Tool

CARTER C. PRICE, KATHERINE OBENSCHAIN, AARON STRONG, ASHLEY DANCER, KAMI EHRICH, JAMES FAN, JOSHUA KLARICH, HSIN SHENG MA, AKSHAYA SURESH, MADISON WILLIAMS, VIRGINIA ZHANG

Federal Spending Mapping Tools

For more information on this publication, visit **www.rand.org/t/TLA4392-2**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our research integrity is grounded in RAND's core values of quality and objectivity. Rigorous quality assurance procedures, conflict of interest screening, and transparency in funding ensure that every study is objective and nonpartisan. Learn more at www.rand.org/integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2026 RAND Corporation
RAND® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit www.rand.org/about/publishing/permissions.

TL-A4392-2

About This Tool

To grasp the implications of federal spending for economic inequality, it is necessary to understand how the U.S. government distributes federal spending across the U.S. population. Although many studies have examined individual means-tested programs, such as food stamps or welfare, it is more challenging to study all federal spending in combination.

This document describes a set of tools to help map the majority of federal spending (more than 98 percent) to the people, communities, and businesses that receive those funds and the populations that benefit from the programs supported by that spending. By mapping spending on federal programs to those who receive the funds and benefit from them, we can greatly improve precision when estimating the effects of federal spending.

This document is written to be understood by a general audience, but we intend it primarily for quantitative analysts interested in the federal budget and distributional effects. The tools described here are intended for use by researchers and analysts familiar with either the R or Python programming languages. We can make the latest version of the computer code relevant to the tools available upon request. These are initial versions of the tools, and we intend to update them over time.

RAND von Furstenberg Family Budget Model Initiative

This work was conducted as part of the RAND von Furstenberg Family Budget Model Initiative, which developed tools for analysis of the U.S. tax code and federal spending and for forecasting the effects of proposed legislation. The RAND von Furstenberg Family Budget Model Initiative operates within RAND Education, Employment, and Infrastructure, a division of RAND that aims to improve educational opportunity, economic prosperity, and civic life for all. For more information, visit www.rand.org/budget-model or email EEI@rand.org.

Funding

The working paper version of this document was funded by the Builder's Initiative. Subsequent funding came from the RAND von Furstenberg Family Budget Model initiative. The RAND von Furstenberg Family Budget Model initiative is generously supported by the Diller-von Furstenberg Family Foundation and Pershing Square Philanthropies. Initial funding for the budget model was provided by the Nick and Leslie Hanauer Foundation.

Acknowledgments

We are grateful for the contributions of many of our colleagues and others to this work. We greatly appreciate insights from Kathryn Edwards on the early concepts for these tools. We would also like to thank Melanie Zaber of RAND and Martha Gimbel of Yale University for their reviews of early phases of this work, Federico Girosi of RAND and Steve Miller of the University of Colorado for their reviews of the working paper, and Vegard Nygaard of RAND for his review of this document. We are grateful to Stephanie Rennane and Heather Schwartz for their inputs as part of RAND's quality assurance process. Joshua Klarich provided superlative technical support for software engineering throughout this effort. Any flaws or errors in this document are the responsibility of the authors.

Summary

Representing more than 20 percent of U.S. gross domestic product (GDP),[1] federal spending plays an important role in the economy. The federal government spends money on national defense, to build infrastructure, to provide income support to the elderly, to cover medical spending for more than 100 million Americans,[2] and to provide a wide variety of other services for individuals, communities, and the general population. Some of this spending reaches beneficiaries directly, while other spending does so through indirect channels. The distribution of this spending across U.S. populations is key to understanding the effect that federal spending has on inequality and other economic outcomes of interest.

We developed tools to map federal spending to different segments of the U.S. population. These tools are part of the RAND von Furstenberg Family Budget Model Initiative, which aims to inform fiscal policymaking by incorporating analyses of the distributional implications of federal spending. By mapping spending on federal programs to those who receive the funds and benefit from them, we can greatly improve precision when estimating the effects of federal spending. Similarly, by more closely examining where federal spending actually goes, economic analysts can more precisely measure the effects of policies. The goal of the RAND Budget Model is to improve policymaking by providing more-detailed and more-granular information about the effects of policies. These mapping methodologies advance this goal by connecting federal spending to its beneficiaries.

This is the first version of the tools, and later iterations will include additional capabilities. Readers can request our computer code. The datasets are too large to cost-effectively host on RAND's website at this time, but we direct readers to the original datasets throughout this work. Consistent, well-defined data are as vital for policy research as for any other field, but the relevant data are frequently scattered across a variety of sources. This type of scattered data can lead to inconsistencies and limit the usefulness of the data. This document details the methodologies for creating an open-source dataset using public sources to present a comprehensive look at the U.S. population. Such a dataset is a valuable resource for policymakers, researchers, and modelers. This work should benefit research beyond the RAND Budget Model studies.

[1] U.S. Office of Management and Budget and Federal Reserve Bank of St. Louis, "Federal Net Outlays as Percent of Gross Domestic Product (FYONGDA188S)," August 5, 2024.

[2] Katherine Keisler-Starkey, Lisa N. Bunch, and Rachel A. Lindstrom, *Health Insurance Coverage in the United States: 2022*, U.S. Census Bureau, September 2023.

Contents

About This Tool iii
Summary v
Figures and Tables viii

Chapter 1. Introduction 1
How We Map Federal Spending to Population Segments 2
Limitations 2
Organization of This Document 3
Chapter 2. How We Collected and Categorized Federal Spending Data 4
USAspending.gov Provides Comprehensive Federal Expenditure Data 4
How We Map Spending to Population Characteristics 6
Key Data Limitations That Affect Analysis Quality 11
Chapter 3. How We Combined National Survey Data to Profile Program Recipients 12
How We Prepared the Survey Data for Analysis 14
Creating Demographic Groups for Statistical Matching 14
How We Match Cells Across Datasets to Ensure Consistency 16
How We Merge Records While Preserving Income Distribution 17
Key Challenges in Matching Survey Records Across Datasets 19
Chapter 4. How We Integrated Spending Data with Our Other RAND Budget Model Datasets 21
Category A: Individual-Related Federal Spending 22
Category B: Business-Related Federal Spending 26
Category C: Nontargeted Federal Spending: Infrastructure, Defense, and Debt Service 31
Chapter 5. How We Estimate Knock-On Effects of Spending 35
Chapter 6. Examples of How to Use These Federal Spending Data to Estimate Economic Impact 38
Example 1: Food and Nutrition Programs Reveal Complex Distribution of Federal Aid to Beneficiaries 39
Example 2: School Lunch Program Payments Flow Primarily to Food-Processing Companies, Not Direct Recipients 41
Example 3: SNAP Benefits Generate Broader Economic Impacts Beyond Direct Recipients 42
Additional Applications Show Tool's Broader Value 46
Chapter 7. These Tools Enable Better Federal Spending Analysis 48
Three Key Improvements Will Enhance Tool Capabilities 48

Appendix A. Overview of Relevant USAspending.gov Data Elements 50
Appendix B. Methods Used to Merge Demographic Data with Federal Spending Records 52
Variables in the CPS and the ACS for Cell Construction and Merging 52
How We Calculate Distances Between Demographic Groups Using Gower's Method 55
Using Silhouette Scores to Validate Cluster Quality 55
Why We Selected PAM as Our Clustering Algorithm 56
Appendix C. Interactive Heat Map 57
The U.S. Department of Agriculture Price Support Program Shows Regional Funding Disparities 57
Highway Construction Funds Concentrate in Urban Areas 58
Appendix D. Version History 60

Abbreviations 61
References 62

Figures and Tables

Figures

Figure 3.1. Cell Quality in the CPS and the ACS 17
Figure 3.2. Illustrative Example of the Merging Approach to Create the Synthetic Dataset 19
Figure 6.1. Food Assistance Spending by Program and State, FY 2023 41
Figure 6.2. National School Lunch Program Spending by Industry and State, FY 2023 42
Figure 6.3. Payments to Industry from SNAP 43
Figure 6.4. Payments to Households by SNAP 43
Figure 6.5. Per Capita Payments to Households from SNAP 44
Figure 6.6. Total GSP of SNAP Payments to Industry 45
Figure 6.7. GSP Impacts of SNAP Payments to Households 46
Figure C.1. Interactive Heat Map of the Farm Service Agency's Price Support Program 58
Figure C.2. Interactive Heat Map of the Federal Highway Administration's Spending 59

Tables

Table 2.1. Key USAspending.gov Data Variables 9
Table 2.2. Attribute Algorithm Examples 10
Table 3.1. Variables Used for Cell Creation 14
Table 4.1. Mapping Spending to RAND Budget Model Datasets and Models 21
Table 4.2. Spending Type Variables for Category A 22
Table 4.3. Recipient Variables 24
Table 4.4. Quantity Variables 24
Table 4.5. Mapping Approaches 25
Table 4.6. Spending Type Variables 27
Table 4.7. Recipient Variables for Category B 28
Table 4.8. Quantity Variables for Category B 30
Table 4.9. Mapping Approaches 31
Table 4.10. Examples of Spending Variables for Category C 33
Table 4.11. Examples of Obligations and Outlays for Category C 34
Table B.1. Comparison of Variables That Could Be Used to Match the CPS to the ACS or the SCF 52
Table D.1. Version History 60

Chapter 1. Introduction

Representing more than 20 percent of U.S. gross domestic product (GDP),[3] federal spending shapes the economy. The federal government spends money on national defense, to build infrastructure, to provide income support to the elderly, to cover medical spending for more than 100 million Americans,[4] and to provide a wide variety of other services for individuals, communities, and the general population. Some of this spending reaches beneficiaries directly, while other spending does so through indirect channels. Understanding the distribution of this spending across populations is key to understanding the effect that federal spending has on inequality and other economic outcomes of interest.

This document describes our approach to mapping federal spending to relevant segments of the U.S. population. We first process federal spending data from USAspending.gov. We clean and harmonize the data (described in Chapter 2). We use these data to characterize each federal spending item's recipients (entities such as individuals, firms, and places that directly receive the funds) and beneficiaries (populations targeted by the programs supported by the funds). We then identify the relevant segments of the U.S. population based on the key attributes of the recipients and beneficiaries, such as demographic characteristics, geography, and industry. Finally, we integrate these spending data with a synthetic dataset that we construct from nationally representative surveys—including the Current Population Survey (CPS),[5] American Community Survey (ACS),[6] and Survey of Consumer Finances (SCF)[7]—using characteristics of individuals, their employers, and their communities. The resulting dataset provides a more complete picture of the U.S. population that can be used to better inform policymaking. Instead of a one-off matching that is typically used for policy analysis, this approach is general purpose, transparent, and easily reproducible. Additionally, integrating several sources of data allows us to capture the distributional effects of federal spending across a variety of demographic, economic, and geographic dimensions. These tools are part of the RAND von Furstenberg Family Budget

[3] U.S. Office of Management and Budget and Federal Reserve Bank of St. Louis, "Federal Net Outlays as Percent of Gross Domestic Product (FYONGDA188S)," August 5, 2024.

[4] Katherine Keisler-Starkey, Lisa N. Bunch, and Rachel A. Lindstrom, *Health Insurance Coverage in the United States: 2022*, U.S. Census Bureau, September 2023.

[5] IPUMS CPS, homepage, University of Minnesota, undated; Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry, "IPUMS CPS," Version 11.0, dataset, IPUMS, University of Minnesota, 2024.

[6] Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler, "IPUMS USA," Version 15.0, dataset, IPUMS, University of Minnesota, 2023.

[7] Board of Governors of the Federal Reserve System, "2022 Survey of Consumer Finances," dataset, 2023.

Model Initiative, which aims to inform fiscal policymaking by incorporating analyses of the distributional implications of federal spending.

Consistent, well-defined data are as vital for policy research as for any other field, but the relevant data are frequently scattered across a variety of datasets. Sometimes data elements are collected in multiple datasets, but the results are not consistent. In other cases, the aggregation of data simplifies them for certain uses but restricts applicability for more-detailed or more-specific purposes. The analysis in this document is an initial step to creating an open-source dataset built from publicly available datasets that presents a consistent and comprehensive look at the U.S. population. This type of dataset is as important for policymakers when seeking to understand the ground truth as it is for modelers seeking to accurately represent the world. Thus, it is our hope that the work will be useful beyond just the RAND Budget Model project.

How We Map Federal Spending to Population Segments

We first processed federal spending data from USAspending.gov.[8] This data source provides detailed and timely information about each federal spending item, including the award date, dollar amount, funding agency, purpose, recipient, beneficiary, and the locations and industries of the recipient and beneficiary. We clean and harmonize the data and characterize recipients and beneficiaries (described in Chapter 2).

Although USAspending.gov data include relevant details about recipients and beneficiaries, these data lack all necessary information about the demographics and socioeconomic characteristics of target populations. We address this limitation by constructing a synthetic population dataset, which we use to match spending data to different segments of the U.S. population based on the characteristics of individuals, their employers, and their communities. As there is not a single, authoritative dataset with all necessary data elements, we develop a method to statistically match data from several nationally representative surveys, including the CPS, ACS, and SCF. Finally, we link the spending information to the synthetic dataset based on key demographic characteristics, geography, and industry to describe how federal spending is distributed throughout the economy.

Limitations

The datasets we selected represent the best publicly available sources, but they have several limitations. The USAspending.gov data do not always provide consistent details about the recipients of the funds, and those details do not always include the same information as the surveys. Similarly, the matching of the surveys is limited by the common factors in the datasets. The implication is that there is some fundamental uncertainty in the estimates.

[8] USAspending.gov, "SQL Database Downloads," webpage, undated-d.

Additionally, this is the first version of the tools. Future versions will improve on the approach as we include additional datasets and refine the approach based on comments and criticism.

Organization of This Document

Chapter 2 describes the processing of the spending data. Chapter 3 describes the processing of the data from nationally representative surveys and the development of a statistical matching approach to construct a synthetic population dataset. Chapter 4 discusses the process for integrating the USAspending.gov data with the synthetic dataset. In Chapter 5, we describe how these data can be used for macroeconomic analysis using an input/output (I/O) model, and in Chapter 6 we present specific examples of using these data in I/O model to explore the macroeconomic effects of specific spending programs. Chapter 7 discusses practical use cases for these data and a roadmap for future directions of these tools. Appendix A describes more data elements from the USAspending.gov dataset that may be relevant for mapping spending to the synthetic dataset. Appendix B describes technical details for the statistical matching process. Appendix C presents the interactive heat map with the geographic distribution of spending information. Appendix D contains the version history for this document.

Chapter 2. How We Collected and Categorized Federal Spending Data

We outline the process of collecting and processing federal spending data for further analysis. Our principal data source is USAspending.gov's federal spending database. We first provide some background on USAspending.gov and its data contents. Then, we describe how we prepare these data for integration with the synthetic population dataset. We conclude with a few examples.

USAspending.gov Provides Comprehensive Federal Expenditure Data

At the core of our analytical framework is the comprehensive federal spending data available through USAspending.gov. This repository contains detailed information on nearly all federal expenditures.[9] The USAspending.gov website was created by the U.S. government in 2001 to improve federal spending transparency. Today, most federal agencies must report spending data to USAspending.gov under the Digital Accountability and Transparency (DATA) Act of 2014, codified in 31 U.S.C. § 6101.[10] The site provides various resources and interactive tools to review and analyze federal spending.

USAspending.gov's Strengths and Limitations for Tracking Federal Funds

Government agencies have historically struggled to track spending, given the number of different agencies, procedures, and transactions, among other variables. Despite advances, such as the DATA Act, selecting an appropriate data source or set of data sources still requires weighing several pros and cons for each option. The most-relevant advantages and disadvantages of USAspending.gov are listed below.

The advantages are as follows:

- **Authoritative:** USAspending.gov is the U.S. government's official open data source for federal spending. As the official source, USAspending.gov is subject to reviews by other government entities and private organizations.
- **Frequently updated:** Agencies update USAspending.gov data monthly or quarterly. Federal agencies must submit financial data to the site every month or quarter. The specific data resource on the site used in the tool is updated every month to reflect the

[9] For any given release, a small number of Treasury Account Symbols may be missing, and there may be a small number that are unlinked. For more details, see USAspending.gov, "Using This Table," webpage, undated-e.

[10] Agencies with small budgets might not report their spending information through USAspending.gov. See Bureau of the Fiscal Service, U.S. Department of the Treasury, "About the Data Transparency Program," webpage, last updated November 16, 2023; U.S. Code, Title 31, Section 6101, Definitions.

most recent agency submissions. These frequent updates enable near-real-time analysis of federal spending changes.

- **Consolidated inputs:** USAspending.gov merges inputs from a variety of sources that would be burdensome to collect and process separately. The sources include
  - Federal Procurement Data System (FPDS)
  - System for Award Management (SAM)
  - Government Treasury Account Symbol Adjusted Trial Balance System (GTAS)
  - agency financial systems.
- **Rich in data:** With more than 400 data elements, USAspending.gov contains detailed information on each spending item, pooled from various government systems. Information is available for awarding entities and recipients, as well as for overhead spending.

