USDA Economic Research Service
U.S. DEPARTMENT OF AGRICULTURE

Generating Synthetic Data for the National Household Food Acquisition and Purchase Survey: A Case Study

Joe Rodhouse, USDA Economic Research Service
Jingchen (Monika) Hu, Vassar College

FCSM 2024 Research and Policy Conference
October 24, 2024

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

FoodAPS

- The National Household Food Acquisition and Purchase Survey (FoodAPS).
  - First nationally representative study of household food purchases and acquisitions.
  - Fills a critical data gap that informs policymaking on key national priorities, including…
    - Health and obesity, hunger, and nutrition assistance

FoodAPS Data Collection

- Fielded April 2012 – January 2013.
- Total of 4,826 responding households.
- Sampling based on household income and participation in the Supplemental Nutrition Assistance Program (SNAP).
- Oversample of low-income households and SNAP participants
- Study with multiple components: screener, initial interview, 7-day food acquisition diary, final/closing interview
- Respondents asked to provide detailed information on household composition, income, program benefits, and food acquisitions and food security

FoodAPS Utility

- FoodAPS is designed to support research about:
  - Socioeconomic factors that impact food access, food choices, food security, and health
  - The impact of foods acquired through USDA and other food and nutrition assistance programs
  - Interrelationships between food acquisitions, food demand, and well-being
  - Enables geographic/mapping research to study local food environments
- Detailed information about respondents is collected to enable research that can be used to design/improve food policy.
- ERS and external researchers are actively involved in rigorous research, using FoodAPS data to examine food demand relationships that previously could not be investigated in detail because the requisite data did not exist.

FoodAPS Data Access

- FoodAPS has public-use data (PUD)
  - https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/
  - Public use files stripped of data that pose a possible disclosure risk
- Access to restricted-use data (RUD) possible, but can be lengthy
  - https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/data-access/
  - Research projects using restricted-use data must be approved before access can be granted
  - Final research product must pass disclosure review before release/distribution

Utility of FoodAPS PUD

- FoodAPS PUD can support a wide variety of high-quality research
- However, several important research variables, such as individual income, are suppressed
- Hypothetical Use Case:
  - A researcher or analyst that wants to use FoodAPS to research how often people acquire food and whether that is impacted by their income and other demographic information.
    - Not possible with the public-use data
    - While they can apply for access, perhaps this route is untenable (e.g., a policy analyst who needs to answer analysis questions quickly but doesn't have access to RUD)
    - However, income can be a disclosure risk, good to suppress it
  - Can we increase the utility of FoodAPS PUD without compromising confidentiality and analytical validity?

Synthetic Data: A Viable Alternative

- Generating a synthetic version of income (and other identifying variables) could enhance the research utility of the FoodAPS PUD.

- What is synthetic data?
  - Related to imputation for nonresponse
  - Instead of missing values -> real values are replaced with plausible values
  - Can create fully or partially synthetic datasets

- Whereas traditional statistical Disclosure Control (SDC) methods may reduce analytical validity, synthetic data allows all variables to remain available
  - Suppression -> nonignorable missing information; renders some analyses undoable
  - Recoding -> loss of info in tails; spatial analysis reduced; ecological fallacies

Synthetic Data: Pros and Cons

- Generating and Releasing Synthetic Datasets to the public has several advantages:
  - Respondent identification is virtually impossible
  - All variables, including sensitive ones, can be fully available
  - Valid inferences can be obtained
- Disadvantages:
  - Strong dependence on modeling
  - May require heavy investment in time and resources
  - Not every variable is a good candidate to synthesize from an analytical validity standpoint (e.g., fixed state-level program benefit amounts)

Synthetic Data Generation Case Study

- What if we wanted to make a fully synthetic dataset that allows data users who only have access to the FoodAPS PUD to analyze how income impacts frequency of food acquisitions?

- Synthpop package in R, synthesizing using the CART method
  - Synthesize individual-level vars sequentially as: sex, age, race, education level, marital status, employment status, last month's income, and number of days during the week they have at least one food acquisition event.
- Case Study Objectives:
  - Observe the original and synthetic data distributions for individual income
  - Observe the relationship between income and # days w/ food acquisitions using original and synthetic data
    - Would similar conclusions about relationship (or lack thereof) be found?
  - Observe disclosure risk of the generated synthetic data

Synthetic Data Generation Method Results

- Use synthpop() in R to synthesize original variables, method = CART, number of iterations (m) = 10
- No records in the original data and synthetic data match; respondent re-identification with the synthetic data not possible.
- The distributions between the observed values and synthetically generated values appear similar
- The one exception being max values for last month's income (Orig value significantly higher than the Syn)

