Social accounting matrix for New Jersey 2022 with methods and sources for other U.S. states

María T. Álvarez-Martínez, Universidad Pablo de Olavide

Michael L. Lahr, Rutgers, The State University of New Jersey

João F.D. Rodrigues, Digital Solutions, The Netherlands

Version generated 2026-05-26 21:40 UTC

ABSTRACT. In this paper, we show how to build a social accounting matrix (SAM) for a U.S. state. We use New Jersey in 2022 as an example. In so doing, we define how SAMs are different from extended input-output tables. We cursorily discuss how iterative proportional fitting can be applied to fill in missing data elements of key component datasets. Ultimately, we focus on general structural features that are either additional to or missing in most subnational SAMs compared to their national equivalents, particularly with respect to institutions. This elaboration follows from available data resources, albeit specifically for the case of U.S. states. We document which SAM elements can be constructed from existing federal statistical sources and which cannot, and we note the data elements not currently available from official statistical agencies. We also discuss how aggregation is best applied when such SAMs are used in computable general equilibrium (CGE) models.

1. Background & Summary

A social accounting matrix (SAM) is a balanced matrix that captures the circular flow of income within an economy for a specific period. SAMs complete information provided by input-output (IO) tables with the data included in national accounts and show the interaction among production processes, details about the generation of income, and how this income is used.

Stone (1947, 1956, 1961) laid the foundations for SAMs, which have matured considerably since. Pyatt and Round (1985) developed the canonical multi-country treatment; Round (2003) provides a pedagogical introduction. The *System of National Accounts* (SNA-93, United Nations, 1993; SNA-08, European Commission et al., 2009; SNA-2025, European Commission et al., 2025) are guidelines from the United Nations for constructing SAMs for a national or regional economies. The recent 2025 update retains the basic SAM framework but elaborates on aspects (globalization, digitalization, well-being) that have recently come into prominence. In the last 30 years, the construction and application of SAMs has expanded rapidly, largely in support of computable general equilibrium (CGE) models, which enable the evaluation of a large variety of economic issues. According to the SNA-93, section 20.105, the structure of SAMs is flexible and depends on the purposes of both the analyses to be undertaken and the availability of data to construct them.

Traditionally, SAMs have been built for a single country or region to evaluate specific issues related to public policies (Álvarez-Martínez and Polo, 2012), tourism (Jones, 2010) or agriculture and energy (Cardenete et al., 2014). In recent years, researchers have developed a veritable boom of SAM databases. Examples are the World Input-Output Database (WIOD) project (Timmer, ed., 2012), the WIOD SAMs adjusted with data from Eurostat for the European Union 27-member states (EU-27) (Álvarez-Martínez and López-Cobo, 2018) and the BioSAMs

(Mainar et al., 2021). A common characteristic of these databases is their development for EU member states. To date, few SAMs have been published for U.S. states. Most regional SAMs developed in the US have been prepared using prepackaged software and data. A prime issue with such prepackaged offerings is knowing aspects of the SAMs that are fabricated versus those grounded in official data. Our prime contribution is to articulate all this in as disaggregate a manner as possible in terms of industries, taxes sources, and types of government sectors. This document presents how to build a 2022 SAM for New Jersey (SAMNJ22), including the data sources.

2. Methods

SAMNJ22 takes as its core a 402-industry 2022 industry-by-industry direct requirements matrix for New Jersey. It is a rendition of the 2017 U.S. BEA national benchmark Supply and Use Tables (SUTs) that is updated and regionalized using a sequence of federal data products. We outline the procedure in the next paragraph and detail it in subsequent sections.

A key data source to our approach is Quarterly Census of Employment and Wages (QCEW) data from the U.S. Bureau of Labor Statistics (BLS). We deploy it to supply state industry details: number of establishments, count of employees, and the total payroll of employees.[1] In turn, after some enhancements to these data so they reflect other state data elements from BEA, we ultimately use QCEW state industry detail to identify state industry output for each of the industries in the benchmark SUTs. We then apply BEA's state value-added data on 65 sectors to adjust the national direct requirements, so they reflect state productivity differences. We estimate 402-industry detail for state-level household expenditures using state-level data that BEA reports for 17 commodities. Other state final demand sectors are mostly estimated as shares of the national equivalent using state value totals. To adjust for interstate trade leakage (commodity inflows and outflows), we econometrically estimate regional purchase coefficients (RPCs) using US Bureau of Transportation Statistics (BTS) commodity-shipment data by mode for goods-producing industries. To help validate the resulting state accounts, we adjust them via an iterative proportional fitting approach to make sure they sum to the national equivalent.

2.1 Source data and the technology update

As noted above, five data sources do most of the work. We also use 2022 state aggregate GDP, personal income, and population estimates to share out components of final demand and to set effective tax rates on households. Except where noted, all source data used in the accounts are for the 2022 reference year; the principal exception is the structural production functions, which are drawn from the 2017 benchmark SUTs and updated to 2022 control totals. The below describes the process.

National SUTs Updating.
The production functions are those from BEA's 402-industry 2017 benchmark SUTs. The 71-

[1] We do not use the more commonly cited County Business Patterns (CBP) data. QCEW is preferred to CBP because (1) it is an annual census of employed workers (CBP is a count of workers after a year's first quarter only), (2) it covers more workers (CBP omits self-employed, agricultural, and several other categories).

industry annual BEA national SUTs provide the aggregate control totals to which one can update the benchmark accounts.[2] We apply a block updating approach in the spirit of St. Louis (1989).

2.2 Handling QCEW disclosure problems

QCEW reports include information on industries down to as much as six-digit North American Industry Classification System (NAICS) codes.[3] For 2022, nation totals are available for 1,394 industries. The data set also delineates sectoral data for up to four types of establishment ownership (private industry and three types of governments: federal, state, local). The data are available for the nation, 51 states (including the District of Columbia), and 3,164 counties. As all U.S. federal providers do, BLS suppresses aggregate information on private industries that have three or fewer establishments to protect individual privacy, safeguard corporate trade secrets, ensure national security, or otherwise discourage simple data snooping. Further, if one industry is suppressed, data for a sibling industry (one at the same level of industry disaggregation) necessarily is also suppressed, even though it may have a sufficient count of establishments to permit disclosure of its information. Information for the parent industry (that at the next level of NAICS aggregation) is typically reported, however. We data mine to estimate values for all such disclosure problems in the QCEW data set.[4]

The point is that we cannot treat nondisclosed values as zeros. We impute them so they respect reported parent sector values (i.e., higher-level aggregates). A fortunate thing is that counts of establishments in the QCEW are always reported if there are any establishments at all. With this in mind, we make an initial estimate of jobs for suppressed industries for each combination of geography and ownership type. We do so by using the industry's jobs/establishment ratio at the next higher geographic level; for states this means we use national ratios. We then use a variation on iterative proportional fitting that holds the higher-level aggregates fixed and distributes the discrepancy across the unsuppressed cells, a reliability-weighted reconciliation (see, Rodrigues, 2014; Rodrigues and Lahr, 2018).