USAspending.gov has these key limitations:

- **Missing data:**
  - The DATA Act does not require some smaller agencies to report to USAspending.gov, and some required agencies report inconsistently.[11] Spending for these missing agencies will need to be sourced outside USAspending.gov in our tool.
  - Agencies fail to report some federal spending each month. This can be due to a late entry by an agency or some error. For example, USAspending.gov documented that 1.6 percent of spending data were unreported in the first two quarters of fiscal year (FY) 2024.[12]
  - In other cases, the dataset lacks the necessary information to link spending to other sources. Multiple data sources create redundancy in the dataset. Data cleaning can address many of these concerns.
- **Unreliable subaward data:**[13] The U.S. Government Accountability Office (GAO) has reported concerns with the quality of the subaward data published on USAspending.gov's website. GAO documented duplicative, incomplete, or incorrect records.[14] Given this issue, in the first phase of this project, we use only parent award records.

We explored several other data sources. Directly sourcing from SAM, FPDS, and GTAS would give us greater control over the data collection methods but would ultimately be duplicative of the USAspending.gov product. We also considered using the Federal Audit Clearinghouse.[15] Although this source has a wealth of high-quality information, the inconsistent

---

[11] U.S. Government Accountability Office (GAO), *Federal Spending Transparency: Opportunities to Improve USAspending.gov Data*, November 7, 2023a.

[12] USAspending.gov, "Explore the Spending Landscape," webpage, undated-c.

[13] A subaward is when the primary recipient of funds transfers some (not all) of those funds to another entity in return for the completion of a portion of the work, similar to a subcontract.

[14] GAO, *Federal Spending Transparency: Opportunities Exist to Improve COVID-19 and Other Grant Subaward Data on USAspending.gov*, November 16, 2023b.

[15] Federal Audit Clearinghouse, homepage, undated.

reporting time frames between different state and local governments and the private sector would make analysis challenging. Finally, we reviewed USAFacts as a possible data source.[16] However, outside its annual report, the site releases data updates at unpredictable intervals. The annual report does not provide the detail necessary for this project's goals.

We selected USAspending.gov for its timeliness and efficiency. Workarounds, such as supplementing the dataset with other sources to address incomplete data, exist for its most concerning disadvantages. Additionally, USAspending.gov's accuracy has improved over recent years and is on a positive trajectory.[17]

How We Collect and Process USAspending.gov Data

For this project, we download the entire USAspending.gov database (amounting to several terabytes), made available to the public as a PostgresSQL archive.[18] We refresh our download monthly, keeping pace with updates by USAspending.gov. We prune and modify this comprehensive database to meet the specific needs of this project. The database contains detailed information on spending items, including outlays, obligations, and account balances. Each entry in the database represents a single spending item or payment made by the federal government to a particular recipient. We characterize each spending item to prepare for its linkage to the synthetic dataset.

Appendix A provides a list of relevant data elements in the USAspending.gov dataset to categorize spending.

How We Map Spending to Population Characteristics

We map each spending item to several attributes describing the populations that receive or benefit from it. Then, as we discuss in the next chapter, we use these key attributes to generate a representative population in a synthetic dataset. These population attributes include group characteristics (e.g., age, veteran status, income level), geography (e.g., city, state, urban), and industry of employment (i.e., North American Industrial Classification System [NAICS] code). See Table 3.1 in Chapter 3 for more information on these attributes.

How Federal Dollars Flow: From Direct Recipients to Ultimate Beneficiaries

For each spending item, USAspending.gov data fields help us identify who receives and benefits from each spending item. We then characterize the key attributes of recipients and

[16] USAFacts, homepage, undated.

[17] GAO, *Federal Spending Transparency: Opportunities Exist to Further Improve the Information Available on USAspending.gov*, November 8, 2021.

[18] USAspending.gov, undated-d.

beneficiaries in terms of demographic, socioeconomic, and geographic attributes, which we later use to map government spending to the synthetic population dataset.

Recipients are populations who directly receive the spending item. *Beneficiaries* are populations intended to benefit from the spending through direct receipt or other channels. For example, Medicaid spending is directly made to health care providers (the recipients) to provide care to eligible low-income individuals (the beneficiaries). For some spending items, the recipient and beneficiary are the same, such as the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC); Social Security; and other direct-assistance programs. In other instances, the beneficiary is less clearly defined, such as with national security spending.

We use the detailed fields in USAspending.gov data, as well as program eligibility criteria for assistance awards, to identify key attributes of recipients and beneficiaries. Consider payments to a U.S. Department of Veterans Affairs (VA) hospital for hospital staff earnings. The recipients would be the health care workers and administrators who directly receive the payments, identified in the data by the field "object class" and the value "personnel compensation and benefits." The beneficiaries would be the veterans receiving care at VA hospitals, identified in the data by the field "budget function" and the value "veterans benefits and services."

Key Data Fields That Reveal Who Gets Federal Money

We use these key USAspending.gov fields to track recipients and beneficiaries:

- **Spending type:** There are three broad spending types: contract awards, assistance awards, and non-award spending.
  - *Contract awards* provide payments to recipients in exchange for services or goods to the government, such as infrastructure, IT services, and defense systems. The recipient is the contractor or service provider. The beneficiary depends on the nature of the program and its intended use but is likely a different entity.
  - *Assistance awards* provide financial aid to advance federal objectives, such as public health insurance, poverty reduction, disaster relief, and scientific research. These awards include grants, loans, direct payments, cooperative agreements, insurance, and technology investment awards. The recipient is frequently the beneficiary.
  - *Non-award spending* is other types of spending to support government operations, programs, and obligations, such as government employee salaries and debt service. The recipient and beneficiary may be the same entity, but the program associated with the payment will typically benefit some other population.
- **Object class:** Object classes categorize spending into one of 35 categories of spending "according to their initial purpose, not the end product or service."[19] Major categories are "rents and fixed charges," "contractual services and supplies," "personnel compensation

[19] U.S. Office of Management and Budget, *Preparation, Submission, and Execution of the Budget*, Circular No. A–11, 2023.

and benefits," and "acquisition of assets." For example, the system classifies payments to federal employees as "personnel compensation and benefits," regardless of whether the employee is building a new government structure, providing health care, or improving prekindergarten programs. Since the object class describes the initial purpose of funding, this field is helpful when assigning attributes to the recipients. In the VA example previously described, "personnel compensation and benefits" identifies that the funding is for hospital staff payments.

- **Assistance listing number:** The federal government maintains a comprehensive library of the federal programs that provide assistance awards, and each distinct program has its own assistance listing number. In the USAspending.gov database, all assistance award spending includes the associated assistance listing number. Available on SAM.gov, each program's listing documents the eligibility criteria.[20]
- **Treasury Account Symbol (TAS):** The TAS is a unique identifier assigned by the U.S. Department of the Treasury to track government financial transactions. This code includes an agency identifier and main account number, as well as a subaccount and fiscal year identifiers. In the USAspending.gov database, the TAS allows users to pinpoint the specific accounts from which agencies disburse funds by federal agencies.
- **Budget function:** Budget functions categorize spending based on purpose. There are 20 budget functions, each with one or more subfunctions. Functions include "national defense," "agriculture," "transportation," and "Medicare." Agriculture subfunctions include "farm income stabilization" and "agricultural research and services." Budget functions inform who the beneficiaries are, as they describe the aim of the funding beyond the direct recipients. In the previously described VA example, the beneficiaries fall under the budget function "veterans benefits and services."
- **Program activity codes:** These codes link spending to specific programs within government agencies or departments. For example, program activities within the U.S. Department of Education include "student financial assistance," "school improvement programs," and "higher education initiatives."
- **Recipient classification:** These codes categorize recipients by type. Types include state and local government, business types (small business, large business), nonprofit organizations, educational institutions, individual recipients, and tribal governments.
- **NAICS code:** The data provide the six-digit NAICS codes for recipients. The first two digits represent the major business sector, the third digit represents the subsector, the fourth digit represents the industry group, and the last two digits designate the industry. NAICS codes serve as a key attribute for recipients of contract award spending, enabling us to characterize the type of organizations and employees receiving and benefiting from the spending.
- **Place of performance:** The data include geographic details on a recipient's location: state, county Federal Information Processing Standard (FIPS), five-digit ZIP code, metropolitan area, and address. We observe the location of the recipient's place of performance, or "principal place of business, where the majority of the work is performed."[21]

[20] SAM.gov, homepage, U.S. General Services Administration, undated-b.

[21] USAspending.gov, "Data Dictionary," dataset, undated-b.

- **Action date:** The data provide the date and fiscal year when each spending item was issued or signed by the government or a binding agreement was reached.

Additional details about the key variables are in Table 2.1.

Table 2.1. Key USAspending.gov Data Variables

USAspending.gov Variable	Examples of Values	Purpose
Spending type	• Assistance award • Contract award	Identify financial assistance awards vs. other types
Budget function	• Medicare (570) • Income security (600) • Social security (650) • Veterans benefits and services (700) • Disaster relief and insurance (453)	Classify spending by federal objective
Object class	• 25.1—Advisory and Assistance Services • 41.0—Grants, Subsidies, and Contributions	Classification system used by the federal government to categorize spending based on the type of goods, services, or activities purchased
TAS	• Agency Identifier: 86 (U.S. Department of Housing and Urban Development) • Main Account Code: 0300 (Community Development Block Grants)	Links spending to specific U.S. Treasury accounts
Recipient type	• Individuals • Small business • For-profit organization	Identify the type of entity receiving the spending
Recipient NAICS code	• 31–33 (Manufacturing) • 61 (Educational Services) • 92 (Public Administration)	Classify recipients by industry
Recipient place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport	Identify geographies where recipient's work is performed[a]
Primary place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport • Primary place of performance scope = multi-state, state-wide, county-wide, city-wide, single ZIP, foreign	Identify geographies where the majority of the recipient's work is performed

[a] Although the recipient place of performance accounts for the official mailing address of the funding recipient, it may differ from the recipient's primary place of performance. For example, a large firm may have its official headquarters in one state, but the actual funds will be spent at a production facility located in another state. It is important to understand both value variables and identify which one will be most accurate when estimating the impacts of any spending. It is likely that the primary place of performance is the most accurate when understanding where the funding is going. An associated variable to primary place of performance includes the scope, such as multi-state, state-wide, county-wide, or city-wide. That will also be useful for estimating wider impacts of the funds.

How We Use Data Fields to Track Federal Spending Impact

Depending on the spending item, we use a combination of the above fields as the building blocks for assigning key attributes to recipients and beneficiaries. As another example, the assistance award's budget function informs its target population, detailed through program eligibility criteria, program activity code, and geography. In some cases, the details of the target population for the program are documented in a record, but the record often lacks structure for easy extraction in an automated fashion (a future version may use some form of natural language processing to allow for supervised automation).

Table 2.2 provides three examples of how we map spending items from the raw data files to the synthetic population data using key attributes of recipients and beneficiaries. The table shows three different records from the spending data: Jennie-O Turkey Store, Inc., for the National School Lunch Program; Sundance, Inc., for the Paycheck Protection Loan Program; and salary payments to Army officers. In each case, there is a description of the recipient and beneficiary, as well as their key attributes. Although this is not enough information to say much about the details of the populations in question, in the next chapter, we will augment this information with survey data.

Table 2.2. Attribute Algorithm Examples

Spending Item: Initial Input Fields	Fields Used	Attributes
The U.S. Department of Agriculture's obligation to Jennie-O Turkey Store, Inc. • Spending type: contract award • Object class: supplies and materials • Budget function: food and nutrition assistance	Recipient: • Location: Willmar, MN • NAICS code: 311615, poultry processing Beneficiary: • Program activity: child nutrition programs, food and nutrition service	Recipient: • Location: Willmar, MN • Industry: poultry processing Beneficiary: • Group characteristic: income < 130% poverty line[a]
Loan to Sundance, Inc. • Spending type: assistance award • Object class: grants, subsidies, and contributions • Budget function: paycheck protection loan	Recipient/beneficiary: • Location: Brighton, MI • Assistance listing number: 59.073, Paycheck Protection Loan Program	Recipient/beneficiary: • Location: Brighton, MI • Group characteristic: small-business employees
Salary payments to active-duty Army officers • Spending type: non-award • Object class: military personnel • Budget function: U.S. Department of War—military[b]	Recipient: • No other fields needed Beneficiary: • No other fields needed	Recipient: • Active-duty Army officers Beneficiary: • General population

[a] For this example, we extracted the eligibility criteria text ("income < 130% poverty line") from this specific record manually, but in future iterations we hope to use some form of automation so that it can be done at scale.
[b] The Department of War is designated the Department of Defense under Public Law 81-216, National Security Act Amendments of 1949.

Key Data Limitations That Affect Analysis Quality

This spending analysis has several key limitations:

- Smaller agencies might not report spending, although missing data account for only 1.6 percent of total spending—a minor limitation for our purposes.
- The data do not include tax expenditures, including economic impact payments.
- Limited subaward data constrain our understanding of spending's geographic distribution. However, we may be able to impute some characteristics of the subawardees in a future version of the data analysis.
- Agencies update the data frequently, although not all programs are updated concurrently. This creates a trade-off between the completeness and the timeliness of the data. Because the initial applications of these data are retrospective rather than forecasting, we can use the prior quarter's data and be assured that these data are mostly complete. In future versions, we may interpolate or impute the current spending data based on prior trends.

Beneficiary details remain limited, so we need to use additional data to understand their characteristics. The next chapter describes how we construct a synthetic dataset that can be used to link spending to individual characteristics.

Chapter 3. How We Combined National Survey Data to Profile Program Recipients

We process and combine demographic data from nationally representative survey data. We use the March 2022 CPS Annual Social and Economic Supplement as the primary source of demographic data. The CPS collects detailed information about several types of income, employment details (such as sector, hours per week, and weeks per year), family characteristics, and demographic information from around 100,000 households.

We then use statistical matching to link records to this dataset. Statistical matching involves identifying and using similar variables in one or more datasets to create one larger, combined synthetic dataset.[22] Because of the dataset size differences and potential for missed matches, one-to-one matching is not possible; in most cases, statistical matching is done at the group level through the creation of cells with similar records, such that the distribution of characteristics across cells is maintained.

The approach described here is similar to the approach used in RAND's Comprehensive Assessment of Reform Efforts (COMPARE) microsimulation model.[23] COMPARE is a microsimulation model used to explore policy options related to health insurance changes. COMPARE uses the CPS as the core dataset for demographic and income information. However, the CPS does not have the detailed information about medical expenditures needed for a health insurance model, and so the COMPARE team used a cell-based statistical matching approach to link records from the CPS with demographically similar records in the Medical Expenditure Panel Survey (MEPS).[24] Records in the CPS and MEPS are grouped into cells based on such factors as age group, income group, health insurance status, and health status. The cells' sizes and characteristics are compared to ensure adequate consistency, and then records within cells are matched, accounting for the record weights, to produce a synthetic dataset that captures the relevant information from both the CPS and the MEPS.

For this project, we also used the CPS as the core dataset and statistically matched additional datasets as needed. For example, we combined the 2022 CPS with the ACS from the same year. This combination allowed us to capture the richer employment and income details available in

[22] Aura Leulescu and Mihaela Agafiţei, *Statistical Matching: A Model Based Approach for Data Integration*, Eurostat, European Commission, 2013.

[23] Federico Girosi, Amado Cordova, Christine Eibner, Carole Roan Gresenz, Emmett B. Keeler, Jeanne S. Ringel, Jeffrey Sullivan, John Bertko, Melinda Beeuwkes Buntin, and Raffaele Vardavas, "Overview of the COMPARE Microsimulation Model," RAND Corporation, WR-650, 2009.

[24] Agency for Healthcare Research and Quality, "Medical Expenditure Panel Survey," webpage, undated.

the CPS and the more-granular geographic details from the ACS.[25] The ACS includes detailed information on households' and respondents' city of residence, corresponding to Census Designated Places (jurisdictions) across the nation. Although the CPS has better coverage of the entire geography of the United States, it includes more-detailed geographic information for only a small sample of large cities. In contrast, the ACS captures both large and small cities nationwide. By merging the two datasets, we can create a more comprehensive and complex dataset with detailed geographic identifiers—including metropolitan areas, counties, and cities—attached to each demographic unit based on its characteristics.

For adults (respondents age 18 or older), we construct cells using state, urbanicity, race, marital status, sex, age group, presence of children in the household, employment status, education group, veteran status, disability status, and citizenship (see Table 3.1 for details). For minor cells (consisting of records for which respondents are under age 18), we exclude irrelevant variables, leaving state, urbanicity, race, sex, and citizenship. For cells that have fewer than five records,[26] we merge characteristics to create adequately sized cells while always preserving the state variable (i.e., we never merged on state). Because the income distribution is of central importance to our analysis, we order the records by income within each cell.[27] These ranked records are then matched and merged with records from a second dataset, accounting for the record weights, to produce a synthetic dataset. The rest of this chapter describes the details of the cell creation and merging.

[25] There is an important time component to consider when comparing these datasets. If the data come from different periods, there may be significant differences in time-sensitive factors. For example, depending on the rate of inflation, dollar-denominated data elements might not be directly comparable between datasets. This can be corrected using inflation adjustments, but that might not be feasible for all variables.

[26] The threshold of five records is somewhat arbitrary, but the rationale is to ensure that at least the 25th percentile, median, and 75th percentile could all be calculated for each cell. Technically, this can be done with a cell size of four, but in that case, there would not necessarily be a median record.

