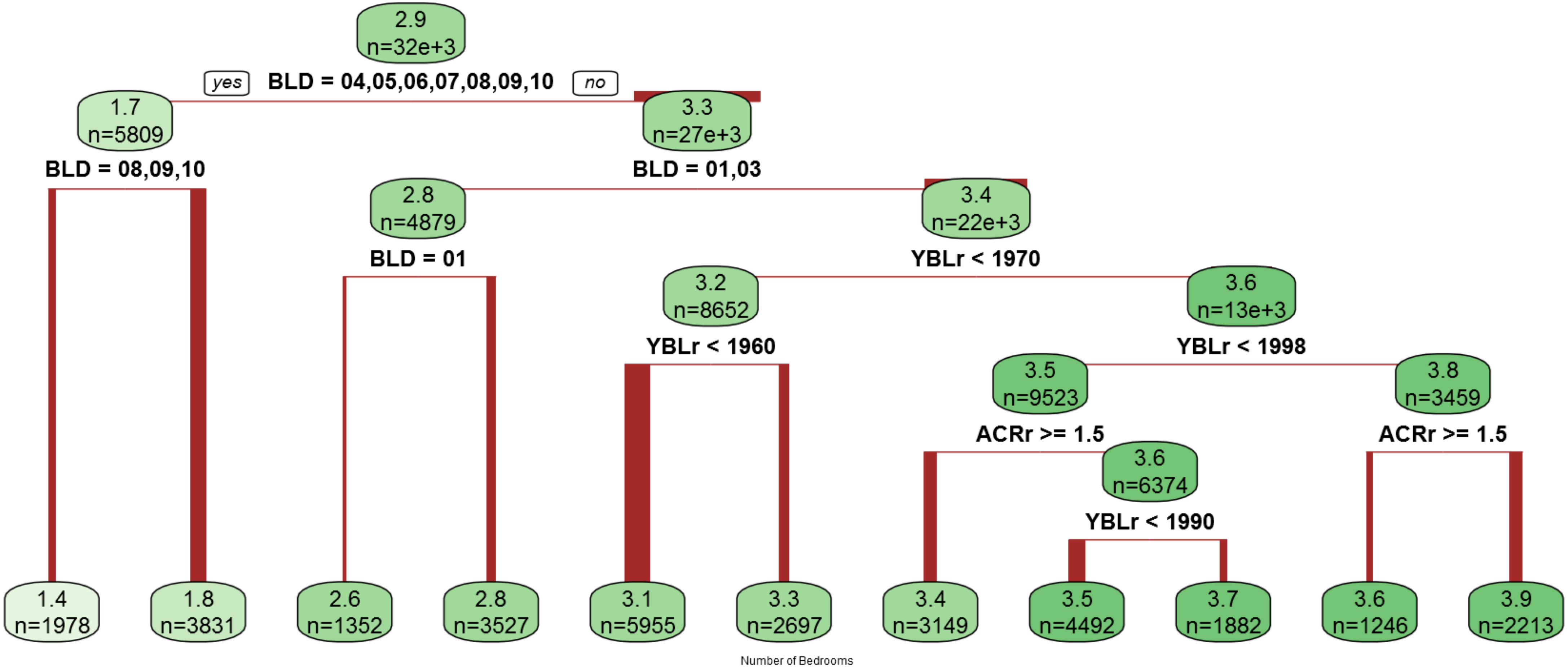
Classification and regression trees make binary splits of a variable based on predictors and homogeneity criteria.

Graphically, we represent the splits as a tree with data in the leaves.

CART can capture non-linear relationships and interactions automatically.

Synthetic data is drawn as a Bayesian bootstrap of leaf values.

Here’s a tree grown on ACS public microdata



Trees with too many leaves can overfit

For prediction we want accurate fits, so we need more than a sapling.

Why not just allow the most leaves we can grow?

Leaf values are actual data, so we have to consider risk of value reproduction.

Continuous predictors can grow lots of leaves and can produce overly precise splits.

Regardless of risk paradigm, we prefer to avoid reproducing the original data.