After estimating suppressed jobs for all detailed sectors, we focus on suppressed annual payroll figures using the payroll/job ratio for suppressed values. Again, we do so using values from the geographic level above the focal one within the spatial hierarchy.

Two difficulties occasionally arise prior to balancing that must be handled in some fashion. First, QCEW very occasionally reports zero jobs that receive nonzero wages. This contradiction must be resolved before the cell enters the matrix-balancing step. Second, the

---

[2] We prefer these SUTs for updating over those available at a more disaggregated level from the US BLS. This is because BEA staff report that BLS tables do not embody either BEA value added updates or Census Bureau technology updates

[3] These data are collected by each state and microdata from the Federal-State Unemployment Insurance (UI) programs are the primary source. States receive a Quarterly Contributions Report (QCR) from all private sector employers, as well as from state and local governments covered under the UI program. Federal government employers report via the Report of Federal Employment and Wages (RFEW), which contain only employment and wages data by worksite (establishment).

[4] While our focus here is on states, we also estimate values to fill-in detailed QCEW industry disclosure problems for all 3,164 U.S counites as well. Indeed, the sum of county-level estimates in a state sets a lower bound for any nondisclosed state-level data.

published QCEW aggregates do not always reconcile across scales: state totals do not always sum to the nation.[5]

2.3 Disclosure problems in BEA data

State industry labor compensation and earnings data, jobs data, and value added for four components (compensation, indirect business taxes, subsidies, and GOS) for 2022 are also subject to disclosure problems. All are published by BEA and at industry levels far more aggregated than are the benchmark SUTs. We inform the suppressed elements in compensation of employees using aggregated-equivalent QCEW payroll data before balancing as above. Subsequently, we use that result to inform the labor compensation component of value added, which is generally identical to the compensation of employees just estimated. We then turn to labor earnings. Balancing is, of course, requisite in each case. Indirect business taxes and subsidies tend to be fully disclosed. GOS need not be disclosed but can be estimated as a residual when the industry value added total is reported. Matrix balancing again is applied so that industry sums of the components across states equal national totals.

2.4 Residual data issues

A handful of issues are tough to routinize. Federal SUT categories include overseas spending. National totals therefore are tough to match to a sum of state-level equivalents absent an explicit residence adjustment. Some QCEW grand totals do not add up across geographic scales even after balancing. We suspect this reflects the cumulative effect of suppression, ownership-class reclassification, and poor address matching by state labor department staff rather than measurement error. Oddly, a nontrivial number of Use-table elements switch between zero and non-zero between the 2017 benchmark and 2022. Some of these switches are substantive (a sector enters or exits). Others reflect revisions to BEA's reporting conventions. BEA's data on personal consumption expenditures (PCE) by state show non-zero entries (occasionally negative, in religious and grantmaking organizations) where SUT household consumption records zeros. Net taxes on government activity in value-added accounts are non-zero where the SUT shows zeros. Where these conflicts arise, we hold the SUT structure fixed, document the affected cells, and let the residuals fall into balancing error rather than push them into the productive accounts.[6]

2.5 Bridging QCEW industries to SUT industries

QCEW industries and SUT industries are close but not identical. The bridge matrix that BEA publishes is not always practical for use by researchers outside of BEA itself. It suggests splits of detailed NAICS, uses NAICS not available in the QCEW, or develops sectors that unnecessarily present amalgamated technology. The last makes models that use them somewhat less useful.

We confront such issues in three ways. First, several QCEW government enterprises (hospitals, utilities, education, transit, farms) have private-sector equivalents in the SUT that the published bridge does not capture. We allocate them to the private-sector equivalent.[7] Second, BEA spreads several very detailed NAICS across more than one SUT sector. Construction is one such case; automobile manufacturing is another. Construction poses the bigger problem as

---

[5] And the sum across counties for an industry do not always equal that of the state.

[6] In some instances, the alternative would be to introduce ad hoc reallocations. We avoided this. Such reallocations would be impossible to justify either from theory or from the source data.

[7] Our thought here is that practical uses of an amalgamous government enterprise sector are few.

QCEW uses labor-oriented NAICS[8] while BEA SUTs use NAICS affiliated with types of infrastructure.[9] Thus, any resulting practical bridge is apt to be workmanlike and somewhat unprincipled. As a rule, however, we assign QCEW information to a single SUT industry. In this way, the content of a given SUT sector is well defined. Third, NAICS codes drift over time. Some are renumbered, and others are formally retained but folded into a sibling (neighboring) industry, such that the original record is empty of information (effectively implying it is zero).[10] Still others split or merge in ways that do not propagate identically across datasets. Each case is identified and carefully handled with an eye favoring consistency in assignment.

2.6 Allocating BEA state sectoral data to more-disaggregate SUT industries

Once all data are aligned and "filled in" at the state level, we "enhance" QCEW SUT aggregates of jobs and payroll, so they sum to the far more complete BEA state information on jobs and compensation reported only in a somewhat more aggregate set of sectors.[11] The same is done for sectorwise state value-added data by component. We do this by assuming the SUT-detail QCEW-informed data report the proper subindustry mix within the more-aggregate state-based BEA data equivalents. We, thus, spread the state-based BEA data across the SUT detail according to SUT subindustry shares.