[27] Income here is the sum of wage and salary income, farm and nonfarm income, and income from dividend, rent, and interest, which are reported in both the CPS and the ACS. The intent of ranking by income is to preserve, as best as possible, the income distribution when matching records. We considered doing an alternative approach, such as adding income groupings to the set of characteristics used for creating cells. However, different programs have different income thresholds, making it difficult to identify the "best" income thresholds for grouping. For example, Medicaid expansion is applicable to people under 138 percent of the federal poverty level (FPL), but the Affordable Care Act's premium subsidies are available for those earning above 100 percent of the FPL (and below 400 percent) (see Medicaid and ACA eligibility at the time of writing in 2026; HealthCare.gov, "Medicaid Expansion & What It Means for You," webpage, undated; Louise Norris, "Will You Receive an ACA Premium Subsidy?" HealthInsurance.org, last updated January 27, 2026). So, when thinking about health care, 100 percent of FPL, 138 percent of FPL, and 400 percent might seem like reasonable thresholds given the underlying policies of interest. However, the population between 100 and 138 percent of FPL is relatively small and would frequently be merged regardless of intent. Thus, we decided for this initial version to use income to sort the records within cells and then, as will be discussed later, measure the degree to which the income distribution is preserved.

How We Prepared the Survey Data for Analysis

As stated previously, the CPS and the ACS are large-scale datasets that provide detailed socioeconomic and demographic information about the U.S. population. We process the relevant variables for cell creation (see Table 3.1) and create categorical buckets for variables, such as age, that are continuous.

Table 3.1. Variables Used for Cell Creation

Variable	Number of Subcategories
Employment	3 (employed, unemployed, non–labor force)
Education	4 (less than high school diploma, high school diploma, some college, college degree[s])
States	51 (all states plus Washington, D.C.)
Urbanicity	3 (0,[a] in metro, not in metro)
Sex	2 (male, female)
Married	2 (married = 1, single = 0)
Have kids	2 (true, false)
Age	3 (minor, working age, retired)
Race	5 (White, Black, Asian and Pacific Islander, American Indian, multiple)
Disability	2 (not disabled, disabled)
Veteran status	2 (yes, no)
Citizenship	2 (citizen, noncitizen)

[a] Zero indicates a missing response.

For variables with an overabundance of categories, we consolidate. For all sources of income used to calculate taxable income (labor income, capital income, and retirement income), we set to 0 all records with error or omitted-data codes (e.g., 99999, −999). Tax IDs are also created to determine family relationships within households (e.g., parentage, spouse in residence) and to determine whether an adult has at least one minor-age child. Annotated code for data cleaning is available upon request.

Creating Demographic Groups for Statistical Matching

As described above, we take a cell-based approach to merging the datasets, and this section describes how the cells are constructed. Because the ultimate intent of this work is to provide reliable mapping from spending to the beneficiaries and recipients, we construct cells to preserve

as much of the demographic and economic information as possible. The objective of cell construction is as follows:

- We start with records of individuals or households in a state.
- We create cells that have at least five individuals or households that share similar characteristics across 11 demographic variables; we define an *identity* as a unique combination of these 11 variables.[28] The key variables used in cell construction are education, marital status, urbanicity (in metro or not in metro), race, sex, age, indication of whether or not the individual or household has children, employment status, veteran status, disability, and citizenship (described in Table 3.1).[29]
- We maximize as many final cells as possible while ensuring that each cell meets the minimum size threshold of five records and that individuals or households within an identity are not split across cells.

We denote key characteristics, c^i, where the number of options under each characteristic is given by $|c^i|$. Note that $|c^i|$ represents the total distinct number of identities in the dataset. To keep individuals who are part of the same identity, the maximum possible number of cells is the same as the total number of identities. That is,

$$max\ C = |c^1| \cdot |c^2| \ldots |c^n|,$$

where n is the total number of characteristics included.

In our context, $n = 12$: the 11 demographic characteristics listed above plus the geographic state. In this context, $max\ C = 3 \cdot 3 \cdot 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 5 \cdot 3 \cdot 3 \cdot 2 \cdot 51 = 1{,}982{,}880$ total identities, or $38{,}800$ identities per state (including Washington, D.C.).

In practice, many identities will contain fewer than five individuals, so these identities require clustering during cell construction. There will also be a large number of empty identities, where no individuals have that set of characteristics in that dataset. Naturally, the number of empty identities will increase as the size of the dataset decreases. In our context, the CPS dataset has a total of 114,158 records, translating to 37,175 represented identities. Of these, 31,997 have fewer than five records and thus need to be clustered.

---

[28] The cells are constructed by the Partitioning Around Medoids (PAM) clustering method, described in more detail in Appendix B. This method results in cells that have the minimum amount of dissimilarity among individuals in the cell. The similarity of individuals within a cell depends on the identities represented in the cell and the geographic state. For example, in New York, where more-distinct identities are represented, it is more likely that any cell will have enough individuals that are exact matches on ten or more characteristics, but in a smaller-population state, such as Alaska, this is less likely. More-dissimilar individuals will need to be grouped together, so there will be more cells in which the individuals match on only a few characteristics.

[29] Note that the variables chosen for cell construction must exist in both the base dataset and the dataset that you want to merge into this dataset. The specific variable name matches for the CPS and the ACS can be found in Table B.1. Not all of the variables of interest are identical between the CPS and the ACS, so we did a careful mapping to ensure that we could construct cells that correctly captured the relevant information. The colors represent the degree to which the variables are similar to each other, with green meaning that they are exactly the same, yellow meaning slight differences, and orange meaning major differences but potentially still possible to match on.

Deciding which identities to cluster together requires (1) a measure of the distance between identities and (2) an algorithm by which to cluster identities that are nearest to each other into cells. We choose to maintain state integrity and therefore cluster identities together within states, using only the demographic characteristics to determine distance. We use the Gower distance to calculate distances and the PAM clustering algorithm for clustering (see Appendix B for more details).[30] The end result of this is a final set of cells in the CPS, each of which has at least five records.

How We Match Cells Across Datasets to Ensure Consistency

After creating cells in the CPS, we assign the same labels to records in the ACS by matching records according to the key demographic variables and the geographic state. For each identity in the ACS, we check whether there is an exact match represented in the CPS. If there is, then the records of that identity in the ACS receive the same cell label as the records of that identity in the CPS. If there is no exact match, then the identity is fuzzy matched in the following way:

1. We calculate the Gower distances between the candidate identity from the ACS and all identities in the CPS.
2. We assign the cell label of the identity in the CPS that is the minimum distance from the candidate identity in the ACS.
3. In case there are multiple identities in the CPS that are equidistant to the candidate identity in the ACS, we break ties by comparing the median income in each of the equidistant groupings in the CPS against the median income of the candidate identity in the ACS.[31] The cell label of the identity in the CPS with the smallest absolute difference in income from the candidate identity in the ACS is assigned to the candidate identity in the ACS.

After assigning cell labels to all identities in the ACS in this way, we can check the performance of the match by comparing the quality of the cells in each dataset. One commonly used method to measure the quality of cells is the silhouette score. The silhouette score evaluates the quality of clustering by measuring how similar an object is to its own cluster compared with other clusters. This score ranges from −1 to 1, with higher values representing a better fit. For more details, see Appendix B.

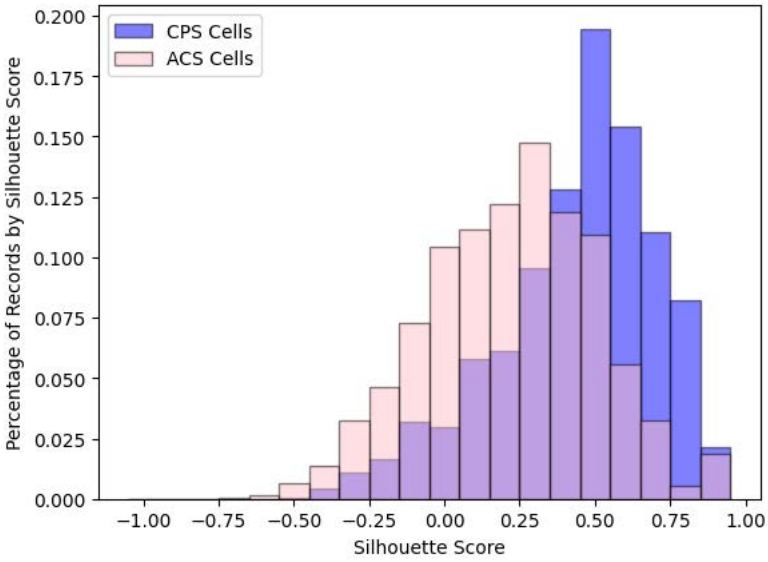
We calculate the silhouette scores for all cells in both the CPS and the ACS. Naturally, we would expect that the silhouette scores for cells in the CPS should be weakly better than for those in the ACS, as the clustering was done directly on the CPS and then walked over to the ACS. Figure 3.1 shows histograms of the silhouette scores for the cells in the CPS (blue) and the ACS (pink) in Alabama when the CPS data are clustered via PAM and the ACS is fuzzy matched to get cell labels. We can see that, as expected, the silhouette scores are on average higher for the

[30] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, Vol. 27, No. 4, 1971.

[31] Income was chosen as the tiebreaker, because the income distribution is a primary interest of this study.

cells in the CPS, indicating that the records in these cells match their cells better than ACS records match theirs.

Figure 3.1. Cell Quality in the CPS and the ACS



NOTE: This example is of cell construction in Alabama. The blue bars represent the silhouette scores of cells in the CPS dataset after clustering via PAM. The pink bars represent the silhouette scores of cells in the ACS dataset after they have been matched to the CPS dataset to obtain cell labels.

How We Merge Records While Preserving Income Distribution

Once we have constructed cells in the CPS and applied them to the ACS for a given state, we begin the record-merging process between the appropriate cells in the CPS and the ACS. Because we are particularly interested in preserving the income distribution, we sort the records in each cell by income. Records are merged by taking into account the record weight—which is the estimated number of people within the U.S. population that a record likely represents.

The pseudocode for our record merging within a cell approach is below. The general approach is to group records that are identical across most characteristics and then, within each group, match individual records in the ACS with records in the CPS such that the multiple ACS records may be linked to a single CPS record or vice versa, depending on the record weights. Here, C_{CPS} is the cell of CPS records, and i is an individual record in that cell. w_i is the weight of record i. C_{ACS} is the cell of ACS records corresponding to C_{CPS}, and j is an individual record within C_{ACS}. w_j is the weight of record j. We construct the synthetic record k by linking records from the ACS and the CPS:

For i in C_{CPS}:
  Set w' equal to weight of record i (w_i)

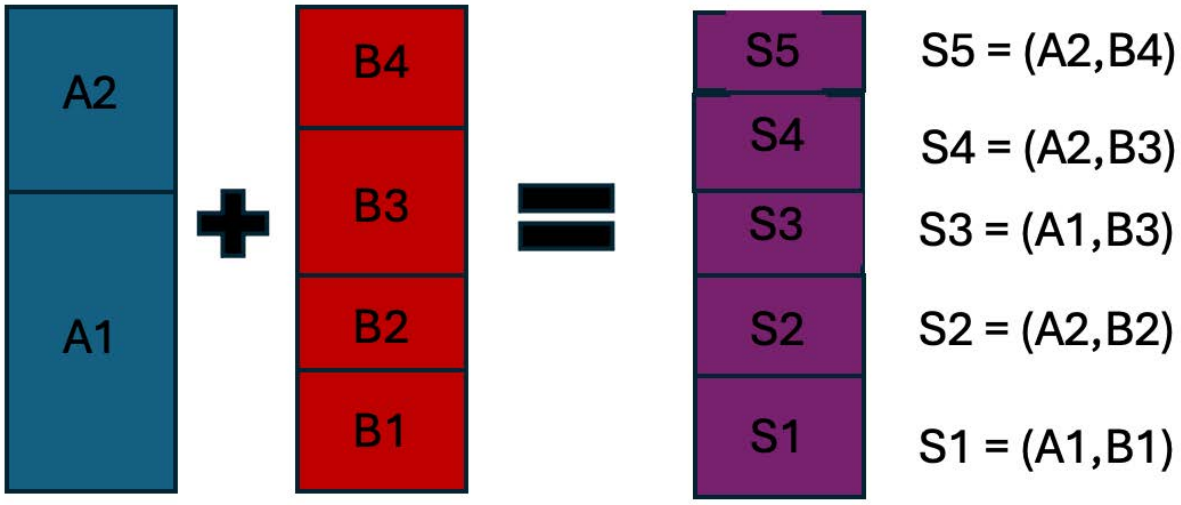
While $w' > 0$:
For j in C_{ACS}:
If $w' \geq$ weight of record j (w_j):
Create synthetic record k as the concatenation of the weight, and the characteristics in the CPS and the ACS (w_j, record i, record j)
Reduce w' by w_j: $w' = w' - w_j$
If $w' <$ weight of record j (w_j):
Create synthetic record k using (w', i, j)
Split record j into two parts: j and j'
Update weights:
$w_j' = w_j - w'$
$w_j = w'$
Set w' to 0: $w' = 0$.

Within each cell, we step through each record in the CPS and match it to records in the ACS until the sum of the ACS weights matches or surpasses that of the CPS record. If there is surplus weight in the last added record from dataset B, we split that record and reweight.

Figure 3.2 shows how two cells in datasets A and B would be joined in a synthetic dataset, S, using this approach. In this example, the cell in dataset A has two records, and the corresponding cell in B has four records. The values on the right represent the characteristics of the records in the synthetic dataset. So, for example, S1 uses relevant characteristics from A1 and B1.[32] After applying this approach to the CPS and the ACS, the resulting merged synthetic dataset now contains characteristics from both the CPS and the ACS.

[32] Note that, in this example, the cells in datasets A and B are of similar total weight, so the records in the synthetic cell maintain the overall total weight of both datasets. If this were not the case, the weight of one of the datasets would not be maintained.

Figure 3.2. Illustrative Example of the Merging Approach to Create the Synthetic Dataset



Key Challenges in Matching Survey Records Across Datasets

Since the CPS and the ACS survey different individuals, we must match records based on available characteristics (i.e., matching with incomplete information). This creates the potential for mismatches. Mismatches occur most often in underrepresented groups. For example, it is well known that the income distribution of the top earners is not properly captured in survey data.[33]

Ideally, cell weights would match across datasets. However, in practice there are many reasons why this might not happen. In the case of the CPS and the ACS, the distributions of demographic characteristics are not the same across the datasets. For example, the share of record weights in the CPS corresponding to people who are disabled is roughly 10 percent, while the fraction of share weights in the ACS corresponding to people who are disabled is roughly 16 percent. Mismatches like these mean that the total weights across cells will not be perfectly preserved after merging.

Since our approach merges cells to match records, the data are no longer reliable for distinguishing conditional distributions between groups whose cells were merged.[34] So, if a pair of cells is merged to ensure a large enough sample size and if those cells differ on one variable, then the resulting data should not be used to describe differences along that variable between the

[33] Carter C. Price and Kathryn A. Edwards, "Trends in Income from 1975 to 2018," RAND Corporation, WR-A516-1, 2020.

[34] For example, if "less than high school" and "high school" are merged to create a cell for the synthetic dataset, then any distribution that is estimated from the synthetic dataset conditional on "less than high school" or "high school" will not necessarily accurately match the conditional distributions for the original datasets.

populations (at least for members of that cell). Hypothetically, if the cell for low-income, rural women were merged with middle-income, rural women for a given state, the synthetic dataset would not be useful for describing differences between rural women by income in that state. This does limit the suitability, but, because merges happen when there are too few records or there are mismatches between the datasets, any statistical analysis of the data along those lines would likely not have been reliable under any circumstance. Although there is an inherent uncertainty in this matching, we can use goodness-of-fit metrics, such as total record weights, total weighted income, and income-weighted distributions, to assess the extent of mismatches.[35] These measures could also describe the suitability of the synthetic data for different types of analyses.

[35] We consider three goodness-of-fit measures for our cell matches. The first is total weights: For each cell, we want the sum of the CPS record weights to equal the sum of the ACS record weights. This measure ensures that we are preserving the population size represented by each weight. The second is total weighted income: Because we want to ensure that we are accurately representing income, within each cell, we compare the record-weighted sum of personal income between the matched datasets. This ensures that the income totals for each cell (and the full synthetic dataset) are close. Essentially, this serves the same function as a conservation law in physics—in the matching, income is neither created nor destroyed. The third is income-weighted distribution: For the purpose of modeling the effects of policy and assessing the effects of policy, it is important to accurately represent the income distribution. To measure the degree to which we are preserving the income distribution, we look at the record-weighted median, the 25th percentile, and the 75th percentile. By preserving these moments, we can accurately describe the income distribution in the synthetic population.

Chapter 4. How We Integrated Spending Data with Our Other RAND Budget Model Datasets

We map federal spending to individuals and households in the Synthetic Population Dataset, as well as to firms, industries, and geographies. The Macro Model, Personal Income Tax Model, and Corporate Tax Model can use these data to explore demographic, economic, and geographical variability. The integration process depends on the type of spending and researchers' analytic objectives, as outlined below. Table 4.1 describes how the three categories of spending variables can be mapped to survey data.

