Model-Assisted State Expenditure Estimates

Clayton Knappenberger
Yezzi Angi Lee
U.S. Bureau of Labor Statistics
Division of Consumer Expenditure Surveys

Outline

- Consumer Expenditure Surveys (CE)
- Project Goal
  - Existing Products
  - Provide Additional States
- Model-Assisted Method
  - Why use MAEs?
  - Auxiliary Data Used
- Models Explored
  - Cross-validated Errors
- Results/Comparisons/Limitations

Consumer Expenditure Surveys

- Two surveys providing data on expenditures, income, and demographics of US consumers

Quarterly Interview	Weekly Diary
Large purchases	Small purchases
Recurring payments	Frequent spending
Three-month recall	Contemporaneous
Rotating panel	Rotating panel
Four waves	Two waves

Project Goal

- CE Sample is meant to represent the US non-institutional civilian population
- Currently publish
  - 4 Regions, 9 Divisions, 5 States, and 23 major urban areas
- **Users frequently ask us for States**
  - **Can machine learning help us?**

Existing State-level products

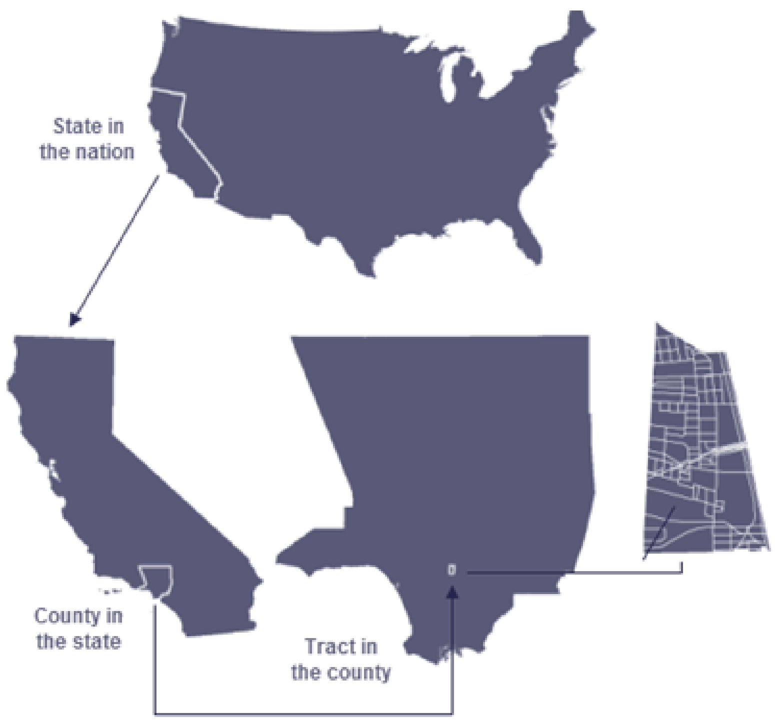
- CE currently provides estimates for 5 States
  - Large and representative samples