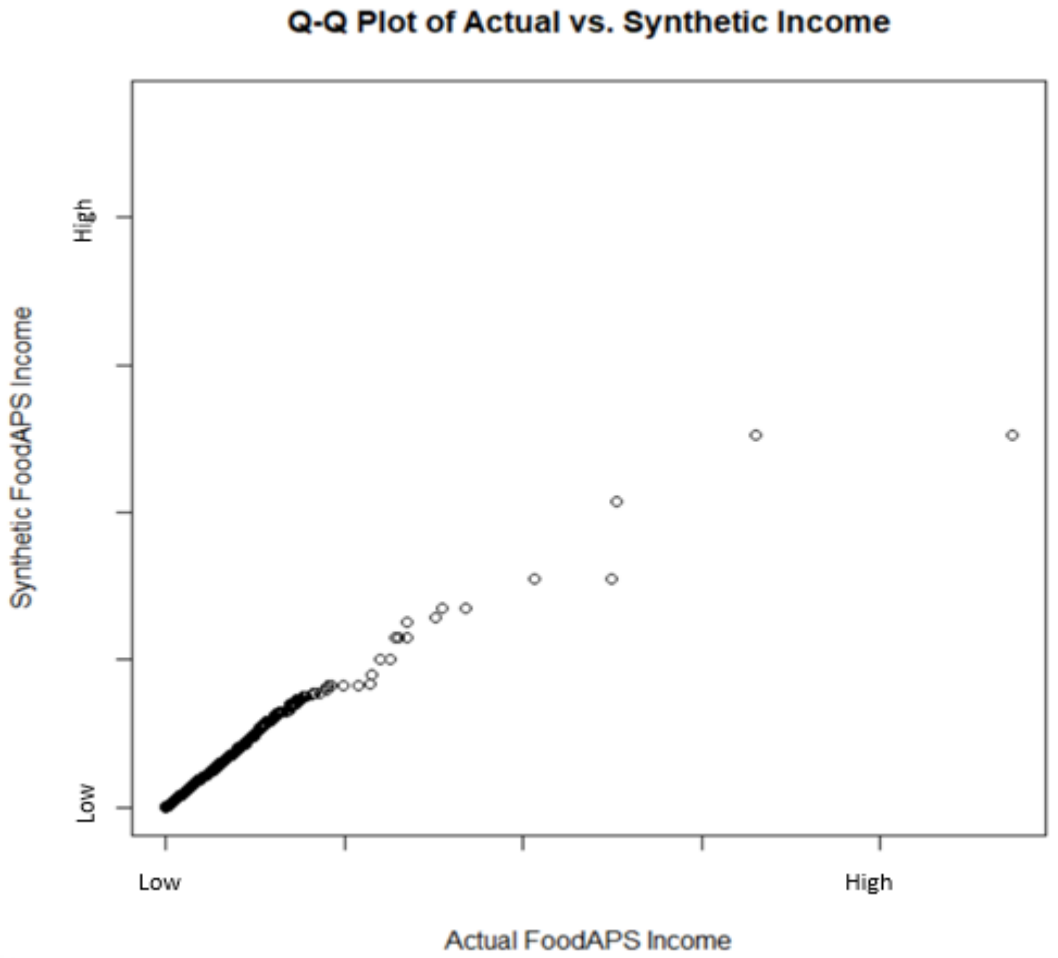
Income			Number of Days w/ Food Acquisitions		
Value	Original	Synthetic	Value	Original	Synthetic
Min.	0	0	0	2238	2272
1st Qu.	0	0	1	1854	1843
Median	>900	1000	2	1827	1800
Mean	1620	1586	3	1808	1778
3rd Qu.	>2000	2123	4	1860	1910
NA's	4070	4156	5	2016	1989
			6	1541	1559
			7	1173	1166

Synthetic Data Generation Methods Results

- Simple analysis: Average Reported Income for last month
- Get the point and variance estimate for m=10 synthetically generated datasets
- Compute final point and variance estimates and compare to the original data

	Lower 95% CI	Mean Income	Upper 95% CI
Original	1577.24	1620.00	1662.77
Synthetic	1598.59	1615.16	1631.74