Table 4.1. Mapping Spending to RAND Budget Model Datasets and Models

Category	Mapping Destination	Examples	USAspending.gov Variables	Mapping Destination Variables
A. Spending directly made to or that directly benefits individuals or households	Synthetic Population Dataset Personal Income Tax Model	All assistance awards: • SNAP, WIC • Medicaid • Social Security Other direct spending to individuals or households: • Disaster relief	• Spending type • Awarding top-tier agency • Awarding subtier agency • Assistance listing code • Place of performance	• Demographics, income or assets (match program eligibility criteria) • Geography
B. Spending directly made to or that directly benefits a firm, industry, or geography	Synthetic Firm Dataset Macro Model Corporate Tax Model	• Contracts, loans, or subsidies to firms • Grants to research institutions	• Recipient type • Recipient NAICS • Place of performance	• NAICS • Geography
C. Spending that does not clearly benefit any populations or segments of the economy but is relevant for macroeconomic analysis	Macro Model Corporate Tax Model	• Community development • Transportation grants • Defense spending • Interest on public debt • Foreign aid • Law enforcement		• Total government spending • Local or national GDP growth • Local or federal budget balance

NOTE: SNAP = Supplemental Nutrition Assistance Program.

Category A: Individual-Related Federal Spending

Category A captures spending made directly to beneficiaries, which can be mapped to individuals and households in the **Synthetic Population Dataset** and the **Personal Income Tax Model**.

Category A includes all financial assistance made to beneficiaries through safety net programs, such as Medicare, Medicaid, SNAP, WIC, Social Security, and unemployment insurance. Examples are awards made by the Department of Agriculture through programs under the Department of Agriculture's Food and Nutrition Service, such as SNAP, WIC, and the National School Lunch Program. Category A also includes other direct spending made to individuals or households. These include other assistance awards, which provide financial aid to advance federal objectives, such as disaster relief.

How to Find and Track Direct Beneficiary Payments in Federal Data

Table 4.2 shows key variables for tracking spending in the USAspending.gov database.

Table 4.2. Spending Type Variables for Category A

USAspending.gov Variable	Examples of Values	Purpose
Spending type	• Assistance award • Contract award	Identify financial assistance awards
Awarding top-tier agency, awarding subtier agency	• Department of Health and Human Services, Social Security Administration • Department of Agriculture, Food and Nutrition Service	Classify spending by agency responsible for specific safety net programs
Assistance listing code	• SNAP (awarding agency = Department of Agriculture, Food and Nutrition Service) • National School Lunch Program (awarding agency = Department of Agriculture, Food and Nutrition Service)	Classify spending by safety net program within an agency
Budget function	• Medicare (570) • Income security (600) • Social security (650) • Veterans benefits and services (700) • Disaster relief and insurance (453)	Classify spending by federal objective
Budget bureau name	• Centers for Medicare and Medicaid Services (awarding top-tier agency = Department of Health and Human Services)	Classify spending by budget bureaus within an agency
Business categories	• Individuals	Identify spending made to individuals

USAspending.gov Variable	Examples of Values	Purpose
Award description	• Funding for disaster relief for households affected by earthquakes, tornadoes, or other natural disasters	Provide context about purpose and scope
AssistanceListingNumber, AssistanceListingTitle	• Various	Map to an assistance listing in the Catalog of Federal Domestic Assistance
AssistanceType	• 02 = block grant, 03 = formula grant, 04 = project grant, 05 = cooperative agreement, 06 = direct payment, etc.	Identify the type of assistance to understand how funds are allocated
DisasterAwardObligations, DisasterEmergencyFundCode, DisasterEmergencyFundName, DisasterFaceValueLoans, DisasterNumberAwards, Disaster Recipient, LocalAreaSetAside, NationalInterestAction	• Various	Identify spending for disasters, the disaster type, if it was a loan, number of awards, and if recipient was an individual or city/state
InherentlyGovernmentalFunctions	• Closely associated, critical function, other function	Identify whether spending is for a critical government function
MajorProgram	• Major program	Identify whether spending is for a designated major program within the agency
Product_or_Service_Code	• Various	Identify the product or service procured
ProgramActivitiesFundingThisAward, ProgramActivityCode, ProgramActivityName	• Various	Tie spending to a program

Table 4.3 details recipient characteristics and locations. Table 4.4 shows spending amounts, timing, and types.

Table 4.3. Recipient Variables

USAspending.gov Variable	Examples of Values	Purpose
Recipient type	• Individuals	Identify spending made directly to individuals
Recipient place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport	Identify geographies where recipient is located
Primary place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport • Primary place of performance scope = multi-state, state-wide, county-wide, city-wide, single ZIP, foreign	Identify geographies where the majority of the recipient's work is performed
HousingAuthoritiesPublic/Tribal	• F = false, T = true	Identify whether the spending goes to a housing authority
IndianTribeFederallyRecognized	• F = false, T = true	Identify whether the spending goes to a federally recognized Indian Tribe
LaborSurplusAreaFirm	• F = false, T = true	Identify whether the recipient firm is in a labor surplus area (high unemployment area)

Table 4.4. Quantity Variables

USAspending.gov Variable	Examples of Values	Purpose
Fiscal year	2023, 2024, 2025	Track year of award for longitudinal analysis
FaceValueofDirectLoanOrLoanGuarantee	Various	Identify whether it is a loan and how much the principal of the loan is for
ObligatedAmountFromCOVID19Supplementals, ObligatedAmountFromIIJAsupplemental	Various	Identify whether any funding came from these two specific emergency funding pools
OriginalLoanSubsidyCost	Various	Estimate the long-term cost to the government of a direct loan or loan guarantee
TotalLoanValue	Various	Identify a loan and how much the total loan (interest and principal) is for

How Synthetic Population Data Map Government Benefits to Individual Recipients

We research the eligibility criteria and participation rates for each of the government programs relevant to the RAND Budget Model. These consist of income and asset limits, household composition, demographics (e.g., age limits), and other relevant metrics. Using these

eligibility criteria, as well as information in USAspending.gov variables, we map Category A spending to representative individuals and households in the Synthetic Population Dataset. We analyze patterns in Category A spending by population group, geography, and program to inform scenario design in the Personal Income Tax Model. The specific approaches are described in Table 4.5.

Table 4.5. Mapping Approaches

Source Dataset and Model	Source Dataset Variable and Model Parameter	Examples of Values	Purpose
Synthetic Population Dataset, Personal Income Tax Model	Income, assets	• Household income = 123% FPL • Net worth = $345	Determine whether the synthetic identification meets income or asset limits for government programs, and model personal income tax implications of changes in program enrollment and benefit levels
Synthetic Population Dataset, Personal Income Tax Model	Demographics, household composition	• Sex = female • Age = 65 • Has school-aged children = yes	Determine whether synthetic ID is part of the target population for government programs, and model distributional impacts of changes to Category A spending on personal income tax
Synthetic Population Dataset, Personal Income Tax Model	Geography	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport	Determine whether synthetic ID benefits from spending to specific geographies, and model geographic variation in impacts of changes to Category A spending on personal income tax

How This Analysis Informs Policy Decisions About Federal Programs

The RAND Budget Model measures the distribution of government spending across the population. This matters because there is substantial geographic variation in spending through U.S. safety net programs. By mapping spending to synthetic individuals and households, researchers can answer such questions as: How does spending in a specific government program vary between states and over time? What is the impact of legislation to increase or reduce federal funds to a specific government program on state budgets? In which geographies and among which populations might enrollment in government programs increase, and what are the federal budgetary implications?

Additionally, not everyone who qualifies takes up safety net programs, and the rate of take-up varies by demographic group. That means that not every eligible person is actually a beneficiary. With more accuracy, we can link participation data from the Survey of Income and Program Participation (SIPP) to the Synthetic Population Dataset. The SIPP is a nationally representative dataset that includes measures of eligibility and participation in all major government programs at the individual- or household-month level. Thus, this extension of the

tool can be used by analysts of benefits programs to better understand the distribution of program participants and the population eligible for a program who do not participate.

Category B: Business-Related Federal Spending

Certain spending items link directly to firms, industries, and geographies. Category B includes all spending made directly to firms or other private for-profit entities. Examples include contracts with IT firms to improve government software, loans to small businesses from the Small Business Administration to expand operations, grants to renewable energy firms to develop green technology, and subsidies to agricultural firms for crop stabilization or price stabilization. We map Category B spending to representative firms in the **Synthetic Firm Dataset** based on firm characteristics and to industries and geographies in the **Corporate Tax Model**.

How We Track Business-Related Federal Spending Using USAspending.gov Data

We use NAICS codes, geography, and other variables described in Table 4.6 in the USAspending.gov data to identify Category B spending and characterize the recipient firms.

We use the variables in Table 4.7 to characterize the firms that receive Category B spending, such as the industry, region, and type of business.

Table 4.6. Spending Type Variables

USAspending.gov Variable	Examples of Values	Purpose
Award type	• Fixed-price contract • Assistance	Distinguish between contracts (procurement) and assistance (grants, loans)
AssistanceType	• 02 = block grant • 03 = formula grant • 04 = project grant • 05 = cooperative agreement • 06 = direct payment, etc.	Identify the type of assistance to understand how funds are allocated
BusinessFundsIndicator	• NON = not Recovery Act • REC = Recovery Act	Identify whether the award was applicable to the American Recovery and Reinvestment Act of 2009
DOD_Acquisition_Program	• Various	Identify whether the spending matches a DOW spending category (will be more relevant for future work)
EmergencyAcquisition	• B = humanitarian • C = presidential declaration	Identify whether the spending is related to an emergency declaration or aid (e.g., COVID-19)
EPA-DesignatedProduct, RecoveredMaterials/Sustainability	• A = meets requirements	Identify whether the spending purchases goods that meet Environmental Protection Agency requirements (will be more relevant for future work)
Research	• Small Business Research Program, etc.	Identify whether the spending is for research and does not directly go to any program
PlaceOfManufacture	• Within the United States, partly within the United States, not within the United States	Identify whether manufacturing is located in the United States

Table 4.7. Recipient Variables for Category B

USAspending.gov Variable	Examples of Values	Purpose
Recipient type	• Small business • For-profit organization • Private institution of higher education	Identify type of entity
Recipient NAICS code	• 31–33 (Manufacturing) • 61 (Educational Services) • 92 (Public Administration)	Classify spending recipients by industry
Recipient place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport	Identify geographies where spending is targeted
Primary place of performance	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport • Primary place of performance scope = multi-state, state-wide, county-wide, city-wide, single ZIP, foreign	Identify geographies where the majority of the recipient's work is performed
Business categories	• category_business, corporate_entity_tax_exempt, other_than_small_business • category_business, small_business, woman_owned_business	Identify whether spending is made to for-profit firms, nonprofits, local governments, other entities, or a special category, such as small business, woman-owned
Business types	• Contracts and grants, for-profit organization • Department of Transportation– certified Disadvantaged Business Enterprise, for-profit organization • Construction firm, contracts, for-profit organization	Identify spending made to for-profit firms rather than nonprofits, local governments, and other entities
Diverse and disadvantaged business types	• AmericanIndianOwnedBusiness, AsianPacificAmericanOwnedBusiness, EducationalInstitution, EmergingSmallBusiness, HispanicAmericanOwnedBusiness, HistoricallyUnderutilizedBusinessZone HUBZoneFirm, MinorityOwnedBusiness, ServiceDisabledVeteranOwnedSmall Business, Self-CertifiedSmall DisadvantagedBusiness, VeteranOwnedBusiness, WomenOwnedBusiness, etc.	Identify spending made to businesses owned by diverse and typically disadvantaged groups
Type of set-aside	• None • Small business • Section 8(a) Business Development program participant	Identify whether the spending was available only to certain businesses (typically small or disadvantaged)

USAspending.gov Variable	Examples of Values	Purpose
CityLocalGovernment, CountyLocalGovernment, Inter-MunicipalLocalGovernment, MunicipalityLocalGovernment, SchoolDistrictLocalGovernment	• F = false • T = true	Identify whether the spending goes to a city or local government, state, or private enterprise
CommunityDevelopmentCorporation, CommunityDevelopedCorporationOwned Firm	• F = false • T = true	Identify whether the recipient firm is part of a community development corporation (local-level impact)
InternationalOrganization	• False • True	Identify whether the recipient is international (spending is not used within the United States)
TransitAuthority	• False • True	Identify whether spending goes to an organization related to transportation

Researchers and cost analysts must distinguish between federal contract awards (*awards*) or indefinite delivery vehicles (IDVs). An award has an individual total obligated amount, and the total spending will appear in the USAspending.gov data once the contract is awarded. An IDV is a type of contract that allows the government to buy goods and services over a longer period (e.g., five years) from one or more contractors without specifying the exact quantity or delivery dates up front. IDVs have a maximum amount that can be spent, called the *obligated amount*. It is not known up front when, how much, to whom, and for what the obligated amount will be spent. In the USAspending.gov data, IDVs have a total obligated value and outlay amounts—the amount that has been spent to date. To understand the total amount of spending and potential impacts of an IDV, researchers must account for the total obligated value rather than the outlay amounts. We do this in different ways depending on analytical needs; however, one way could be to assume that the government will spend the full obligated amount and spread the total obligated value equally across the IDV's period of performance.

The USAspending.gov data also include variables identifying the source and amount of associated spending from nonfederal sources, as described in Table 4.8. For example, the terms of a federal grant provided to a Tribal Nation may require a percentage of funds to be spent by the Tribal Nation related to the grant. In this case, we consider this additional funding when estimating the full impact of a federal grant program.

Table 4.8. Quantity Variables for Category B

USAspending.gov Variable	Examples of Values	Purpose
Fiscal year	2023, 2024, 2025	Track year of award for longitudinal analysis
Award or IDV flag	AWARD, IDV	Distinguish between awards and IDVs. IDVs should also include the variables BaseAndAllOptionsValue and BaseAndExercisedOptionsValue and IDV_Type and MultipleOrSingleAwardIDV
BorrowingAuthorityAmountTotal_CPE, BudgetAuthorityAppropriatedAmount_CPE	Various	Identify whether spending was allowed to be financed by borrowing or if it was part of an appropriation (legally obligated)[a]
CorporateEntityNotTaxExempt, CorporateEntityTaxExempt	F = false, T = true	Identify whether spending is tax-exempt
ContractAwardType	A = blanket purchase agreement (BPA) call, B = purchase order, C = delivery order, D = definitive contract	Identify the contract type to understand how funds are distributed
TotalSubsidyCost	Various	Identify whether spending is a subsidy
Total obligation	715876.20	Identify obligations or spending commitments
Total outlays	30933.08	Identify outlays or actual spending made

[a] This variable may be relevant to the Macro Model and other RAND Budget Model modules in identifying where funds may be coming from to finance the spending and program funding. In particular, understanding whether spending is financed by borrowing can add to an analysis of federal debt.

How We Match Federal Spending to Business Recipients Across Models

We map government spending to representative firms in the Synthetic Firm Dataset using firm characteristics (e.g., NAICS code, public or private sector, geography) as described in Table 4.9. We also aggregate spending by fiscal year, geography, and NAICS code and serve as inputs for analyzing production, GDP growth, and corporate taxes in the Corporate Tax Model and Macro Model.

Table 4.9. Mapping Approaches

Source Dataset and Model	Source Dataset Variable and Model Parameter	Examples of Values	Purpose
Synthetic Firm Dataset, Corporate Tax Model	• NAICS • Public or private sector	• Manufacturing (31–33) • Computer and Electronic Product Manufacturing (334) • Public, private	Determine whether the synthetic firm is in same industry and sector as spending recipients, and measure differential exposure to spending by industry and sector
Synthetic Firm Dataset, Corporate Tax Model, Macro Model	• Geography	• State = OR • Congressional district = 05 • ZIP5 = 97365 • City = Newport	Measure differential exposure to spending by geography

How Federal Spending Analysis Informs Policy Impact Assessments

The RAND Budget Model analyzes how economic growth varies across geography, industry, and labor force segments. For example, when evaluating manufacturing tax legislation, researchers can analyze

- county-level manufacturing employment data from the Bureau of Labor Statistics and the U.S. Census Bureau
- projected GDP impacts compared with historical trends
- tax revenue effects on manufacturing firms
- labor market outcomes for manufacturing workers and manufacturing-heavy regions.

Category C: Nontargeted Federal Spending: Infrastructure, Defense, and Debt Service

Measuring total federal spending and its components enables analysis of GDP growth and budget balance. Category C captures spending without direct beneficiaries, such as transportation grants, community development, interest payments, and defense spending. These include transportation grants, community development, interest on public debt, and defense spending. Category C includes these types of spending.

Three Types of Federal Spending Shape Economic Impact Analyses

The Macro Model distinguishes between mandatory, discretionary, and supplemental spending. Mandatory spending includes outlays made to individuals, businesses, nonprofits, and state and local governments through federal benefit programs and represents roughly 60 percent

of total federal spending.[36] The largest mandatory programs are Social Security, Medicare, and Medicaid. Researchers use awarding agencies, assistance listing codes, budget functions, budget bureau names, and TASs to identify mandatory spending for major health care programs (e.g., Medicare, Medicaid), Social Security (e.g., Old-Age, Survivors, and Disability Insurance), income security programs (e.g., SNAP, Supplemental Security Income, Earned Income Tax Credit), certain veteran programs, and other programs.[37]

Discretionary spending makes up about 25 percent of total federal spending, although that is expected to decline over time as costs for Social Security and Medicare grow. Roughly half of discretionary spending goes to national defense, and the rest funds the administration of other agencies and programs.

Congress approves supplemental spending for urgent needs that cannot wait for annual appropriations. Examples include COVID-19 spending, additional disaster relief funds, emergency humanitarian assistance, and spending for other urgent purposes, as described in Table 4.10.

[36] Kathleen FitzGerald, Ann E. Futrell, Susanne Mehlman, and Emily Stern, *Common Budgetary Terms Explained*, Congressional Budget Office, December 2021.