Formally, for a generic state series reported by BEA for *b* aggregate sectors and spread to the 402 SUT industries by their QCEW-informed shares,

$$\mathbf{x} = \hat{\mathbf{w}}\, {}^{\mathbf{b}}\mathbf{\Gamma}^{402\prime} \left({}^{\mathbf{b}}\mathbf{\Gamma}^{402}\, \hat{\mathbf{w}} \right)^{-1} \bar{\mathbf{x}} \qquad (1)$$

where $\mathbf{x}$ is the 402-vector of the series at SUT-industry detail; $\bar{\mathbf{x}}$ is the b-vector of the same series as BEA reports it for its aggregate sectors; $\hat{\mathbf{w}}$ is the 402×402 diagonal matrix of QCEW-informed SUT-industry weights; and the aggregator (left- and right-superscripted by its output and input dimensions) is the $b \times 402$ zero–one matrix that sums the SUT industries belonging to each BEA aggregate sector. The bracketed inverse turns the weights into within-sector shares, so each BEA aggregate is allocated across its SUT industries in proportion to their QCEW weights; summing the allocated values back over each sector returns the BEA total.

2.7 From national to state-level coefficients

To enable an easier understanding of the technological framework that we apply, we now review some basic SUT principles and define notation. Upper-case bold letters represent matrices and lower-case bold letters represent vectors.

---

[8] For example, general, site preparation, structural, systems, finishing, heavy and civil engineering.

[9] For example, both new single- and multi-family structures, nonresidential structures; utility, highways and streets, and bridges; education hospital and health structures, and both residential and nonresidential maintenance and repair.

[10] For example, NAICS 333315 was a perfectly good photographic-and-photocopying-equipment manufacturing code in 2007. It became 333316 in 2017 and is now empty.

[11] U.S. BEA discontinued publishing jobs data by industry for states and counties starting for the year 2023. Indeed, pre-existing data on jobs by industry for 1929-2022 are now archived by BEA and, hence, difficult to find.

Figure 1. Scheme of Supply-Use IO Tables

	Commodities	Industries	Final Demand	Total Output
Commodities		$\mathbf{U}$	$\mathbf{F} - \mathbf{m}$	$\mathbf{q}$
Industries	$\mathbf{S}$			$\mathbf{g}$
Value Added		$\mathbf{W}$		
Total Input	$\mathbf{q}'$	$\mathbf{g}'$		

Note: The supply table **S** is commodity-by-industry; column sums of **S** equal industry output **g**, and row sums of **S** equal commodity output **q**.

First, we estimated the Use coefficients matrix **B**:

$$\mathbf{B} = \mathbf{U}\,\hat{\mathbf{g}}^{-1} \tag{2}$$

where **U** is the 2017 U.S. Use matrix and $\hat{\mathbf{g}}$ is the diagonal matrix of U.S. industry output—the column sums of **S** (equivalently, the column sums of **U** plus value added).

Second, we estimated the commodity-output proportions matrix **C** as follows:

$$\mathbf{C} = \mathbf{S}\,\hat{\mathbf{q}}^{-1} \tag{3}$$

where **S** is the U.S. Supply table and $\hat{\mathbf{q}}$ is the diagonal matrix of total output by commodities—the row sums of **S**. The (i, j) entry of **C** is the share of commodity i's total supply produced by industry j. From **B** and **C** we estimate the industry-by-industry direct input coefficient matrix, **A**, for New Jersey using St. Louis's (1989) proposal in supply-use form:

$$\mathbf{A} = \mathbf{C}'\,\hat{\boldsymbol{\rho}}^{us}\,\mathbf{B} \tag{4}$$

where $\hat{\boldsymbol{\rho}}^{us}$ is the elementwise (Hadamard) ratio $(\mathbf{q} - \mathbf{e})/(\mathbf{q} - \mathbf{e} + \mathbf{m})$, where $\mathbf{e}$ is international exports at the national level. This re-export-adjusted domestication (Lahr, 2001) modifies the national technology to account for international-trade leakages for the State of New Jersey; it is a slight modification of the Miller and Blair (2022) taxonomy that adjusts for re-exports, a more pronounced issue at a subnational level.

There are two points that are important to note in the above. First, we use industry-by industry technology. Second, our approach allocates international imports to states using compensation shares and international exports using output shares. We discuss each in turn in the following sections.

2.7.1 *Choice of technology assumption*

Equation (4) is the industry-technology variant of the supply-use transformation—Model B in the Miller and Blair taxonomy—which assumes that each industry operates a single input recipe regardless of the mix of commodities it happens to produce. The alternative commodity-technology assumption holds the recipe constant by *commodity* rather than by industry and yields a commodity-by-commodity table; the two formulations are compared in Table 1.

Table 1. Industry- and Commodity-Technology Forms of A under Trade Leakage

Technology assumption	Industry-by-industry / commodity-by-commodity A
Commodity technology (Model A)	$\mathbf{A}^{CC} = \hat{\boldsymbol{\rho}}^{us}\mathbf{B}\,\mathbf{C}^{-1}$
Industry technology (Model B)	$\mathbf{A}^{II} = \mathbf{C}'\,\hat{\boldsymbol{\rho}}^{us}\,\mathbf{B}$

Source: Authors' compilation. $\hat{\boldsymbol{\rho}}^{us}$ is the diagonal matrix of all international imports, following Lahr's (2001) domestication scheme, a slight modification of the Miller and Blair (2022) taxonomy that adjusts for re-exports. $\mathbf{C}^{-1}$ exists only when $\mathbf{S}$ is square (industries and commodities at the same aggregation, as in SAMNJ22).

We adopt the industry-technology form on four grounds. First, the resulting $\mathbf{A}$ is guaranteed non-negative, whereas $\mathbf{C}^{-1}$ can carry negative entries when secondary production is large relative to a commodity's total supply, producing direct-requirements coefficients with no economic interpretation. Second, we prefer industry-by-industry accounts because almost all U.S. subnational data are available with industry detail only. In fact, the only subnational data available in the U.S. are those focused on household consumption and interregional trade. A third reason for selecting industry-by-industry configuration over its commodity-by-commodity equivalent is that most policies are directed to firms rather than their products. That is, most analyses at the subnational level are likely to be industry- rather than product-oriented. Fourth, perhaps least importantly, most CGE modeling platforms expect state-level SAMs to be industry-by-industry. Ultimately, we think it is best practice to adhere to the industry-by-industry scheme in a subnational setting.