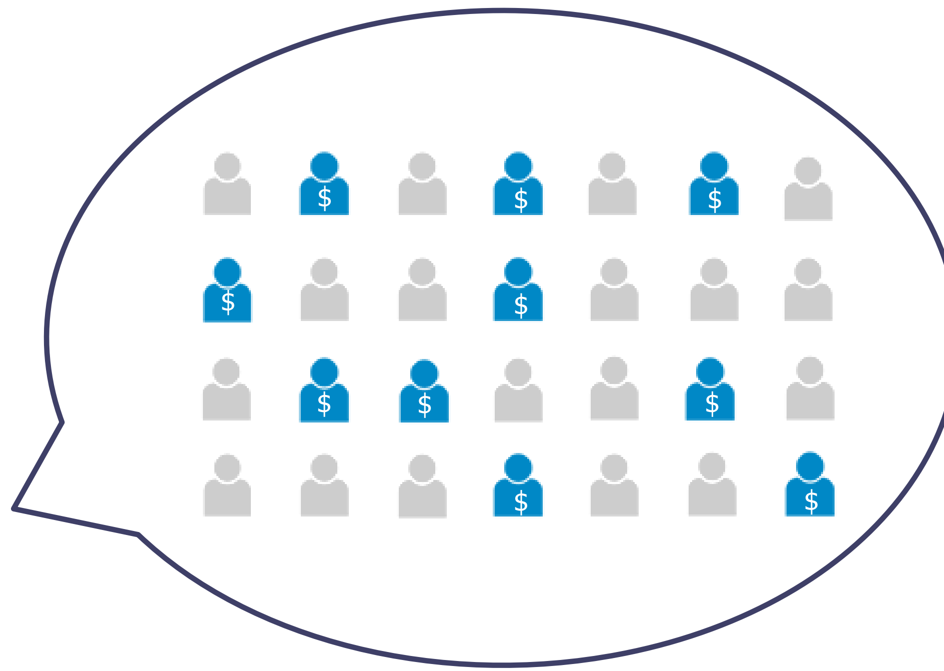
Provide Additional States!

- Feasibility study using Gradient Boosting Machines

Model-Assisted Method

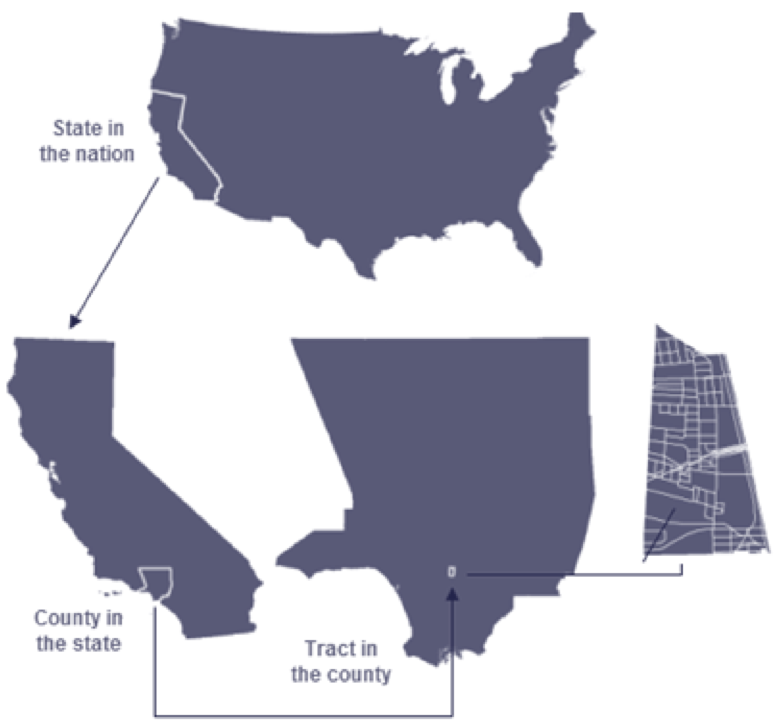
- Using a model to combine sample data with auxiliary data from areas not sampled





Model-Assisted Method

- Model predicts expenditures for each area in the auxiliary data giving us total coverage