[37] Dan Ready, Jorge Salazar, and Caitlin Verboon, "Mandatory Spending in Fiscal Year 2024: An Infographic," Congressional Budget Office, March 20, 2025.

Table 4.10. Examples of Spending Variables for Category C

USAspending.gov Variable	Examples of Values	Spending Category
Assistance listing code	• SNAP (awarding agency = Department of Agriculture, Food and Nutrition Service) • National School Lunch Program (awarding agency = Department of Agriculture, Food and Nutrition Service)	Classify spending by safety net program within an agency
Budget function	• Medicare (570) • Income security (600) • Social Security (650) • Veterans benefits and services (700) • Disaster relief and insurance (453)	Classify spending by federal objective
Budget bureau name	• Centers for Medicare and Medicaid Services (awarding top-tier agency = Department of Health and Human Services)	Classify spending by budget bureaus within an agency
TAS	• Agency identifier: 86 (Department of Housing and Urban Development) • Main account code: 0300 (Community Development Block Grants)	Unique identifier for federal accounts managed by the U.S. Treasury that links spending to specific Treasury accounts
Supplemental		
BusinessFundsIndicator	• NON = not Recovery Act • REC = Recovery Act	Identify whether the award was applicable to the Recovery Act (COVID-10)
DisasterEmergencyFundCode, DisasterEmergencyFundName,	• CARES Act (Emergency) • DEFC = "O"	Identify supplemental spending for disasters
EmergencyAcquisition	• B = humanitarian • C = presidential declaration	Identify whether the spending is related to an emergency declaration or aid (COVID-19)
ObligatedAmountFromCOVID19 Supplementals, ObligatedAmountFromIIJA supplemental	• Various	Identify whether any funding came from these two specific emergency funding pools

NOTE: DEFC = Disaster Emergency Fund Code.

How Spending Commitments Differ from Actual Expenditures

Researchers may need to distinguish between obligations and outlays. Obligations legally bind spending commitments and help with analyzing longer-term spending needs over specific periods. Outlays represent actual government expenditures. USAspending.gov has complete data on all agencies' outlays since FY 2022. Examples can be seen in Table 4.11.

Table 4.11. Examples of Obligations and Outlays for Category C

USAspending.gov Variable	Examples of Values	Purpose
Total obligation	715876.20	Identify obligations or spending commitments
Total outlays	30933.08	Identify outlays or actual spending made

How Budget Analysis Informs Debt Ceiling and Economic Policies

The RAND Budget Model decomposes GDP growth across geographies, budget functions, and other dimensions. This would allow researchers to better understand the channels through which new legislation could affect macroeconomic outcomes and to construct more-accurate scenarios. For example, Congress could pursue several strategies to lower the debt ceiling by reducing spending for certain types of programs or budget functions. In each case, researchers can model breaks in historical trends in such federal spending. Researchers can also model potential impacts on economic outcomes for individuals and households by mapping this spending to the Synthetic Population Dataset by program eligibility criteria and geography.

Chapter 5. How We Estimate Knock-On Effects of Spending

In this chapter, we provide guidance to economic analysts on estimating how spending to either firms or households affects the broader state or regional economies. This approach distinguishes between subsidies to firms and transfers to households, as these have different economic consequences. The methodological discussion in this chapter is technical, but these classes of models estimate the knock-on effects of federal spending based on how funds flow from the recipient into the community and broader economy.

There are two main approaches for considering knock-on effects: I/O models and computable general equilibrium (CGE) models. These are similar approaches that differ in how firms and households may adjust as the economy changes. I/O models assume that each firm has a recipe for production such that, to double output, all inputs must be doubled. That is, there is no substitution between inputs, but those inputs are perfect complements to production. Similarly, households consume a fixed bundle of goods based on income. If income doubles, consumption of all goods doubles. Alternatively, CGE models are more flexible with respect to how firms and consumers respond to changes in prices. For example, if labor is relatively cheaper than capital, then firms will substitute labor away from capital. The goal of both sets of models is to understand how policy changes or shocks affect not just the targeted sector but all sectors in the economy.

There are advantages and disadvantages to each approach. I/O models are relatively straightforward from an implementation standpoint and for lay economists to understand how the models function. They effectively produce economic multipliers whereby a shock of $1 to an industry with a multiplier value of 1.5 translates into a $1.5 total effect on the economy that can be measured in terms of GDP, total output, employment, or state, local, and federal taxes. These models can also be used to estimate the effects of both subsidies to firms and transfers to households. This provides a sense of how large a spending impact a shock has on the economy as a whole. I/O models provide good estimates for spending levels that are small relative to the size of the economy under consideration. I/O models do poorly when policy changes have large impacts or when those impacts are in the form of price changes, such as tax rates.

Alternatively, CGE models are more complex to implement and require more data to calibrate the parameters necessary for implementation. Importantly, an I/O model is a special case of a CGE with a particular choice of model parameters that do not allow for substitution. But CGE models are able to handle changes in prices and, more importantly, allow firms and households to change their consumption composition based on these price changes, as well as consider policy changes in the form of taxes.

The type of policy change will inform which type of model will be used. In general, where I/O models can be used, we will use them because they are more transparent and require less

economic intuition to interpret the results, and these models can be easily scaled to larger or smaller policy changes. Additionally, for national-level impacts, the Bureau of Economic Analysis provides all the data necessary for the calibration of a national I/O model for three separate levels of detail with respect to industry disaggregation. For alternative geographic disaggregation, alternative sources of data are necessary to calibrate such models. The bureau's Regional Input-Output Modeling System (RIMS II) estimates are provided at a regional level at a cost but offer only the economic multipliers, discussed below, rather than the full data necessary to calibrate an I/O model.[38] Alternatively, IMPLAN provides the regional data necessary to calibrate an I/O model and the software for implementation.[39] But these I/O data can also be used to calibrate CGE models, if the policy analysis needs to consider price changes or substitution. Therefore, in most cases we use the IMPLAN data to estimate the impacts either directly through an I/O model or indirectly through a calibrated CGE.

How I/O Models Calculate Economic Impact Multipliers

Economists developed I/O models in the early 20th century to better understand the flow of goods and services through a double accounting framework.[40] This allows us to know two important pieces: (1) what inputs are necessary to produce goods and services and (2) how those goods and services are allocated across different industries and household demands. This approach takes advantage of linear algebra to have simple approaches to analyze the effects of changes in the economy. The starting point for analysis is an equation that defines how goods are produced and allocated within the economy:

$$x = Ax + d.$$

In this equation, x is a vector of outputs by industry; A is a transformation matrix that provides the recipes for how goods are produced, where columns represent the production requirements for each industry; and d is household demand. What this means is that if there is an exogenous shock to demand of size Δd, this translates into a change in aggregate output of

$$\Delta d = (I - A)^{-1}\Delta x.$$

The $(I - A)^{-1}$ matrix is known as the *Leontief inverse matrix* and provides the sector-by-sector economic multipliers that can be aggregated to form sector-level economic multipliers.

---

[38] Bureau of Economic Analysis, "Welcome to the RIMS II Online Order and Delivery System," webpage, undated.

[39] IMPLAN, homepage, undated (for the 2025 model year).

[40] The original work by Leontief in 1941 and 1951 provides the basic structure for all I/O models. See Wassily W. Leontief, *The Structure of American Economy, 1919–1929: An Empirical Application of Equilibrium Analysis*, Harvard University Press, 1941; Wassily W. Leontief, *The Structure of American Economy, 1919–1939: An Empirical Application of Equilibrium Analysis*, 2nd ed., Harvard University Press, 1951.

This simplified presentation has several key implications. First, it assumes a closed economy but can be easily expanded to consider imports and exports, where those are more easily viewed as different kinds of households within the model. This is important as we consider different geographies for the analysis of federal spending. For example, a set of 51 state or district models will produce different results from a single nationwide model. Within I/O models, we talk about *regional purchasing coefficients*, which are the proportion of the change in demand that is satisfied within the economy rather than coming from outside sources.

Beyond external demand shocks, I/O models allow for the consideration of changes in income. These may arise because of transfers from the federal government but also as an effect of shocks to demand. In general, we divide the effects of policy changes into three buckets: (1) direct, (2) indirect, and (3) induced. Direct effects are those that come directly from the policy or spending change. The I/O models incorporate these effects; for this project, they would come from aggregations of the USAspending.gov data. The indirect effects are those that arise through the changes in industry demands and are akin to the economic multipliers arising from the Leontief inverse matrix. The induced effects are those that arise from changes in income that increase demand from positive shocks and are commonly referred to as *income effects*.

For work considering federal spending, the economic geography may be important. In most cases, we will consider a national-level model to better understand the aggregate knock-on effects of spending. But our goal is to shed more light on who is affected by spending. As discussed in Chapter 3, we may want to disaggregate by industry, state, both, or some other geography, such as Federal Emergency Management Agency regions or Federal Reserve districts.

In the current version of this work, there is a mismatch between the IMPLAN data and the synthetic datasets described in Chapter 3. The IMPLAN data have nine household types distinguished by income. Although it would be better to use the synthetic data to assign payments to the household types, the individual state-level considerations would need to be used to assign payments to households and then aggregate by household income. This is work that will be added in future versions.

Chapter 6. Examples of How to Use These Federal Spending Data to Estimate Economic Impact

This chapter contains a series of examples of applying the methods described in Chapter 5 for estimating spillover effects to the spending data described in Chapter 2 and the synthetic data described in Chapter 3. Although the examples vary by the USAspending.gov fields, key attributes, and other characteristics, we apply common integration strategies for certain spending types. For example, for assistance awards (e.g., SNAP, Medicaid), integration begins with USAspending.gov data on the assistance listing code, recipient geographic codes, and recipient NAICS codes, as well as eligibility criteria for the government program. In this chapter, we describe two examples for integrating assistance awards with the synthetic dataset and mapping the spending to segments of the U.S. population.

Integration requires matching common elements across both datasets. As described in Chapter 2, the spending data contain details on the geographies (for some types of records, the details go as far as the address) and NAICS codes of recipients and beneficiaries, as well as the fiscal year and month of the spending. The synthetic data described in Chapter 3 have detailed geographic information (typically at the county level) and NAICS codes of individuals, employers, and communities, as well as calendar year and month. Additionally, the synthetic dataset has relevant data elements for determining eligibility for many of the government programs described in the USAspending.gov data. By matching common elements across the datasets, we can link spending to the appropriate populations.

Geographic matching creates a key challenge. Specifically, the location information in the USAspending.gov data corresponds to the physical site of employment, whereas the location information in the ACS and the CPS is about the residence of the respondents. Because people may work for firms in a different county or even a different state from their residence, we use commuting zones to match workers to the spending.[41] This matching approach is well suited for the types of jobs that require some level of physical presence at the worksite but less well suited for remote workers. Further, because this type of work is not uniform across geography, industry, or demographic characteristics, the effect of this limitation is not uniform. As remote work is generally considered "knowledge work," future iterations of this project may be able to better match those workers to spending. Similarly, with the NAICS codes, not every individual working in a specific sector in a geographic area may receive the spending marked in the data. Thus, the appropriate interpretation of a match should be that it represents the types of workers at

[41] C. W. Carpenter, M. C. Lotspeich-Yadao, and C. M. Tolbert, "When to Use Commuting Zones? An Empirical Description of Spatial Autocorrelation in U.S. Counties Versus Commuting Zones," *PLOS One*, Vol. 17, No. 7, July 2022.

firms receiving the funds rather than a specific worker receiving the spending. A more sophisticated construction of synthetic firms could be used to refine this analysis in the future.

Lastly, for beneficiaries of assistance awards, we can examine eligibility criteria by income, family type, location, and so on to identify beneficiary populations. In this initial version of the tools, we cannot identify people who actually participate in nonuniversal programs (i.e., programs that require eligible individuals to apply for benefits or otherwise restrict access). Information on program participation may be added in the future using data from the SIPP and other nationally representative surveys.

Example 1: Food and Nutrition Programs Reveal Complex Distribution of Federal Aid to Beneficiaries

Figure 6.1 presents the outlays in FY 2023 for several distinct programs under the U.S. Department of Agriculture's Food and Nutrition Service. We filter the USAspending.gov database for observations in which the awarding top-tier agency name is the Department of Agriculture and the awarding subtier agency name is Food and Nutrition Service, and we use assistance listing codes to distinguish between food and nutrition programs. We then use recipients' geographic codes (from the place of performance) to map spending under each program to states. We further inspect the eligibility criteria for each program to inform the key attributes (e.g., income level, sex, having kids) to use when mapping spending to cells in the synthetic dataset. The analysis matches the largest 50 records to the programs and presents them by state. Mapping these programs to individuals in the synthetic dataset is challenging and involves applying the eligibility criteria from the USAspending.gov data to the population to find the most-relevant estimates of the spending distribution by cell.

The distinct Department of Agriculture Food and Nutrition Service programs presented in Figure 6.1 are as follows:

- **SNAP** provides food assistance to low-income families and individuals. SNAP is administered by the states, so the eligibility criteria vary by state.[42] Additionally, because people are required to apply, not everyone who is eligible for SNAP receives benefits.
- The **National School Lunch Program** provides free or reduced-price lunches to students in kindergarten through 12th grade who are in low-income households or who are in schools with a substantial number of students in low-income households.[43] As structured, the synthetic population can be used to flag children in low-income households who meet

[42] Food and Nutrition Service, U.S. Department of Agriculture, "Supplemental Nutrition Assistance Program (SNAP)," webpage, last updated March 12, 2026a.

[43] Food and Nutrition Service, U.S. Department of Agriculture, "National School Lunch Program," webpage, last updated December 11, 2025b.

the eligibility criteria but cannot be used to identify those students who meet the criteria because of the share of students from low-income families.[44]

- The **Nutrition Assistance for Puerto Rico Program** is challenging to map to beneficiaries because not all relevant datasets include Puerto Rico, so the synthetic population is deficient for this purpose.[45]
- **State Administrative Matching Grants for SNAP** go to states to support the administration of SNAP.[46] In this case, the beneficiary is the state government and the taxpayers of the state.
- **WIC** provides food assistance and other supports to some low-income women and their young children.[47] The eligibility criteria are clear and, with the exception of pregnancy, are all observable in the synthetic dataset.

By triangulating USAspending.gov fields on assistance listing codes and geography, as well as program eligibility criteria, we can map the distribution of spending from specific food and nutrition assistance programs to beneficiaries across states and fiscal years. This information enhances analyses of trends in specific assistance programs, contextualizes spending in assistance programs against overall state and federal budgets, and supports evaluations of the economic impacts of government programs.

This is one example of mapping for specific government programs. Because government programs vary in their eligibility criteria, benefit structure, and administration, we may use different mapping approaches for different assistance awards. For example, identifying recipients and beneficiaries for contractual spending will look different from recipients and beneficiaries for grants and other fixed spending.

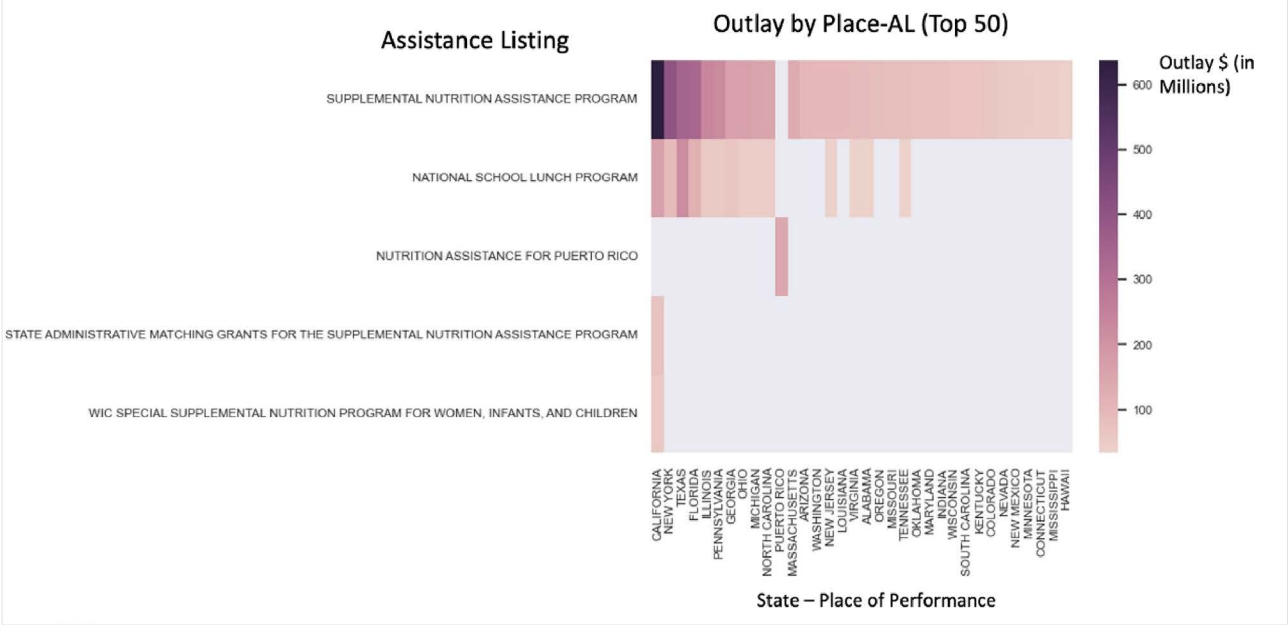
[44] In practice, we can produce estimates for people getting free lunches through this other path by looking at poverty in their areas. This will not be perfect but will likely be sufficiently accurate for most analytic purposes.