2.7.2 *Total industry output*

Armed with the national benchmark SUT updated to 2022 as discussed in Section 3.1 and state-level data in SUT detail, we can now estimate state output for each benchmark industry. We set each industry's state output to the state's share of national labor compensation in that industry.[12] That is, total output by commodities in New Jersey is the state's share of national labor compensation multiplied by output:[13]

$$\mathbf{g}^{nj} = \left(\frac{\boldsymbol{\upsilon}^{nj}}{\boldsymbol{\upsilon}^{us}}\right)\mathbf{g}^{us} \tag{5}$$

where $\boldsymbol{\upsilon}$ is a 1-by-402 vector of labor compensation and the superscripts *nj* and *us*, respectively denote New Jersey and the nation.

2.8 Breaking out tax revenues to jurisdictional levels

The model distributes net taxes across the three levels of government — federal, state, and local — separately for each industry and region. National accounts develop two conceptually distinct categories of taxes: taxes on production (levied on the industry, such as property taxes) are split

[12] Lahr (2001) suggests such an allocation assumes the productivity of labor is spatially constant.

[13] All operations in Equation (5) are Hadamard.

using measured regional revenue shares, while taxes on products, which are referred to here as commodity taxes (such as sales and excise taxes, levied on the commodity) are split using industry-specific rules.

2.8.1 Production taxes

Before the production tax is split across levels of government, it is worth noting how the production-tax amount itself varies across regions. It is not a national figure rescaled by industry size: it reflects real regional variation in taxes on production and imports as measured in BEA state value-added accounts. (The state adjustment of national value-added figures is treated elsewhere in this document.) The allocation below then divides the state-level adjustment across federal, state, and local governments.

Production taxes are split using state-specific government shares derived from revenue data reported annually to the U.S. Census Bureau. The federal component comes from business income tax. The state and local components are built from a fixed, weighted combination of four named revenue lines drawn from the U.S. Census Bureau's Annual Survey of State and Local Government Finances — specifically its 2022 Census of Governments: Finance edition, the full census conducted every five years (U.S. Census Bureau, 2024) —four lines from the Survey are applied to derive a state-local split: the entirety of Corporate income (line 19) and half the sum of Property (line 9), Motor vehicle license (line 20), and Other taxes (line 21). The reasoning is that these lines refer to state and local taxes that fall on production. Clearly, corporate income tax fully lands on businesses, whereas property taxes, motor vehicle license fees, and other taxes are also borne by households.[14] The consumption taxes — general and selective sales, individual income — are deliberately excluded here, since they belong to the commodity-tax side of the model. The three components are then normalized into fractions that sum to 1, and each industry-region's net production tax is allocated according to those fractions. Thus, the state/local split reflects variations specific to each state.

A caveat arises where the net production tax is negative — that is, where subsidies exceed tax revenues received — the entire amount is assigned to the federal level and, hence, state and local shares are set to zero.

2.8.2 Commodity taxes

The commodity-tax amount also varies across states. Unlike the production-tax figure, it is distributed from the national total to states in proportion to its share of national payroll (with state GDP as a fallback if the wage-based split produces inconsistent accounts). A balancing residual is applied where necessary. The reported division of state and local revenue governs the split.

To implement the split, we use the following lines from the Annual Survey of State and Local Government Finances to the relevant industry tax bases:

- **General sales (line 11)** — the bulk of wholesale and retail trade industries, whose tax base is broad sales.
- **Selective sales (line 12)** — one specialized wholesale code.
- **Motor fuel (line 13)** — petroleum refining (324110) and petroleum wholesalers.

[14] A 50:50 split is applied in the absence of guiding information.

- **Public utilities (line 16)** — the electric, gas, and water utility industries (221100, 221200, 221300) and a government electric-utility code.
- **Other selective sales (line 17)** — rental and leasing, professional services, and repair and personal services.
- **Current charges (line 23)** — state-and-local government enterprise codes.
- **Education (line 24) and Hospitals (line 27)** — the government education and health codes respectively, matching those services to the charges governments levy for them.

2.8.3 Household taxes

Households are treated as a separate sector rather than as part of the industry-by-industry allocation described above. The split is structurally similar to the production-tax approach but draws on partly different sources and is built on individual income.

Federal household tax comes from total individual income tax by state, processed from the Internal Revenue Service Statistics of Income table — "Returns with Positive Adjusted Gross Income" (U.S. Internal Revenue Service, 2023b) —that reports total income tax by state; the column the model reads is explicitly headed total income tax in thousands of dollars, and its national total is consistent with 2022 federal individual income tax. The effective federal tax rate on household income is calculated by dividing that figure by aggregate personal income reported by the BEA.

State and local household taxes are drawn from the same Census Bureau Annual Survey of State and Local Government Finances used for production taxes, but with a different set of named lines and weights. In addition to Individual income (line 18), we use the remaining half of the sum of Property (line 9), Motor vehicle license (line 20), and Other taxes (line 21) not counted in production taxes (see above). Both the state and local split of these household figures are then divided by the BEA-reported state aggregate personal income (table CAINC4).

Table 2. Federal Government Revenues and Expenditures

Revenues	
Tax revenues received by the Federal Government (SAMNJ22) in $ millions	
Income tax	**154,944.100**
Data from IRS Data Book 2022	
-Corporate income tax revenues	29,377.431
-Personal income tax revenues*	125,566.669
SSC, employers	**24,504.678**
Data from BEA	24,504.678
SSC, employees	**29,160.205**
Data from BEA	29,160.205
Taxes on production and imports	**16,502.147**
Data from the Annual Survey of State and Local Government Finances, Fiscal Year 2022: New Jersey (U.S. Census Bureau, 2022a)	
-Taxes on production and imports: total revenues (BEA)	70,680.874
-Taxes on sales (State & Local Government Finances: 2022)	-17,965.046
-Taxes on production received by SLGNPISH (State & Local Government Finances)	-36,213.681
Expenditures	

Revenues	
Current transfers paid by the Federal Government (SAMNJ22)	
To households	**105,228.690**
Data from BEA	
To state & local government + NPISH (A+B)	
(A) State & Local Government. Data from State & Local Government Finances:2022	8,411.339
-Intergovernmental revenues	26,019.270
-Intergovernmental expenditure	-17,607.931
(B) NPISH. Data from BEA	**I did not find it**
*Net of SSCH (BEA), $29,160.205 (January 2023)	
Source: BEA and State & Local Government Finances: 2022: New Jersey (January, 2023)	