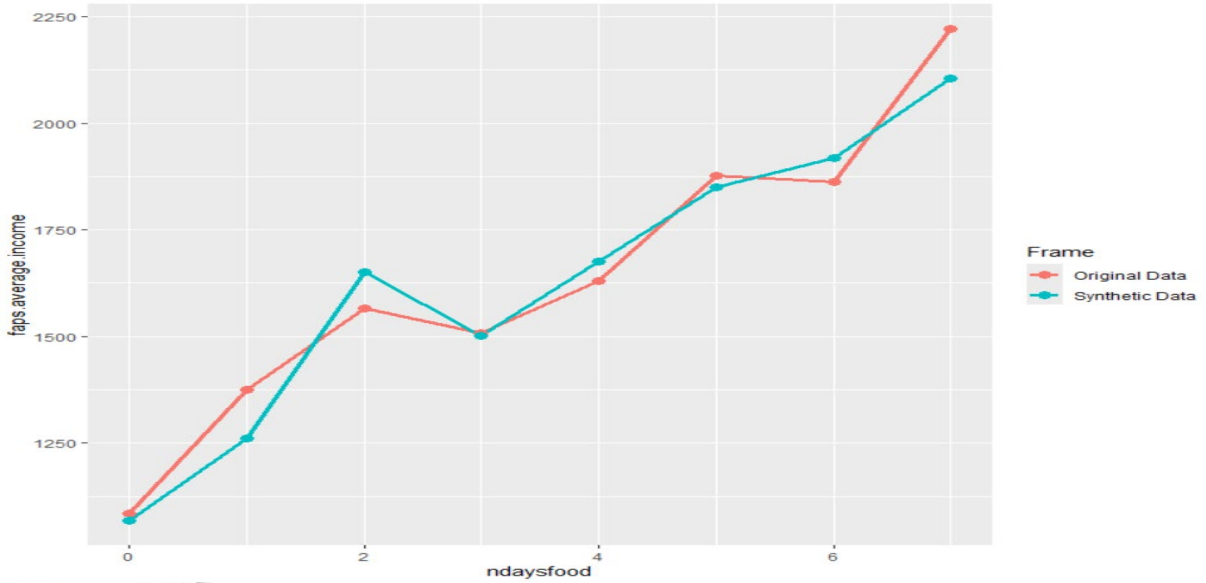
- Estimates look similar, but why are synthetic CI's shorter? Outliers -> Plausible values



Synthetic Data Generation Methods Results

- More complex analysis: Compute average income by Ndays w/ food acquisition and conduct ANOVA test.
- Estimate Ndays w/ food acquisitions with original and synthetic values controlling for income, sex, age, race, education level, and marital status.
- Average income seems similar, both ANOVA tests suggest relationship exist
- Controlling for other demographics, both regressions yield positive and significant coefficients of income (output in Appendix)

Number of Days w/ Food Acquisitions	Average Income	
	Original	Synthetic
0	1074	1058
1	1367	1257
2	1566	1643
3	1501	1498
4	1630	1677
5	1865	1842
6	1858	1908
7	2219	2093
F-value (ANOVA)	20.42	21.9
p-value	< .0001	< .0001



Synthetic Data Summary

- To enhance the utility of FoodAPS public-use data, synthetic data must be plausible, maintain relationships between vars found in original data, and not present a disclosure risk.
- No matches were found between respondent records in the original data and synthetic data, meaning the probability of disclosing a respondent's identity with the synthetic data is virtually zero.
- These preliminary findings are promising, and more utility checks of the synthetic data are planned.
- The potential benefit to the public data users of FoodAPS behooves further engagement in the domain of synthetic data research and generation.

Future Directions

- Cooperative agreement between ERS and Dr. Hu (Vassar College) began October 2024.
- Goal is to implement state-of-the-art methods to generate statistically valid synthetic FoodAPS datasets that meet the criteria needed to be released in public-use files.
- Develop methods that are optimal for the nuances and intricacies of FoodAPS and important economic and policy research, such as food spending, food security, healthy eating indexes, and geography/food environments.
- Develop methods that are optimal for future FoodAPS data releases
  - For example, OMB's revised Statistical Policy Directive 15 requires more robust and standardized race and ethnicity questions, but possible there could be disclosure concerns.
- Perhaps synthetic data could be a viable solution.

Joe Rodhouse
USDA Economic Research Service
Research Survey Statistician

Joseph.Rodhouse@usda.gov

Appendix

Poisson Regression Predicting Number of Days of Food Acquisitions				
	Original Data		Synthetic Data	
	Estimate Direction	p-value	Estimate Direction	p-value
(Intercept)	8.95E-01	<.0001	8.88E-01	<.0001
income	+	<.0001	+	<.0001
sex2	+	<.0001	+	<.0001
age	-	<.0001	-	<.0001
racecat2	-	<.0001	-	<.0001
racecat3	+	0.4838	+	0.6075
racecat4	-	<.0001	-	<.0001
racecat5	-	0.0334	-	0.3376
racecat6	-	0.0069	-	0.0661
racecat7	+	0.0329	+	0.0495
as.numeric(educ)	+	<.0001	+	<.0001
marital2	-	0.01297	-	<.0001
marital3	-	0.4574	-	<.0001
marital4	+	0.3982	+	0.8489
marital5	-	<.0001	-	<.0001