[45] Food and Nutrition Service, U.S. Department of Agriculture, "Examination of Cash Nutrition Assistance Program Benefits in Puerto Rico," webpage, last updated January 24, 2025a.

[46] Food and Nutrition Service, U.S. Department of Agriculture, "SNAP Program Administration," webpage, last updated May 29, 2026b.

[47] Economic Research Service, U.S. Department of Agriculture, "WIC Program," webpage, last updated July 22, 2025.

Figure 6.1. Food Assistance Spending by Program and State, FY 2023

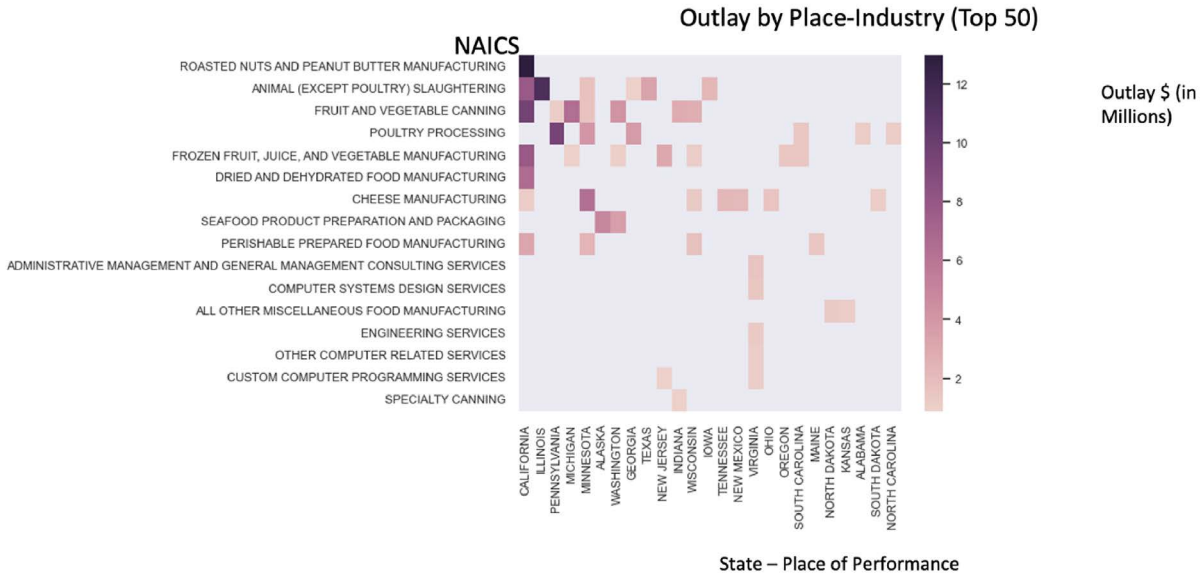


NOTE: AL = assistance listing.

Example 2: School Lunch Program Payments Flow Primarily to Food-Processing Companies, Not Direct Recipients

Figure 6.2 shows the spending by NAICS code and the state of the place of performance for the top 50 spending recipients of the National School Lunch Program in FY 2023. The NAICS codes are all of those present in the records and include a variety of food-processing and agriculture-related codes. The states where the provider is located for the 50 selected records are also displayed. These details about the recipients of outlays are enough for us to identify people based on their states and sectors in the synthetic dataset who are likely to work for the firms that receive payments for these programs. It is important to note that the recipients for this program are typically agricultural companies (e.g., a meat-packing plant) that deliver food assistance rather than the individuals who receive food assistance. Thus, for this program, the recipients are different from the beneficiaries.

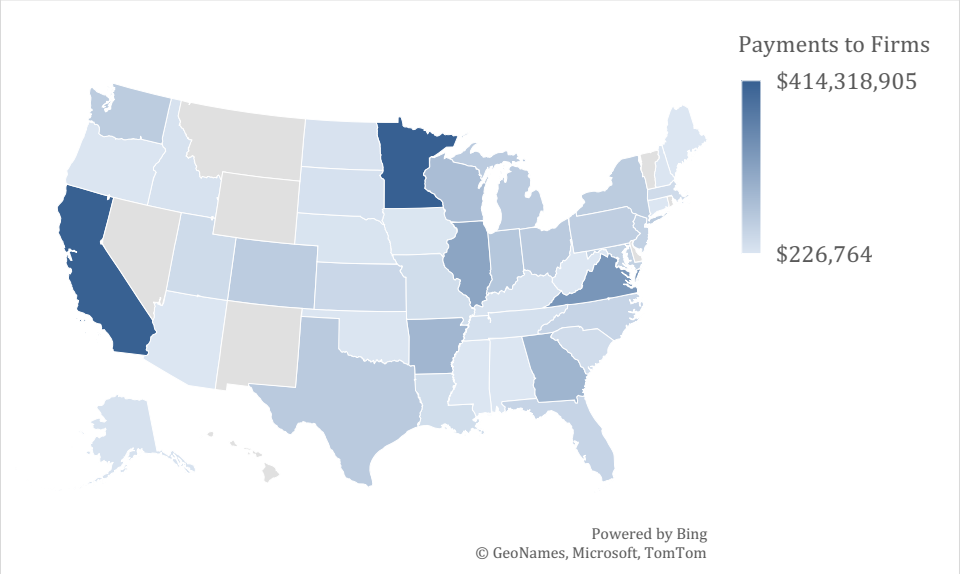
Figure 6.2. National School Lunch Program Spending by Industry and State, FY 2023



Example 3: SNAP Benefits Generate Broader Economic Impacts Beyond Direct Recipients

We analyzed SNAP to provide insights into how shocks to demand for industry output extend beyond the individual firm to the broader state and national economies, as well as how the transfers to households affect the economies. We calibrate 52 models using data from IMPLAN and consider these two broad sets of shocks. We modify the data in Figure 6.1 to consider the aggregate impact of industry spending by state. These are the contracts that SNAP has with individual firms to purchase products that are then aggregated at the state level. Figure 6.3 provides the geographic distribution of the funds. There are two outliers for this: California and Minnesota. These payments are largely dominated by food manufacturing. For other high-recipient states, many of the payments are to IT suppliers and support.

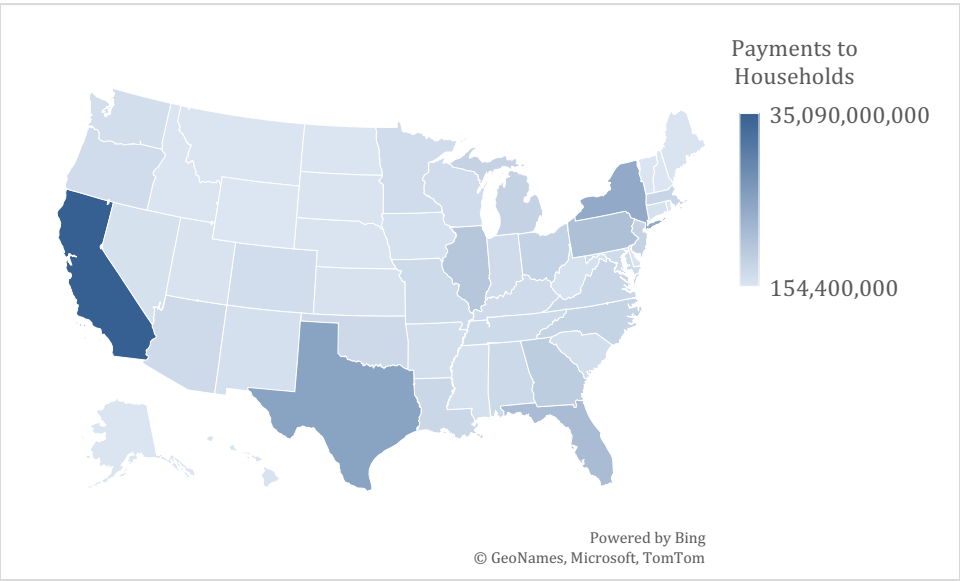
Figure 6.3. Payments to Industry from SNAP



SOURCE: Aggregation of data from USAspending.gov, undated-d, for SNAP payments to industry.

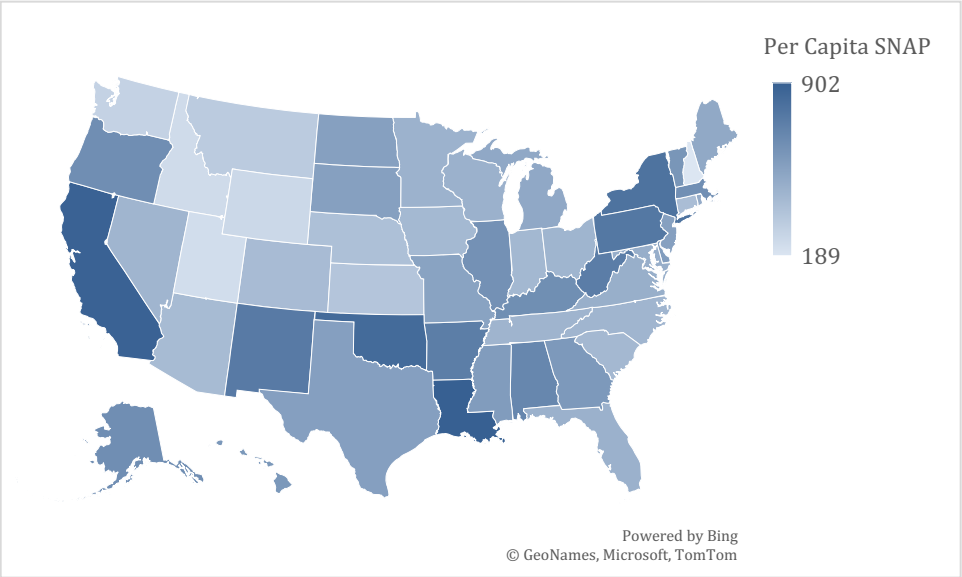
Figure 6.4 provides the distribution of aggregate payments to households through SNAP. As was already discussed, states administer these programs differently and use different criteria for benefit recipients, as well as those who can participate in the program. The high outliers are partially driven by large populations in California, New York, Texas, and Florida. Thus, Figure 6.5 provides these on a per capita basis.

Figure 6.4. Payments to Households by SNAP



SOURCE: Aggregation of data from USAspending.gov, undated-d, for SNAP payments to households.

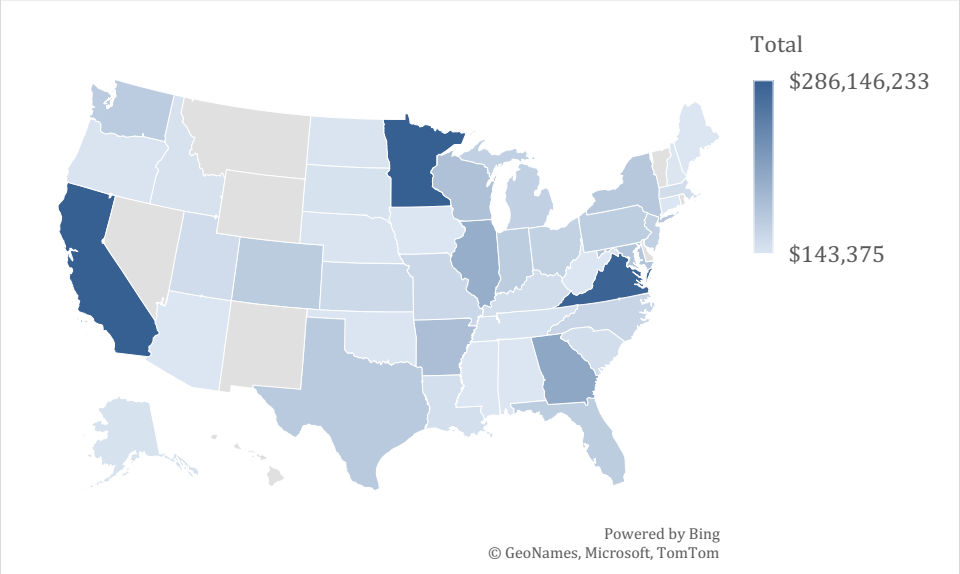
Figure 6.5. Per Capita Payments to Households from SNAP



SOURCE: Features data from USAspending.gov, undated-d.

We use these payments to estimate the aggregate impact on gross state product (GSP), not just directly from these payments but also from the knock-on effects. To implement these changes in an I/O framework, we include the payments to industries separately from the payments to households. To implement the payments to industries, we first aggregate the USAspending.gov data by industry and state and then compare the state economy with and without the aggregate spending amount. A variety of metrics could be used to display these results, including output, employment, GSP, or taxes. We report the increase in GSP driven by SNAP payments as our metric. We do this for two reasons. First, reporting the results in changes to output has a tendency to overestimate the impact by double counting output, since these are all expressed in monetary terms. For example, if we increase the output of cars, we also increase the output of tires. But the value of the car includes the value of the tires. Second, reporting just changes in taxes misses many of the impacts that may be of interest, especially when we are considering federal spending, as is the case in this example. Although we chose GSP, employment would be an equally transparent choice if our interest lies more in the workforce. Figure 6.6 provides the results for the payments to industry. For these results, it is important to note that the impact on GSP is smaller than the payments to industry, as some of the benefits spill outside of the state. For example, in Minnesota, the largest contracts are for food manufacturing, which has livestock as an input, but all the livestock may not be from Minnesota.

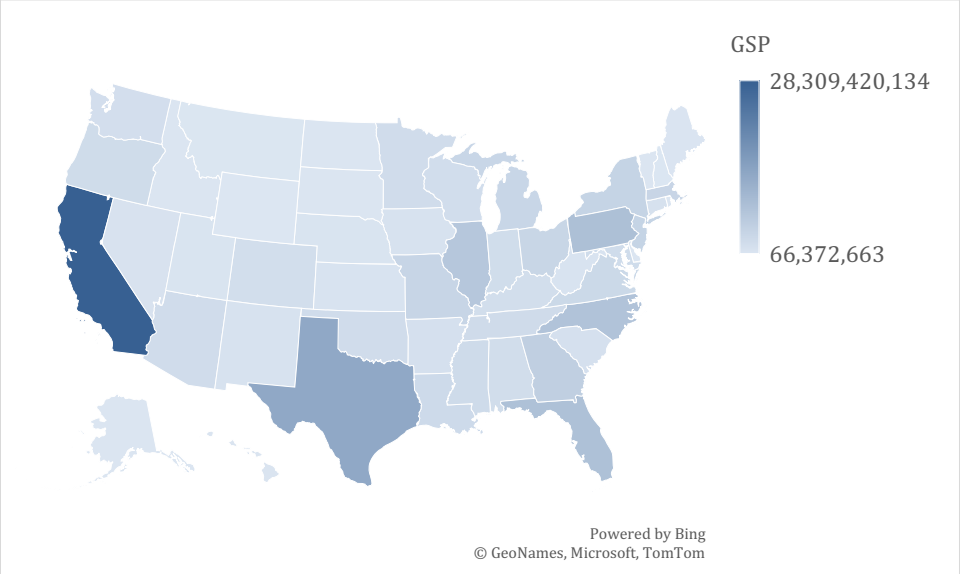
Figure 6.6. Total GSP of SNAP Payments to Industry



SOURCES: Features data from IMPLAN and USAspending.gov, undated-d.

In addition to the payments to industry, we consider payments to households. We assume that the payments from SNAP are a transfer from the federal government to households. For the current version of the tools, we assign the aggregate state-level payments to households and divide them into the two lowest income classes based on the aggregate income proportions for these two income types. While not ideal, this approach provides a reasonable estimate of the impacts of these payments. As noted above, future versions of the tools will assign payments to households and aggregate by household income within states for more-precise state comparisons. Additionally, these payments are considered income rather than money forced to be spent on food. We do this because SNAP payments are substitutes for income—other sources of income are moved away from food if SNAP is not present. Figure 6.7 provides the distribution of gross state product (GSP) because of the household payments from SNAP. As with the payments to industry, there is leakage from the payments to households out of states to other states and potentially internationally, as food and food products are from across the United States and the globe.

Figure 6.7. GSP Impacts of SNAP Payments to Households



SOURCES: Features data from IMPAN and USAspending.gov, undated-d.

We estimated a national-level model and aggregated the results from the state-level model. In total, households received $200 billion in payments across all states. According to the state-level models, this has an aggregate impact on GSP of approximately $143 billion. When we consider a national-level model, the impact on GDP is $263 billion. That is, $120 billion in economic impact flows from payments in one state to increases in GSP in another state. Therefore, it is important to recognize that these benefits may spill out of one state and be captured by another. Moving forward, better integration of these state-level models needs to be considered because of these leakages. This can be accomplished by moving toward multiregion input-output models or a multiregion CGE to obtain better estimates at the subnational level.

Additional Applications Show Tool's Broader Value

The mapping between disaggregated spending data and synthetic population applies to a variety of other use cases. The tools will be suitable for a wide variety of applications, including the following:

- **Demographic characteristics of workers affected by federal investments:** By linking employer information (such as the NAICS code) from the CPS to geographic information in the ACS, we can describe the demographic characteristics of the workforce for firms that receive contract awards.
- **Geographic distribution of spending on beneficiaries of federal programs:** The details of spending and eligibility criteria for a specific program can be linked to the beneficiaries' counties, states, or other geographic information to produce spending maps in total or per capita.

- **Comparison of the role of different programs on populations or locations:** The spending profile by geography or demographics of different programs could be compared to better understand how programs are fragmented, overlapping, or duplicative.