2.9 Interindustry demands

An estimate of the inter-industry demands in New Jersey is obtained by post-multiplying **A** by a diagonal matrix of New Jersey total output:

$$\mathbf{Z}^{nj} = \mathbf{A}\,\hat{\mathbf{g}}^{nj} \tag{6}$$

2.10 *Regional final demand*

Final demand is a 402-by-9 matrix (**F**) of industry-by-final-demand accounts. The nine accounts are: households' consumption, private investment, exports, nondefense federal investment and consumption, investment and consumption for national defense, and state and local government investment and consumption. State-level household consumption is built up from the U.S. Bureau of Economic Analysis's Personal Consumption Expenditures by State (SAPCE) for 2022, which BEA publishes only at a coarse classification of roughly 20 functional expenditure categories. To reach the commodity detail the benchmark accounts require, each state's expenditure category is disaggregated to the 402 benchmark commodities using expenditure shares drawn from the 2022 Consumer Expenditure Survey (U.S. Bureau of Labor Statistics, 2022a). The remaining components are first assembled at the national level from the 2022 benchmark accounts and then allocated to states by component-specific rules. Private investment, exports, and changes in inventories are spread to states in proportion to each state-industry's share of national output: the national ratio of the component to output is applied to state-level output. Government consumption and investment, by contrast, are allocated using dedicated government proxies, with each state receiving a share of the national figure equal to its share of the relevant government-activity proxy. All components are estimated in commodity space and converted to the 402 benchmark industries through the market-share matrix **C**, and personal consumption and government are further disaggregated to counties using personal-income shares.

Table 3. State and Local Government Revenues and Expenditures

Revenues	
Tax revenues received by SLGNPISH (SAMNJ22) in $ millions	
Income tax	**22,793.26**
Data from State & Local Government Finances: 2022 State, New Jersey	
-Corporate income tax revenues	5,959.760
-Personal income tax revenues	16,833.495
Other taxes on production	
Data from State & Local Government Finances: 2022 State: New Jersey	
-Property taxes	32,793.053
-Motor vehicle license taxes	681.061
-Other taxes	2,739.567
Taxes on sales	
Data from State & Local Government Finances: 2022 State, New Jersey	
-General taxes on sales	12,803.267
-Selective taxes on sales	5,161.779
Current transfers received by SLGNPISH (SAMNJ22)	
From corporation (to NPISH)	
Data from BEA	907.756
From federal government	
Data from State Government Finances:2022: State: New Jersey	8,411.339
Expenditures	
Current transfers and subsidies paid by the SLGNPISH Government (SAMNJ22)	
To households (from NPISH)	
Data from BEA	2,218.302
To unemployment benefits	
Data from BEA	1,787.353
To subsidies on production	
Data from BEA in 2022 and from State and local Government finances in 2022 (assistance and subsidies)	**1,922.020**
Source: BEA and State Government Finances: 2022, State: New Jersey	

Formally, with the 20 SAPCE functional categories spread across the 402 benchmark commodities by their Consumer Expenditure Survey shares, state household consumption is

$$\mathbf{h} \;=\; \hat{\mathbf{c}}\,{}^{20}\boldsymbol{\Gamma}^{402\,\prime}\left({}^{20}\boldsymbol{\Gamma}^{402}\,\hat{\mathbf{c}}\right)^{-1}\mathbf{p} \tag{7}$$

where $\mathbf{h}$ is the 402-vector of state household consumption by benchmark commodity (the household-consumption column of F); $\mathbf{p}$ is the 20-element vector of state PCE by SAPCE functional category; $\hat{\mathbf{c}}$ is the 402×402 diagonal matrix of Consumer Expenditure Survey expenditures by benchmark commodity; and the aggregator (left- and right-superscripted by its output and input dimensions) is the 20×402 aggregation matrix that sums the benchmark commodities belonging to each SAPCE category. Post-multiplying by the bracketed inverse converts those commodity weights into within-category shares, so each category total is split across its commodities in proportion to their survey expenditures.

We treat household consumption as a single column, but the social-accounting framework readily supports more. A SAM extends the IO core by closing the loop from value added through institutions back to final demand, and a common elaboration splits the household account into groups—by income class, age of householder, or intra-regional location—so that both the consumption column and the labour-income row are disaggregated in parallel. Doing so lets the model trace how a policy shock reaches different kinds of households and how their spending feeds back through the regional economy, which is what equity- and distribution-oriented analyses require. Lahr et al. (2021), for instance, build a regional CGE on accounts of this kind—four household groups by income level, with matching labour detail—to study how Philadelphia's sugar-sweetened-beverage tax and its daycare subsidies move low-income parents into the labour market. Others have parsed households by income, age, and location (Kim, Kratena, and Hewings, 2015; Kim, Hewings, and Kratena, 2016; Hewings et al., 2017; Cazcarro et al., 2022; Kratena, 2021, 2024). We keep a single representative household here, since disaggregating both labour compensation by industry and consumption by household type—and estimating each group's commuting propensities—is data- and budget-intensive; the accounts are nonetheless structured so that the split can be introduced when the data warrant it.

2.10.1 International Imports and Exports

International imports and exports enter SAMNJ22 accounts through the national benchmark Supply and Use tables and the national RPC introduced in Table 1. We use two proxies. International imports for each industry are allocated to New Jersey by its share of national labor compensation,

$$\mathbf{m}^{nj} = (\boldsymbol{\upsilon}^{nj} / \boldsymbol{\upsilon}^{us})\, \mathbf{m}^{us}, \tag{8}$$

and international exports by its share of national industry output,

$$\mathbf{e}^{nj} = (\mathbf{g}^{nj} / \mathbf{g}^{us})\, \mathbf{e}^{us}. \tag{9}$$

Such allocations treat all states equally with respect to international imports, even though intuition suggests producers in states with major ports are more apt to secure imports and dispatch exports than producers in landlocked states.

A direct, data-based treatment of foreign trade is a natural future refinement, and some relevant subnational statistics do exist. International exports for New Jersey are available from the *Census Bureau* for four-digit NAICS (U.S. Census Bureau, 2022b), by commodity and by state of origin. The data on international imports currently match the 2017 NAICS. The *Census Bureau* also publishes data on six-digit Harmonized System (HS) commodity imports by state, though for our purposes these enter only at the four-digit NAICS level. A future version of the state accounts could provide far better estimates of foreign trade than the rough approximations we currently produce.