Model-Assisted Method

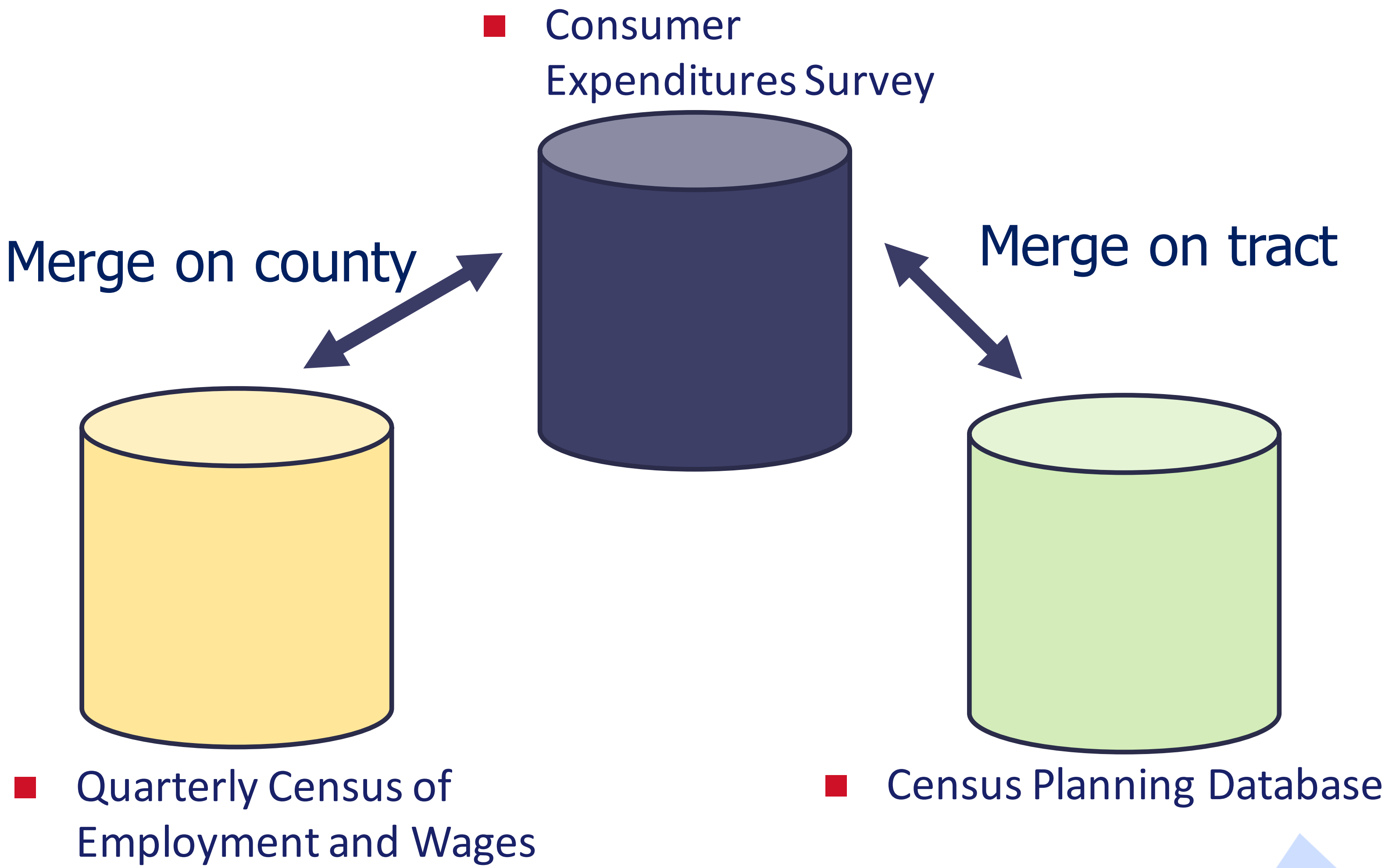
$$DIFF(y,\widehat{M}) = \sum_{k \in U} \widehat{m}(x_k) * N_k + \sum_{k \in S} \frac{y_k - \widehat{m}(x_k)}{\pi_k}$$

1. Predicted Expenditures (m)
2. Number of HH (N) in the tract (i)
3. Reported Expenditures (y)
4. Selection probability (π)
5. Survey correction

Why Use MAEs?