By pairing these geographic spending distributions with an I/O model, economic multipliers, or other macroeconomic model, the overall economic effect of this spending can be estimated for individual communities.

Chapter 7. These Tools Enable Better Federal Spending Analysis

In this document, we have explained how we constructed tools to map federal spending to the beneficiaries and recipients. In their current state, these tools display the geographic distribution of spending with no additional processes. Similarly, we include the spending by sector using the NAICS codes and map that spending to people in the synthetic population dataset to describe the funds going to workers in different sectors of the economy.

These federal spending tools in their current state apply to a wide variety of policy questions related to the targets, topics, and effects of federal spending, such as the following:

- What are the demographic characteristics of the recipients of funds for a given program?
- What population benefits from the spending associated with a given program?
- How does spending in one community compare with the national average?
- What are the spillover effects associated with a given program, and how does it compare with other programs?

Because of the data cleaning, the processed USAspending.gov data can be used more easily than the raw data for understanding the contents of federal spending. Likewise, although not as accurate as if there were a single dataset that captured all the data elements of both the CPS and the ACS datasets, the synthetic population data are useful for geographically disaggregating data elements from other datasets.

To further increase the utility of this work for the RAND Budget Model project and other users, we are planning updates to these approaches, described below.

Three Key Improvements Will Enhance Tool Capabilities

We have three primary directions for next steps for these tools:

- **Improve the processing of USAspending.gov data:** The data from USAspending.gov, in their present form, map well to the recipients by NAICS code or place but do not always contain enough information to map to the beneficiary of the spending. As part of this project, we have done some initial processing to improve those connections, but we need to do more work to make those connections reliable. In future work, we intend to use information about the programs linked to the spending available in the text fields in USAspending.gov, as well as to leverage other data elements, such as budget function codes and TASs. We will also look for other sources of the eligibility criteria for programs so that we can correctly identify populations that may be beneficiaries of different programs.
- **Include additional data sources in the synthetic population:** At this point, we have connected the ACS to the CPS. This combination is useful for many types of questions, but to better identify beneficiaries in the future, we intend to include data from the SIPP regarding program participation and the SCF for information about household debts and

assets. Because these are survey data that are similar to the CPS and the ACS, we will use a similar approach to that described in Chapter 3 to combine these datasets. Initial data on variables to be used for matching and merging can be found in Table B.1. We will also use data from the Internal Revenue Service's Statistics on Income (SOI) to better match income distributions of taxable data.[48] Because the SOI has aggregated descriptive statistics, we will use distributional adjustments similar to those used in a previous RAND study to make small modifications to the synthetic data to better match the SOI distributions.[49]

- **Enhance the mapping between spending and targets:** The linkages between the USAspending.gov data and the synthetic population are constrained by the content of both. Therefore, by doing the modifications described above, we can enhance our ability to accurately map the spending lines to the beneficiaries and recipients. Additionally, as described in Chapter 4, we do not currently identify beneficiaries for many types of federal programs, and that will be a focus for future updates to this analysis.

[48] Internal Revenue Service, "SOI Tax Stats—Statistics of Income," webpage, last updated March 17, 2026.

[49] Price and Edwards, 2020.

Appendix A. Overview of Relevant USAspending.gov Data Elements

This appendix lists key data elements available in the USAspending.gov data:

- **Action date and fiscal year:** the date the government issued or signed the award or reached a binding agreement and the fiscal year. The data contain information from fiscal years going back to FY 2008, although earlier datasets capture less data than the current year. We have not assessed the degree of completeness for each fiscal year.
- **Assistance listing number:** the number assigned to an assistance listing in the Catalog of Federal Domestic Assistance and SAM.gov.[50] Assistance listings describe federal programs that provide financial assistance awards.
- **Award type:** categorizes each record's spending into contracts, contract IDVs, grants, direct payments, loans, insurance, or other.
- **Agency:** identifies the awarding and funding agencies.
- **TAS:** identifies the agency and subcomponents linked to the spending record.
- **Location:** specifies geographic details for both recipient and performance location. In both cases, the location includes the country, state, county, city, ZIP code, and congressional district.
- **Recipient:** the name and unique ID of the entity receiving the funds.
- **Recipient type:** a categorical variable that includes broad categories, such as general business, minority-owned business, women-owned business, veteran-owned business, special designations, nonprofit, higher education, government, and individuals. Each of these categories has a variety of subcategories.
- **Award amount:** the dollar value of the award.
- **Award ID:** the unique ID for the award.
- **Assistance listing:** the number in the Catalog of Federal Domestic Awards. The catalog lists grants available to U.S. people and organizations. This will only have values for relevant grant programs.
- **Budget function code:** organizes federal spending by primary government activity, such as national defense, health, or transportation. These codes are aligned with the federal budget classification system, which organizes spending based on the primary purpose or objective of the funds.
- **Disaster Emergency Fund Code:** a code connecting an award to a specific disaster. Only disaster-related supplemental funding records include this code.
- **NAICS:** the NAICS code and name for the award recipient.
- **Product or Service Code (PSC):** the "codes to describe products, services, and research and development (R&D) purchased by the federal government."[51]

[50] SAM.gov, "Assistance Listings," webpage, U.S. General Services Administration, undated-a.

[51] U.S. General Services Administration, *Federal Procurement Data System Product and Service Codes (PSC) Manual: Fiscal Year 2024 Edition*, April 2024, p. 4.

- **Type of contract pricing:** a description of the pricing mechanism for the contract when applicable. This can include pricing approaches, such as cost plus, firm fixed price, or time and materials.
- **Type of set-aside:** a description of what, if any, special set-aside was relevant for the record.
- **Extent competed:** a description of whether the funding was competed and what mechanism was used for the competition, if any.

Appendix B. Methods Used to Merge Demographic Data with Federal Spending Records

This appendix describes the technical details for merging the demographic data.

Variables in the CPS and the ACS for Cell Construction and Merging

Table B.1 lists the key variables in the CPS and the closest variable in the ACS and the SCF. Colors are used to indicate similarity: Green indicates that the variables are essentially identical, variables colored yellow are similar but not identical, and orange indicates that the variables are conceptually similar but with meaningful differences. Also, because the SCF is primarily a household-level dataset, the individual-level income variables in the CPS and the ACS need to be aggregated to the household level to be matched to the SCF.

Table B.1. Comparison of Variables That Could Be Used to Match the CPS to the ACS or the SCF

CPS Variables	ACS Variables	SCF Variables	Notes
1. YEAR (survey year)	1. YEAR (census year)	1. Year (survey year)	
6. ASECWTH (Annual Social and Economic Supplement Household weight)	5. HHWT (household weight)	2. Revised Kennickell-Woodburn consistent weight (household weight)	
8. REGION (region and division)	7. REGION (census region and division)		
9. STATECENSUS (state [census code])	8. STATEFIP (state [FIPS code])		
10. COUNTY (county [FIPS code])	9. COUNTYFIP (county [FIPS code, identifiable counties only])		The CPS to the ACS: The codes are the same, but both datasets say that the county is not always listed in order to protect identity of respondents
11. METFIPS (metropolitan area [FIPS code])	11. CITY (city [identifiable cities only])		The CPS to the ACS: METFIPS provides metropolitan areas, including cities and also suburbs and in some cases exurbs. Some of the city names match the FIPS metro areas, but the codes are different numbers. 846 missing cities

CPS Variables	ACS Variables	SCF Variables	Notes
			for CPS out of 1,163 for the ACS.
12. METRO (metropolitan and central/principal city status)	10. METRO (metropolitan status [where determinable])		The CPS to the ACS: very similar, slightly different codes
15. FAMINC (family income of householder)	41. FTOTINC (total family income)	3. INCOME (total amount of income of household)	The CPS to the ACS: Income is not represented in the same way; one reports actual value, and the other puts income into categorical buckets
18. PERNUM (person number in sample unit)	14. PERNUM (person number in sample unit)		
21. ASECWT (Annual Social and Economic Supplement weight)	15. PERWT (person weight)		
22. RELATE (relationship to household head)	16. RELATE (relationship to household head [general version])		
23. AGE (age)	19. AGE (age)	4. AGE (age of reference person)	
24. SEX (sex)	18. SEX (sex)	5. HHSEX (gender of household reference person)	
25. RACE (race)	21. RACE (race [general version])	6. RACECL5 (alternate class of race of respondent)	
26. MARST (marital status)	20. MARST (marital status)	7. Married (marital status of reference person)	
29. VETSTAT (veteran status)	58. VETSTAT (veteran status [general version])		
31. CITIZEN (citizenship status)	25. CITIZEN (citizenship status)		
32. HISPAN (Hispanic origin)	23. HISPAN (Hispanic origin [general version])	6. RACEL5 (alternate class of race of respondent)	
33. EMPSTAT (employment status)	33. EMPSTAT (employment status [general version])	7. X6670 (job status in detailed dataset)	
35. OCC (occupation)	38. OCC (occupation)	8. OCCAT2 (occupation classifications for reference person)	
38. IND (industry)	39. IND (industry)	9. INDCAT (industry classifications for reference person)	
39. CLASSWKR (class of worker)	37. CLASSWKRD (class of worker [detailed version])	10. OCCAT1 (occupation categories for reference person)	
43. EDUC (educational attainment recode)	32. EDUCD (educational attainment [detailed version])	11. EDCL (education categories of ref person)	

CPS Variables	ACS Variables	SCF Variables	Notes
46. FULLPART (worked full or part time last year)	61. UHRSWORK (usual hours worked per week)		The CPS to the ACS: One provides categorical data, and the other provides integers but is able to convert hourly to categorical
47. INCTOT (total personal income)	40. INCTOT (total personal income)	3. INCOME (total amount of income of household)	The ACS and the CPS contain individual-level data for this income, but the SCF has it at the household level
48. INCWAGE (wage and salary income)	42. INCWAGE (wage and salary income)	12. WAGEINC (household wage and salary income)	The ACS and the CPS contain individual-level data for this income, but the SCF has it at the household level
49. INCBUS (nonfarm business income)	43. INCBUS00 (business and farm income, 2000)	13. BUSSEFARMINC (income from business, sole proprietorship, and farm)	
50. INCFARM (farm income)	43. INCBUS00 (business and farm income, 2000)	13. BUSSEFARMINC (income from business, sole proprietorship, and farm)	
51. INCSS (Social Security income)	43. INCSS (Social Security income)	14. SSRETINC (Social Security and pension income)	
52. INCWELFR (welfare [public assistance] income)	44. INCWELFR (welfare [public assistance] income)	15. TRANSFOTHINC (unemployment, alimony/child support, food stamps, Supplemental Security Income, and other)	
53. INCRETIR (retirement income)	46. INCRETIR (retirement income)	16. SSRETINC (Social Security and pension income)	
55. INCINT (income from interest)	45. INCINVST (interest, dividend, and rental income)	17. INTDIVINC (interest and dividend income)	
59. INCDIVID (income from dividends)	45. INCINVST (interest, dividend, and rental income)	17. INTDIVINC (interest and dividend income)	
60. INCRENT (income from rent)	45. INCINVST (interest, dividend, and rental income)	17. INTDIVINC (interest and dividend income)	
64. POVERTY (original poverty status [PUMS original])	49. POVERTY (poverty status)		

SOURCES: Features information from IPUMS CPS, undated; IPUMS USA, homepage, University of Minnesota, undated; Ruggles et al., 2023; Board of Governors of the Federal Reserve System, 2023.
NOTE: PUMS = Public Use Microdata Sample. Green indicates that the definition of terms matches between the CPS and the ACS. Yellow indicates characteristics with similar, but not identical, definitions between the datasets. Orange indicates that the terms are conceptually similar but do not match between the datasets. Gray indicates that there are no data in the SCF that correspond to the CPS data factor.

How We Calculate Distances Between Demographic Groups Using Gower's Method

Demographic characteristics can be one of two types: categorical or ordinal. When comparing two cells on a categorical dimension, such as race, the cells can be either the same or different. Thus, we calculate the distance between cells *j* and *k* for categorical characteristics using the Hamming distance, given as

$$dist_{categorical}^{j,k} = 1\big(categorical_j = categorical_k\big).$$

When comparing two groupings on an ordinal dimension, such as age, the groupings vary in their similarity. For example, a group of 34-year-olds is more similar to a group of 30-year-olds than to a group of 12-year-olds. In this case, the appropriate distance measure is an absolute difference. To allow comparison with the distances for categorical variables, the difference is normalized by the range of the characteristic:

$$dist_{ordinal}^{j,k} = \frac{\big(ordinal_j - ordinal_k\big)}{range(ordinal)}.$$

We calculate the total distance between groupings *j* and *k* using

$$dist^{j,k} = dist_{categorical}^{j,k} + dist_{ordinal}^{j,k}.$$

This total distance measure is a *Gower distance*.

Using Silhouette Scores to Validate Cluster Quality

We use silhouette scores to evaluate the quality of clustering by measuring how similar an object is to its own cluster compared with other clusters. The silhouette score considers both the cohesion within clusters, defined as the average distance $a(i)$ between a sample individual or household i and all other individuals and households within the same cluster, and the separation between clusters, defined as the average distance $b(i)$ between a sample individual or household and all the individuals and households in the nearest cluster. This measure provides an overall assessment of clustering performance by essentially describing whether an individual or household best fits in the assigned cluster or whether the individual or household would better fit in a neighboring cluster. Silhouette scores are given by

$$s(i) \;=\; \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)},$$

and $s(i)$ ranges from −1 to 1, where a higher score indicates a better fit.

Why We Selected PAM as Our Clustering Algorithm

We begin PAM clustering with a specified *m* number of points called *medoids*, which we randomly choose. The medoids are in an *n*-dimensional space, where *n* is the number of characteristics along which the cells are being constructed. The Gower distance from each record to each medoid is calculated, then each record is assigned to its nearest medoid, resulting in *m* cells. We then calculate dissimilarity scores by taking the Gower distance between each record in each cell and every other record in that cell and summing these distances. The record within each cell that has the lowest dissimilarity score is then assigned to be the new medoid for that cell. The algorithm iteratively refines the selection of medoids and assigns records to the nearest medoid to minimize the sum of dissimilarities until no further improvements can be made. By running PAM over a range of values for *m* and selecting the output from the maximal *m* that still respects the threshold that the minimum cell size is ≥5, we get a final set of cells that has as many cells as possible and a minimum cell size of ≥5.

PAM produces low variance in silhouette widths for the final cells. We tested many other clustering algorithms (such as hierarchical agglomerative clustering) and found PAM to produce the highest-quality final cells,[52] although it also resulted in fewer total cells and is computationally inefficient, making it impractical for datasets that are much larger than the CPS. Additional results are available upon request.

[52] See Anastasia Reusova, "Hierarchical Clustering on Categorical Data in R," *Towards Data Science*, April 1, 2018; Fionn Murtagh and Pedro Contreras, "Methods of Hierarchical Clustering," arXiv, arXiv:1105.0121, April 30, 2011.

Appendix C. Interactive Heat Map

This appendix describes the construction of an interactive heat map using the USAspending.gov data. This interactive heat map lets users visualize spending by fiscal year and the geography of the recipient (state, county, city, ZIP code).

We aggregated the data by awarding top-tier agency, awarding subtier agency, and assistance listing code. The ability to categorize spending by federal agency and program allows for easy integration into scenario design in the RAND Budget Model.

Top-tier awarding agencies represent the federal department awarding the spending. A list of top-tier awarding agencies appears at USAspending.gov.[53] The subtier awarding agency is the division under the top-tier agency (e.g., the Internal Revenue Service is a subtier agency of the U.S. Department of the Treasury, and the Federal Bureau of Investigation is under the U.S. Department of Justice). The assistance listing is the federal program under the subtier awarding agency that provides the spending. Researchers can look up assistance listing codes and the associated top-tier and subtier awarding agencies at SAM.gov.[54]

Our heat map shows the relative amount of spending made to various geographies for the selected attributes. In the next iteration, researchers will be able to download data on every spending item that matches the filters as a CSV file. The heat map currently aggregates all spending types. In future iterations, users will be able to filter by outlays versus obligations, contracts versus loans versus grants, and more. We welcome suggestions from potential users.