2.10.2 Regional Purchase Coefficient Estimation

An RPC is the fraction of a region's demand for a given commodity that is satisfied by production within that same region rather than by imports from other regions. It is a number between 0 and 1: a value near 1 indicates that the region is largely self-sufficient in the commodity, while a value near 0 indicates that the region imports almost all of it. RPCs are the central parameters that trade-regionalize a national IO table. Vectors of regional output, interstate trade flows, and the regional direct-coefficient matrices ultimately depend on them.

For goods-producing sectors we estimate RPCs econometrically, in the Stevens-Treyz tradition of regional purchase-coefficient models (Stevens et al., 1983; Treyz and Stevens, 1985; Lahr, Ferreira, and Többen, 2020) as extended to subnational and multi-region settings (Sargento et al., 2024; de la Torre Cuevas et al., 2026). We fit a beta regression of the known RPC on economic-geography predictors—regional supply and demand, land area, an industry location quotient, a commodity weight-to-value ratio, and commodity- and region-group indicators—using 2022 Freight Analysis Framework (FAF5; U.S. Bureau of Transportation Statistics and Federal Highway Administration, 2022) shipment data. The model accounts for just over half the variance in the known coefficients. The full specification, the rationale for each variable, the estimates and their standardized effects, and the fit diagnostics are in Appendix B (Tables B2–B4).

Trade in services is largely absent from shipment surveys, so the freight-based regression cannot generate RPCs for service industries (cf. Gervais and Jensen, 2019). For these sectors we instead set the RPC equal to the region's supply/demand ratio, censored from above at unity, the long-standing nonsurvey practice in this literature (Treyz and Stevens, 1985; Stevens, Treyz, and Lahr, 1989; Lahr, 2001). Lodging is treated separately, by a formula keyed on the industry's location quotient, for a reason peculiar to some services: a low service RPC need not mean that local demand is met from outside the region. It can instead mean that demand arrives from outside—visitors consuming a place-bound service whose production, capital, and labor stay immovably local (Persky, Doussard, and Wiewel, 2009). The same decoupling of self-sufficiency from mobility applies, with variation, to hospitals, universities, and parts of the entertainment sector. The detailed service and labor-row rules are given in Appendix B.

2.10.3 *Estimating Inflows and Outflows*

A core issue in subnational IO table estimation is how to model intra-national trade. Generally, detailed data on regional inflows and outflows of commodities do not exist. This is particularly the case in the U.S., where interregional trade is neither controlled nor heavily monitored. Thus, traditionally regional economists have tended to use regional data associated with output levels to "regionalize" IO accounts. The parameter estimates used to perform such regionalization, which is applied rowwise on IO tables, are termed RPCs as discussed in Section 3.10.2 above. Key data for estimating them are the levels of production and consumption by industry.

In the case of New Jersey, we therefore first had to estimate the share of demand fulfilled by commodities and services in New Jersey.

$$logit\,(\boldsymbol{\rho}) = \mathbf{X}\boldsymbol{\beta} \tag{10}$$

where $\boldsymbol{\rho}$ is the vector of regional purchase coefficients, whose conditional mean is modeled through the logit link shown; the response is taken to be beta-distributed, with the precision parameter and full likelihood given in Appendix B. The design matrix $\mathbf{X}$ has one row per region-commodity observation, and those rows hold each observation's demand (an element of the vector $\mathbf{d}$), its supply-demand ratio, its weight/value ratio, the region's share of U.S. land area, a vector of geographical binary variables, and another for industries, and $\boldsymbol{\beta}$ is the conformable vector of regression coefficients estimated by beta regression (Appendix B). RPCs for services are estimated using supply location quotients. All values of RPCs range between 0 and 1.

The interindustry demand matrix $\mathbf{Z}^{nj}$ identifies transactions within the nation that fulfill industry needs in New Jersey. To arrive at fully trade-adjusted IO accounts, RPCs are applied to

$\mathbf{Z}^{nj}$ to remove intermediate demands from it that are fulfilled by supplies from other states—imports from other states that are included in the SAMNJ22 as "inflows"—by simply multiplying the IO matrix:

$$\mathbf{Z}_{\boldsymbol{\rho}}^{nj} = \hat{\boldsymbol{\rho}}\,\mathbf{Z}^{nj} \tag{11}$$

as well as final consumption (households, federal government, and state and local government) and investment (private investment, federal government investment, and state and local government investment) by the RPCs:

$$\mathbf{F}_{\boldsymbol{\rho}}^{nj} = \hat{\boldsymbol{\rho}}\,\mathbf{F}^{nj} \tag{12}$$

where $\mathbf{F}^{nj}$ is a 402-by-6 matrix of industries-by-final-demand accounts.

Hence, the vector of regional inflows by producing industry, $\boldsymbol{\mu}$, can be calculated as the difference between the total demands and those fulfilled within the state:

$$\boldsymbol{\mu}' = \mathbf{i}'\mathbf{Z}^{nj} + \mathbf{i}'\mathbf{F}^{nj} - \left(\mathbf{i}'\mathbf{Z}_{\boldsymbol{\rho}}^{nj} + \mathbf{i}'\mathbf{F}_{\boldsymbol{\rho}}^{nj}\right) \tag{13}$$

where $\mathbf{i}$ is a summation vector composed of 1s of appropriate dimension—since it is transposed (′) and used in a pre-multiplication, it sums across rows. Outflows are calculated as a residual, and both inflows and outflows are hand-adjusted to assure that no negative inflows or outflows exist and that total supply equals total demand. State GDP can be calculated with the information provided in the SAMNJ22 using both the income and final-demand approaches; total savings must equal total investment, and foreign savings must match the difference between inflows/imports and outflows/exports plus the residence adjustment.