- Best of both worlds!
  - Unbiased estimate (if either term is unbiased)
  - More precise than just the survey estimate
- Doesn't depend too much on $\widehat{m}$
  - Breidt and Opsomer (2017) show a range of Machine Learning models can work for this

Auxiliary Data Used



Auxiliary Data Continued

Dataset	N. Obs.	N. Vars.	Unit of Observation
CEQ 2017	29,872	N.A.	Consumer Unit
CEQ 2018	28,244	N.A.	Consumer Unit
CEQ 2019	26,462	N.A.	Consumer Unit
CEQ 2020	25,087	N.A.	Consumer Unit
PDB 2019	72,893	124	Census Tract
PDB 2020	72,893	124	Census Tract
PDB 2021	72,893	124	Census Tract
QCEW 2017	3,190	44	U.S. County
QCEW 2018	3,191	44	U.S. County
QCEW 2019	3,191	44	U.S. County
QCEW 2020	3,192	44	U.S. County

Models Explored

- Models
  - Gradient Boosting Machines
  - Lasso
  - K-Nearest Neighbors
- Evaluation metrics
  - Cross-validation RMSE
  - Comparison to existing estimates

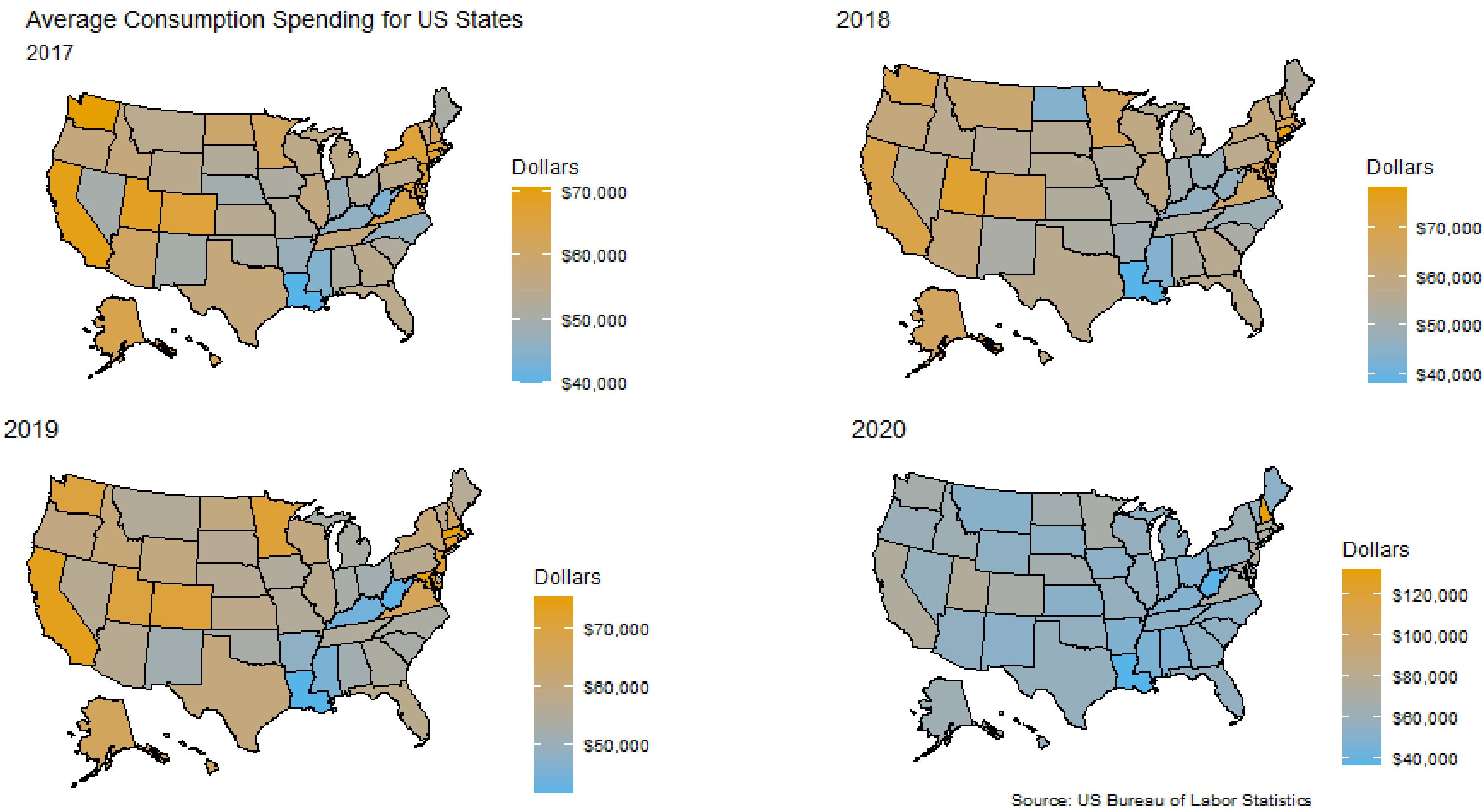
Cross-Validation Errors

2017	5-fold Cross-Validation RMSE					
Model	Total	Food	Housing	Transport	Health	Entertain
GBM	**$13,251.65**	**$1,182.30**	**$3,587.82**	$5,590.90	**$1,346.13**	$2,075.71
Lasso	$14,004.36	$1,320.54	$3,957.22	**$5,580.79**	$1,445.78	**$2,068.02**
KNN	$13,551.45	$1,219.20	$3,661.10	$6,307.92	$1,396.20	$2,380.56

2018						
Model	Total	Food	Housing	Transport	Health	Entertain
GBM	**$11,299.43**	**$1,263.33**	**$3,585.88**	$5,679.56	**$1,358.37**	$2,580.32
Lasso	$12,479.09	$1,446.52	$3,972.41	**$5,661.81**	$1,469.81	**$2,574.83**
KNN	$11,639.90	$1,297.21	$3,693.67	$6,458.11	$1,414.44	$2,904.76

2019						
Model	Total	Food	Housing	Transport	Health	Entertain