We use regressions for continuous variables

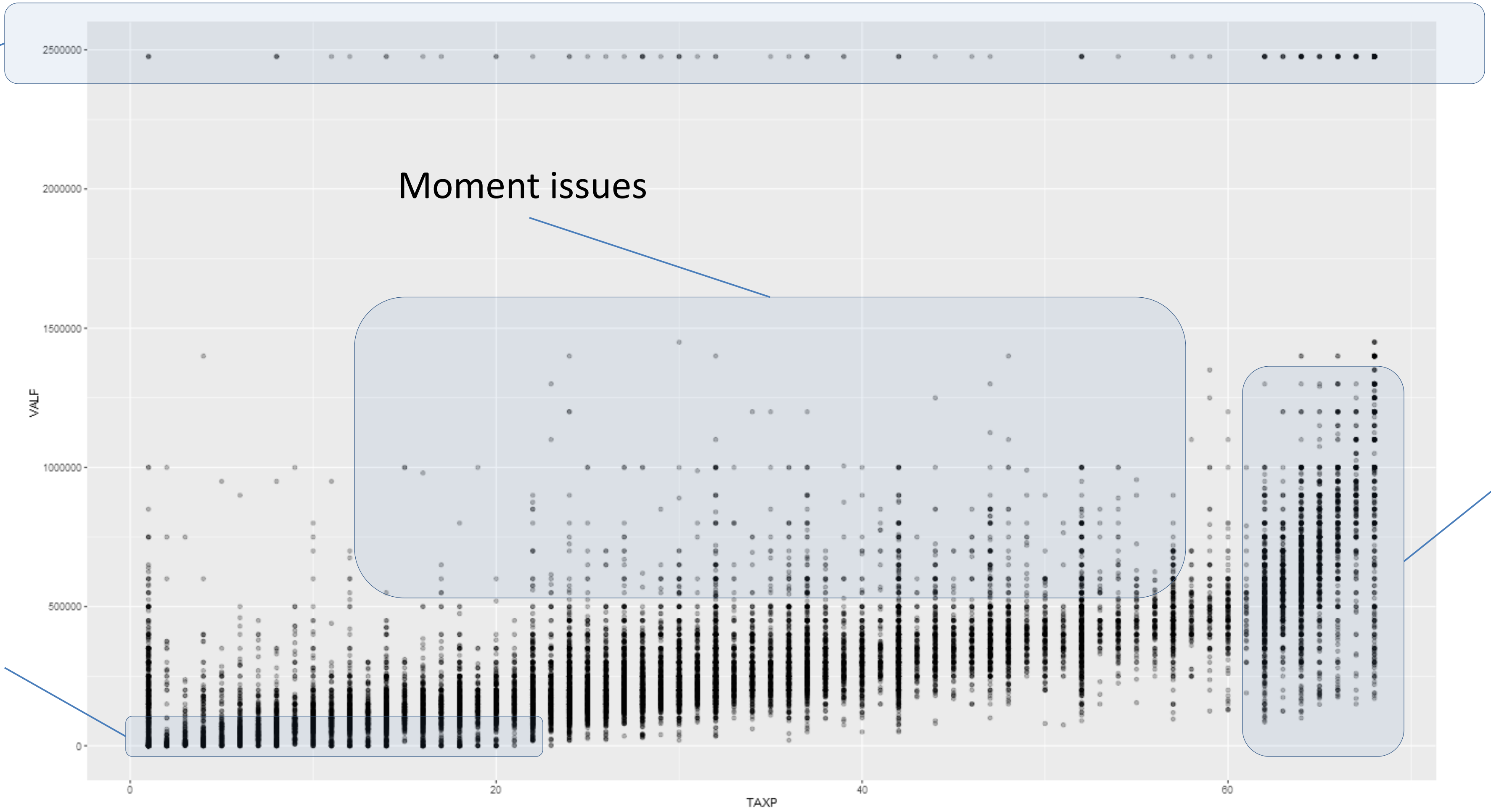
OLS regressions are easy, fast, explainable, assessable, and synthetically proper.

Redrawing an exact record is theoretically impossible and practically unlikely.

Interactions and transformations allow for rich models and control of accuracy.

Proper synthesis via regression demands adherence to model assumptions.

Real data often violate regression assumptions



Censored outliers

Moment issues

Non-linear relationship

Range not (-∞, ∞)

We can still make regression a useful model

Transformations can mitigate some of these issues.

Regression diagnostics can inform these and other fixes.

Ideally solutions can be found that are broadly applicable across geographies.

Regardless, if the data user tries the same regression, good things will happen.

What if analysts are not using trees and regression?

Any gulf in assumptions between analysis and synthesis models can cause issues.

We cannot predict all analyses users might perform on the ACS public-use microdata.

We can look at changes in the public ACS tables.

CART is a greedy search through a table space.

Regression is concerned with conditional means.

Results for tabulations are mixed

We assess unweighted synthetic table counts.

- Generate bootstrap tables
- Find quantile of synthetic table in the bootstraps based on a metric

We see issues but no clear patterns.

Few housing-unit-only tables are published.

Generate random tables for assessment.

Table		Synthetic Table Quantile
Monthly costs		1.00
Units in Structure		0.99
Heating Fuel		0.54
Housing-unit value		1.00
Housing-unit value (detail)		1.00
Number of Rooms		0.98
Number of Bedrooms		1.00
Has a mortgage		0.05
	Second loan	1.00
	Monthly costs	1.00
Owned/Rented		0.31
	Household Size	1.00
	Number of Rooms	0.96
	Number of Bedrooms	1.00
	Number of Vehicles	0.22
	Number of Vehicles (detail)	0.50
	Heating Fuel	0.40
Rent (yes/no)		0.93
	Rent amount	1.00

Open questions

Can we use formal privacy methods on some subset of the variables?

Can we make current methods formally private?

How do we account for survey weights?

How do results look after placing housing units in sub-state geographies?

How can we leverage alternate data sources (administrative records)?

Thank you!

rolando.a.rodriguez@census.gov