2.11 The full set of state benchmark-industry data

To summarize the above effort, we now list the data estimates that we create for each of 402 benchmark SUT industries for each state:

1. Number of establishments
2. Number of jobs
3. Total value added
4. Labor compensation
5. Gross operating surplus
6. Subsidies
7. Federal tax revenues
8. State tax revenues
9. Local tax revenues
10. Labor earnings
11. Output
12. Household consumption
13. International imports
14. International exports
15. Change in inventories
16. Investment
17. Intra-national inflows
18. Intra-national outflows

2.12 *Occupational allocation of jobs*

As noted in a prior section, we have job estimates by SUT industry. Many subnational applications—labor-market analysis, workforce development, and income distribution studies—require employment by occupation. We obtain estimates of occupational employment by mapping state job estimates to national occupation structure (BLS, 2024). First, we use a concordance that re-expresses the 402 SUT industries to the industry classification that BLS uses to report occupational staffing. Then we apply their industry-by-occupation matrix that distributes each industry's jobs to occupations.

BLS's Occupational Employment and Wage Statistics (U.S. Bureau of Labor Statistics, 2024) set of 183 reporting industries is coarser than and imperfectly aligned with BEA's 402 benchmark industries. We therefore match our benchmark jobs vector to those 183 industries[15] and then apply a normalized occupational-staffing pattern of each reporting industry. Formally, that is

$$\boldsymbol{\eta} = \mathbf{O}\ {}^{183}\boldsymbol{\Gamma}^{402}\ \boldsymbol{\varepsilon} \tag{14}$$

where $\boldsymbol{\eta}$ is the vector of jobs by occupation; $\boldsymbol{\varepsilon}$ is the 402-vector of state jobs by benchmark industry estimated in Section 3.11; the aggregator (left- and right-superscripted by its output and input dimensions) is the 183×402 zero–one matrix that sums the benchmark industries belonging to each Bureau of Labor Statistics reporting industry; and $\mathbf{O}$ is the occupation-by-industry staffing matrix, each column of which gives the share of a reporting industry's employment accounted for by an occupation, normalized to sum to one over the occupations at a given level of the hierarchy. The product $\mathbf{O\Gamma}$ is thus a single occupation-by-benchmark-industry operator that can be stored and reused; post-multiplying it by any state's job vector returns that state's occupational profile while preserving total employment.

The occupational classification is the hierarchical Standard Occupational Classification (SOC). We modify BLS's industry-by-occupation matrix, so it has three nested tiers: 22 *major* groups (the broadest summary categories, such as "Management occupations" or "Production occupations"), 92 *minor* groups nested within them, and 173 *broad* occupations nested within the minor groups. The three tiers are mutually consistent: the shares at any tier aggregate exactly to the tier above, so an analyst may work at whatever resolution the application and the reliability of the underlying counts warrant. We eliminate many of the 832 detailed six-digit occupations in the BLS matrix. Many of these eliminated detailed occupations engage so few workers nationwide—and fewer still in any one state—that carrying them adds resolution without really adding much information that is apt to be used. But they can be appended if the need arises. We retain the three coarser tiers because they are fully populated across all 183 reporting industries and, hence, are statistically meaningful at the state scale.

The U.S. Census Bureau's American Community Survey (ACS), a 1% sample self-reporting personal and household microdata set, also uses SOCs in its occupational profile. This lets researchers link the industrial and occupational structure of the SAM to self-reported socioeconomic characteristics of the local resident working population—that is characteristics such as household income, poverty status, sex, race and ethnicity, age, educational attainment,

[15] The problem with construction rears its ugly head again. Here, however, we aggregate the related sectors in BEA and BLS to a single Construction sector.

and the like. Hence, occupation enables as a bridge between production and household sides of the accounts. It therefore is an easy way to connect income-, demographic-, and location-based attributes of households as described in Section 3.9 and is a way to build distributionally and equity-oriented SAMs.

3. Data Records

The SAMNJ22 has been elaborated using the latest benchmark IO tables for the US (2017) and data from the US Bureau of Economic Analysis (BEA; U.S. Bureau of Economic Analysis, 2024), the *Internal Revenue Service Data Book, 2022* (U.S. Internal Revenue Service [IRS], 2023a), and *State Government Finances*: 2022 and 2021 for New Jersey (U.S. Census Bureau, 2022a). SAMNJ22 is a 71-by-71 balanced matrix based on purchasers' prices. It contains one representative household sector, a corporate sector, two sectors of government (federal and a combined state and local, which includes Non-Profit Institutions Serving Households (NPISH)), and the foreign sector. There are two accounts for transfers (unemployment benefits and dividends), six accounts for tax revenues (income taxes, Social Security Contributions (SSC) of employers, SSC of employees, taxes on production, taxes on sales, and taxes on imports), one account for subsidies on production, three investment accounts (federal government, state and local government, and private investment), changes in the industry stocks, savings, two productive factors (labor and capital), and 71 productive sectors or activities. The structure of the matrix is detailed in Figure 2.

3.1 Institutions, transfers, savings and foreign sectors

As mentioned earlier, there are four agents in this economy: households, corporations, federal government, and state and local government. In the following, we describe the main data sources that were used to estimate the income and expenditure figures for each agent.

3.1.1 The representative household

According to the BEA's accounts for New Jersey's gross domestic production, households' incomes accrue from primary factors, capital, and labor income (which includes the residence adjustment), transfers from other institutions (the corporate sector, the federal government and the state and local government) and other transfers (unemployment benefits and dividends), and redistributions of SSC of employers.

Labor income is calculated as the aggregation of *wages and salaries* registered in BEA plus the adjustment for residence ($77.2 billion). Personal current transfer receipts of households from the federal government are the sum of retirement and disability insurance benefits, medical benefits, Income maintenance benefits and veterans' benefits. State and local governments pay unemployment benefits (1787.353million). Households also receive transfers from the corporate sector, NPISH, and employer contributions for employee pension and insurance funds ($45.1 billion).

A share of household income is used to fulfill payments of personal income taxes. BEA and the *Internal Revenue Service Data Book*, 2022 (see Table 4) provide detailed information on federal government revenues.