GBM	**$11,435.33**	**$1,337.12**	**$3,675.72**	$5,777.95	**$1,510.45**	**$1,789.47**
Lasso	$12,433.45	$1,502.79	$3,928.52	**$5,761.16**	$1,615.87	$1,795.59
KNN	$11,860.80	$1,380.31	$3,837.14	$6,569.00	$1,569.00	$1,992.28

Results


Average Consumption Spending for US States
2017
Dollars
$70,000
$60,000
$50,000
$40,000
2018
Dollars
$70,000
$60,000
$50,000
$40,000
2019
Dollars
$70,000
$60,000
$50,000
2020
Dollars
$120,000
$100,000
$80,000
$60,000
$40,000
Source: US Bureau of Labor Statistics

State Weights Comparison

	Total	Food	Housing	Transport	Health	Entertain
California						
2017	107.62%	104.64%	107.75%	113.14%	110.17%	110.17%
2018	104.88%	101.37%	107.19%	105.78%	103.94%	114.34%
2019	107.85%	103.83%	111.99%	112.95%	104.68%	113.00%
Florida						
2017	107.62%	100.54%	107.29%	116.21%	106.49%	143.19%
2018	105.50%	100.52%	107.57%	116.70%	111.17%	103.90%
2019	101.66%	98.70%	103.12%	115.58%	105.75%	98.85%
New Jersey						
2017	90.29%	93.52%	91.38%	87.95%	92.40%	89.57%
2018	93.57%	95.06%	93.38%	104.06%	101.43%	106.38%
2019	97.31%	100.16%	96.54%	101.20%	97.91%	102.36%
New York						
2017	111.40%	101.23%	103.51%	115.79%	99.98%	113.55%
2018	97.81%	96.33%	98.59%	108.40%	94.23%	95.10%
2019	98.89%	102.11%	100.06%	103.91%	103.33%	94.65%
Texas						
2017	103.48%	100.94%	104.16%	105.21%	106.34%	99.99%
2018	99.76%	100.80%	99.81%	102.52%	99.79%	100.85%
2019	99.05%	101.93%	101.78%	98.43%	97.91%	108.19%

Limitations

- Models aren’t very accurate (high RMSE)
- High year-to-year volatility (weird results)
- Lack of auxiliary data
- We didn’t calculate variances

Contact Information

U.S. Bureau of Labor Statistics
Division of Consumer Expenditure Surveys
www.bls.gov/cex