The U.S. Department of Agriculture Price Support Program Shows Regional Funding Disparities

The interactive heat map reveals spending patterns across federal programs. By identifying the awarding top-tier agency as "DEPARTMENT OF AGRICULTURE (USDA)" and the awarding subtier agency as "Farm Service Agency," researchers can filter down to spending by the specific federal agency, as seen in Figure C.1. The Farm Service Agency provides financial assistance to agricultural producers to recover from disasters in the form of emergency farm loans, management programs, emergency farm loans, rural development loans and grants, and more. Researchers can then narrow the spending to a specific program under the agency. For example, they can set the assistance listing to "10.051—Commodity Loans and Loan Deficiency Programs," the agency's price support program to assist agricultural producers. Spending under

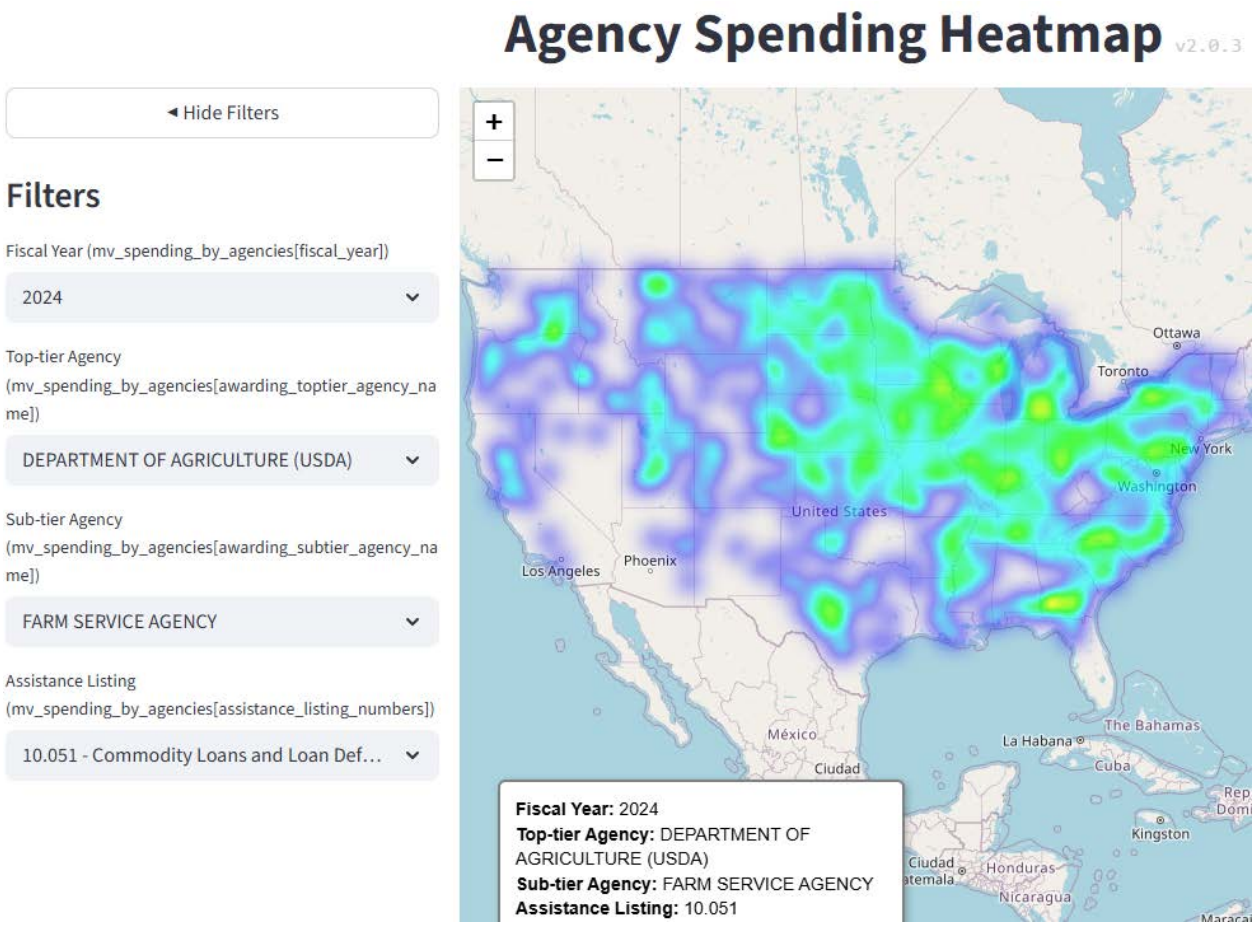
[53] USAspending.gov, "Agency Profiles," webpage, undated-a.

[54] SAM.gov, undated-a.

this program includes direct payments and loans to farmers to help stabilize incomes during periods of low prices.

The map shows how spending from this program flows throughout the United States. Spending concentrates in the Midwest and South, indicating that these regions received relatively higher dollar amounts. Researchers can then download a CSV file, which will list every spending item made under this program in FY 2024, with the dollar amount and geographic codes. These data can then be used for scenario design in the RAND Budget Model.

Figure C.1. Interactive Heat Map of the Farm Service Agency's Price Support Program

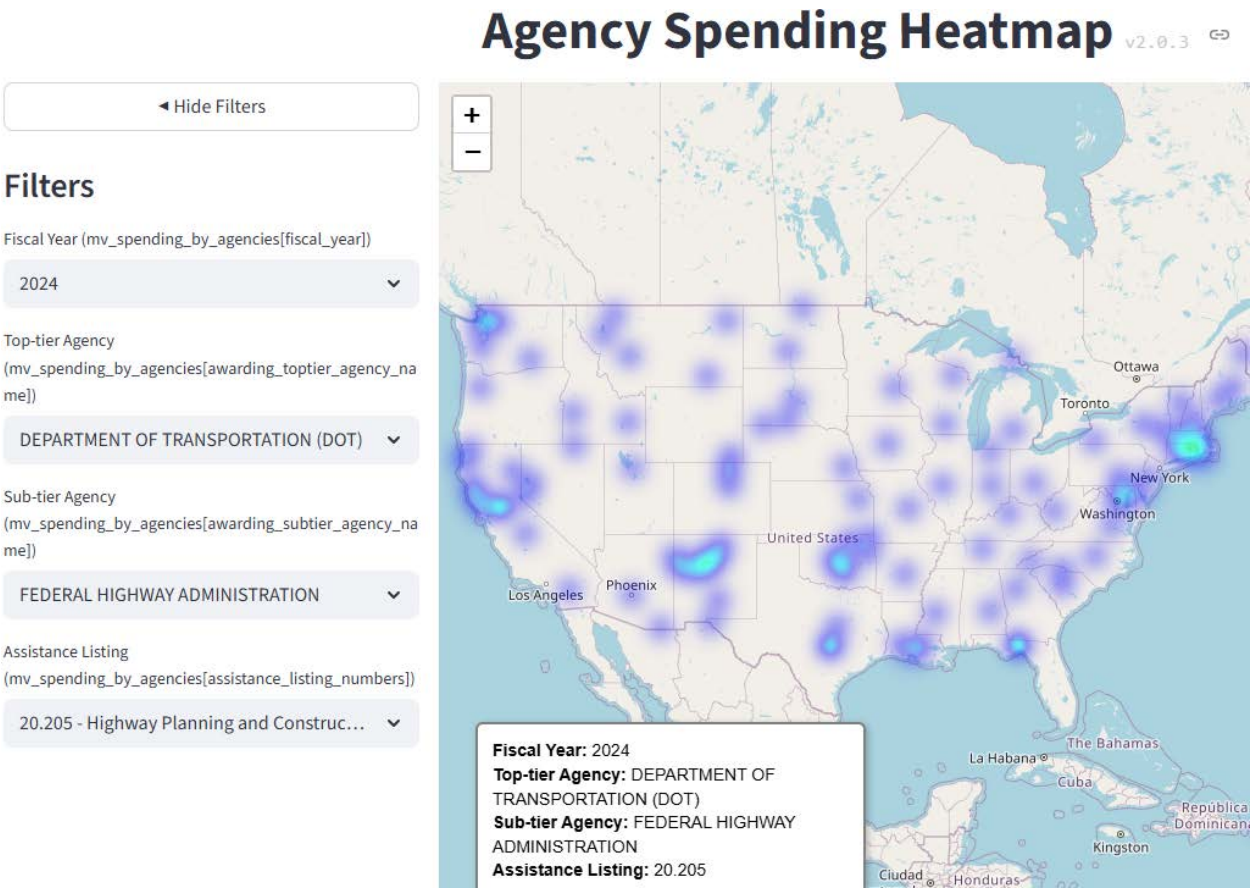


Highway Construction Funds Concentrate in Urban Areas

As another example, researchers can filter spending on highway planning and construction in FY 2024 by setting the awarding top-tier agency as "DEPARTMENT OF TRANSPORTATION (DOT)," the awarding subtier agency as "FEDERAL HIGHWAY ADMINISTRATION," and the assistance listing to "20.205—Highway Planning and Construction," as seen in Figure C.2.

The heat map reveals spending concentrates in urban areas, with more spending in populated metropolitan areas, such as Washington, D.C. Researchers interested in modeling the budgetary implications of infrastructure spending can extract relevant USAspending.gov data from the heat map and incorporate it into models.

Figure C.2. Interactive Heat Map of the Federal Highway Administration's Spending



Appendix D. Version History

Table D.1 describes the version history of this document and the tool.

Table D.1. Version History

Version	Release Date	Major Changes
0	August 19, 2025	Initial document
1.0	July 2026	Added Chapters 4 and 5 and Appendix C Significant updates to Chapter 6 General edits and corrections throughout

Abbreviations

ACS	American Community Survey
CGE	computable general equilibrium
COMPARE	Comprehensive Assessment of Reform Efforts
CPS	Current Population Survey
DATA	Digital Accountability and Transparency
FIPS	Federal Information Processing Standard
FPDS	Federal Procurement Data System
FPL	federal poverty level
FY	fiscal year
GAO	U.S. Government Accountability Office
GDP	gross domestics product
GSP	gross state product
GTAS	Government Treasury Account Symbol Adjusted Trial Balance System
IDV	indefinite delivery vehicle
I/O	input/output
MEPS	Medical Expenditure Panel Survey
NAICS	North American Industrial Classification System
PAM	Partitioning Around Medoids
SAM	System for Award Management
SCF	Survey of Consumer Finances
SIPP	Survey of Income and Program Participation
SNAP	Supplemental Nutrition Assistance Program
SOI	Statistics on Income
TAS	Treasury Account Symbol
VA	U.S. Department of Veterans Affairs
WIC	Special Supplemental Nutrition Program for Women, Infants, and Children

References

Agency for Healthcare Research and Quality, "Medical Expenditure Panel Survey," webpage, undated. As of August 8, 2024: https://meps.ahrq.gov/mepsweb/

Board of Governors of the Federal Reserve System, "2022 Survey of Consumer Finances," dataset, 2023. As of August 14, 2025: https://www.federalreserve.gov/econres/scfindex.htm

Bureau of Economic Analysis, "Welcome to the RIMS II Online Order and Delivery System," webpage, undated. As of July 8, 2026: https://apps.bea.gov/regional/rims/rimsii/

Bureau of the Fiscal Service, U.S. Department of the Treasury, "About the Data Transparency Program," webpage, last updated November 16, 2023. As of August 8, 2024: https://fiscal.treasury.gov/data-transparency/history-overview.html

Carpenter, C. W., M. C. Lotspeich-Yadao, and C. M. Tolbert, "When to Use Commuting Zones? An Empirical Description of Spatial Autocorrelation in U.S. Counties Versus Commuting Zones," *PLOS One*, Vol. 17, No. 7, July 2022.

Economic Research Service, U.S. Department of Agriculture, "WIC Program," webpage, last updated July 22, 2025. As of June 11, 2026: https://www.ers.usda.gov/topics/food-nutrition-assistance/wic-program/

Federal Audit Clearinghouse, homepage, undated. As of August 12, 2025: https://www.fac.gov/

FitzGerald, Kathleen. Ann E. Futrell, Susanne Mehlman, and Emily Stern, *Common Budgetary Terms Explained*, Congressional Budget Office, December 2021.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry, "IPUMS CPS," Version 11.0, dataset, IPUMS, University of Minnesota, 2024. As of October 7, 2024: https://cps.ipums.org/cps/

Food and Nutrition Service, U.S. Department of Agriculture, "Examination of Cash Nutrition Assistance Program Benefits in Puerto Rico," webpage, last updated January 24, 2025a. As of June 11, 2026: https://www.fns.usda.gov/research/snap/examination-cash-nutrition-assistance-program-benefits-puerto-rico

Food and Nutrition Service, U.S. Department of Agriculture, "National School Lunch Program," webpage, last updated December 11, 2025b. As of June 11, 2026: https://www.fns.usda.gov/nslp

Food and Nutrition Service, U.S. Department of Agriculture, "Supplemental Nutrition Assistance Program (SNAP)," webpage, last updated March 12, 2026a. As of June 11, 2026: https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program

Food and Nutrition Service, U.S. Department of Agriculture, "SNAP Program Administration," webpage, last updated May 29, 2026b. As of June 11, 2026: https://www.fns.usda.gov/snap/admin

GAO—*See* U.S. Government Accountability Office.

Girosi, Federico, Amado Cordova, Christine Eibner, Carole Roan Gresenz, Emmett B. Keeler, Jeanne S. Ringel, Jeffrey Sullivan, John Bertko, Melinda Beeuwkes Buntin, and Raffaele Vardavas, "Overview of the COMPARE Microsimulation Model," RAND Corporation, WR-650, 2009. As of October 9, 2024: https://www.rand.org/pubs/working_papers/WR650.html

Gower, J. C., "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, Vol. 27, No. 4, 1971.

HealthCare.gov, "Medicaid Expansion & What It Means for You," webpage, undated. As of June 26, 2026: https://www.healthcare.gov/medicaid-chip/medicaid-expansion-and-you/

IMPLAN, homepage, undated. As of July 9, 2026: https://implan.com/

Internal Revenue Service, "SOI Tax Stats—Statistics of Income," webpage, last updated March 17, 2026. As of June 11, 2026: https://www.irs.gov/statistics/soi-tax-stats-statistics-of-income

IPUMS CPS, homepage, University of Minnesota, undated. As of October 7, 2024: https://cps.ipums.org/cps/

IPUMS USA, homepage, University of Minnesota, undated. As of October 7, 2024: https://usa.ipums.org/usa/

Keisler-Starkey, Katherine, Lisa N. Bunch, and Rachel A. Lindstrom, *Health Insurance Coverage in the United States: 2022*, U.S. Census Bureau, September 2023.

Leontief, Wassily W., *The Structure of American Economy, 1919–1929: An Empirical Application of Equilibrium Analysis*, Harvard University Press, 1941.

Leontief, Wassily W., *The Structure of American Economy, 1919–1939: An Empirical Application of Equilibrium Analysis*, 2nd ed., Harvard University Press, 1951.

Leulescu, Aura, and Mihaela Agafiţei, *Statistical Matching: A Model Based Approach for Data Integration*, Eurostat, European Commission, 2013.

Murtagh, Fionn, and Pedro Contreras, "Methods of Hierarchical Clustering," arXiv, arXiv:1105.0121, April 30, 2011.

Norris, Louise, "Will You Receive an ACA Premium Subsidy?" HealthInsurance.org, last updated January 27, 2026. As of June 26, 2026: https://www.healthinsurance.org/obamacare/will-you-receive-an-aca-premium-subsidy/

Price, Carter C., and Kathryn A. Edwards, "Trends in Income from 1975 to 2018," RAND Corporation, WR-A516-1, 2020. As of October 7, 2024: https://www.rand.org/pubs/working_papers/WRA516-1.html

Ready, Dan, Jorge Salazar, and Caitlin Verboon, "Mandatory Spending in Fiscal Year 2024: An Infographic," Congressional Budget Office, March 20, 2025.

Reusova, Anastasia, "Hierarchical Clustering on Categorical Data in R," *Towards Data Science*, April 1, 2018. As of June 11, 2026: https://medium.com/data-science/hierarchical-clustering-on-categorical-data-in-r-a27e578f2995

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler, "IPUMS USA," Version 15.0, dataset, IPUMS, University of Minnesota, 2023. As of October 7, 2024: https://www.ipums.org/projects/ipums-usa

SAM.gov, "Assistance Listings," webpage, U.S. General Services Administration, undated-a. As of August 14, 2025: https://sam.gov/assistance-listings

SAM.gov, homepage, U.S. General Services Administration, undated-b. As of October 8, 2024: https://sam.gov/content/home

USAFacts, homepage, undated. As of August 8, 2024: https://usafacts.org

USAspending.gov, "Agency Profiles," webpage, undated-a. As of June 26, 2026: https://www.usaspending.gov/agency

USAspending.gov, "Data Dictionary," dataset, undated-b. As of October 10, 2024, https://www.usaspending.gov/data-dictionary

USAspending.gov, "Explore the Spending Landscape," webpage, undated-c. As of August 8, 2024:
https://www.usaspending.gov/explorer

USAspending.gov, "SQL Database Downloads," webpage, undated-d. As of June 26, 2026:
https://onevoicecrm.my.site.com/usaspending/s/database-download

USAspending.gov, "Using This Table," webpage, undated-e. As of June 12, 2026:
https://www.usaspending.gov/submission-statistics/data-sources

U.S. Code, Title 31, Money and Finance; Subtitle V, General Assistance Administration; Chapter 61, Program Information; Section 6101, Definitions.

U.S. General Services Administration, *Federal Procurement Data System Product and Service Codes (PSC) Manual: Fiscal Year 2024 Edition*, April 2024.

U.S. Government Accountability Office, *Federal Spending Transparency: Opportunities Exist to Further Improve the Information Available on USAspending.gov*, November 8, 2021. As of August 8, 2024:
https://www.gao.gov/products/gao-22-104702

U.S. Government Accountability Office, *Federal Spending Transparency: Opportunities to Improve USAspending.gov Data*, November 7, 2023a. As of August 8, 2024:
https://www.gao.gov/products/gao-24-106214

U.S. Government Accountability *Office, Federal Spending Transparency: Opportunities Exist to Improve COVID-19 and Other Grant Subaward Data on USAspending.gov*, November 16, 2023b. As of August 8, 2024:
https://www.gao.gov/products/gao-24-106237

U.S. Office of Management and Budget, *Preparation, Submission, and Execution of the Budget*, Circular No. A–11, 2023.

U.S. Office of Management and Budget and Federal Reserve Bank of St. Louis, "Federal Net Outlays as Percent of Gross Domestic Product (FYONGDA188S)," August 5, 2024. As of October 7, 2024:
https://fred.stlouisfed.org/series/FYONGDA188S