Figure 2. Main Accounts of the SAMNJ22

	Institutions (4)	Transfers (2)	Taxes and Subsidies (6)	Gross Capital Formation (4)	Savings (1)	Primary Factors (2)	Productivity Sectors (x)	Foreign Sector (1)
Institutions	Transfers among institutions	Redistribution of unemployment benefits and dividends	Redistribution of tax revenues among the institutional sectors			Labor and capital income		
Transfers	Unemployment benefits and dividends payments							
Taxes and subsidies	Income taxes and subsidies payments						Contributions, taxes on production, sale taxes and tariffs	
Gross Capital Formation					Redistribution of savings			
Savings	Public and private savings							Foreign savings
Primary factors							Labor and capital income	Adjustments for residence
Productive sectors	Public and private consumption		Subsidies on production	Investment and Stocks variation			Intermediate matrix	Exports
Foreign sector	Transfers and/or savings to the Rest of World (ROW)						Imports	

Table 4. Personal income tax paid by households (SAMNJ22) in $ millions

Federal Government Revenues (IRS Databook, 2022)*	125,566.669
State Government Revenues (Census Gov, 2022)	20,630.297
Personal income paid by households	**146,196.966**

**Net of SSCH (BEA), $29,160.205 (January 2023)*

According to the U.S. Consumer Expenditure Survey (U.S. Bureau of Labor Statistics, 2022a), the ratio of average annual expenditures ($72,967) above income before taxes ($94,003) is 77.6%. Taking this number as a reference for New Jersey we can approximate household savings.

The last available data for New Jersey in the Consumer Expenditure Surveys (US Bureau of labor statistics) also is for 2022. Dividing the average annual household expenditure ($80,688.54) by the income before taxes ($111,034.09) less personal income taxes (it contains some imputed values), ($16,787.28) we obtain a share of 85.6 percent as a result of ($80,688.54/ $94,246.81). Household savings are assumed to be the residual. We model a single representative household here, though the accounts can be extended to distinguish household groups.[16]

3.1.2 The corporate sector

The income of this account accrues from capital services (gross operating surplus, GOS). A share of GOS is used to pay corporate taxes as well as some transfers to households, state and local governments,[17] and dividends. The GOS figure for New Jersey provided by BEA ($281,603 millions of dollars) was reduced in the amount paid to households[18] ($54.7 billion) in the BEA accounts. The amount paid by corporations to state and local governments ($19.526878 billion) is the sum of current charges ($10.976843 billion) for government services and miscellaneous general revenue ($8.550035 billion), which is detailed in the *State & Local Government Finances: 2022, New Jersey*. Income taxes paid by corporations is the sum of the corporate tax received by the state government ($5.959760 billion, *State & Local Government Finances: 2022: New Jersey*) and the federal government ($29.377431 billion, *Internal Revenue Service Data Book, Fiscal Year 2022*). As an intermediate agent, the corporate sector also pays dividends to households. Finally, corporate savings are assumed to be the residual ($ billion). The transfers paid by the Corporate sector to NPISH is obtained from BEA (SAINC35 Personal current transfer receipts) in 2022, and it is ($907.756 million)

3.1.3 Federal government

Federal government revenues are primarily generated via taxes: direct taxes (personal and corporate income taxes; SSC of employers, employees, and self-employees; taxes on production

[16] Many SAMs and extended IO models split the single household account into groups—by income, age of householder, or location—so that consumption columns and labour-income rows can differ across them; see the discussion in Section 3.9.

[17] Recall that this account includes NPISH.

[18] Calculated as an average in the table quarterly personal income by state (BEA), displayed in SQINC4 Personal income by major component. It is the value of proprietors' income

and taxes on imports. Direct tax revenues in New Jersey have been calculated using information provided by BEA and the *IRS Data Book, Fiscal Year 2022 (U.S. Internal Revenue Service, 2023a)*. They are equal to the difference between total revenues from direct taxes and direct taxes collected by the state and local government (individual income taxes [$16.8 billion] and corporate income taxes [$5.9 billion]) and reported in *State & Local Government Finances: 2022, New Jersey*. Data on SSC are from BEA. Taxes revenues from production and imports are estimated as total revenues less the revenues of taxes on products received by the state and local governments, according to the *State & Local Government Finances: 2022, New Jersey*—Property taxes ($32.8 billion), sales taxes in the form of general sales ($12.8 billion) and selective sales ($5.1 billion), motor vehicle license registration ($0.7 million), and other taxes ($2.7 billion). The federal government also receives all taxes on imports.

Federal tax revenues are used to pay for transfers to households and to State and local government: the estimates are detailed in Table 2. Federal consumption is the aggregate of federal nondefense and defense consumption provided by *USAspending.gov* (U.S. Department of the Treasury, 2022). Total direct expenditure is $81.1 billion that is disaggregated into total direct expenditure on defense ($10.4 billion) and nondefense direct expenditures ($70.4 billion). However, these figures include retirement and disability payments and other direct payments to individuals that are already included in the estimated figure of current transfers provided by BEA. To avoid a double accounting, we have subtracted these transfers, and the new values and the resulting value have been split in defense ($4.2 billion) and nondefense expenditures ($28.7 billion) using the percentages of previous figures over the initial total expenditure. Federal savings are derived as a residual.

Next, we disaggregate expenditures into investment and current consumption using the percentages of investment (capital outlay) for state and local government, *State & Local Government Finances: 2022, New Jersey,* ($4.1 billion) over the aggregation of capital outlay and current operation expenditures ($73.1 billion). The coefficients are 0.1 and 0.9, respectively.

3.1.4 State and Local Government and NPISH

State and local government institutions draw income from corporations' transfers to NPISH, net intergovernmental revenues, corporate and individual income taxes, as well as other taxes on production and capital revenues. The revenues are used to finance current transfers to households, unemployment insurance compensation, and subsidies on production (see Table 3). Total expenditures of state and local governments are estimated using data from *State Government Finances: 2022 State: New Jersey*. Estimates of the savings of state and local governments are identified by the residual.

4. Technical Validation

Our overarching validation principle is that the state accounts should reproduce federally published figures wherever a corresponding national or state control total exists. Throughout construction we benchmark the SAMNJ22 against official source data—BEA state value added and personal income, BLS QCEW employment and payroll, and the Census and IRS revenue series—and adjust the accounts to match those totals. The one systematic exception is the state-level international trade block: because direct subnational trade statistics are limited, international imports and exports are allocated to New Jersey using national shares (compensation shares for imports and output shares for exports) rather than matched to an independent state-level benchmark. Results in that block should therefore be read as consistent approximations rather than as validated against an external state figure.

The most basic check is internal accounting consistency. As a social accounting matrix, the SAMNJ22 must balance: total receipts equal total expenditures for every account, so each row sum equals its corresponding column sum. We enforce this with an iterative proportional fitting (RAS) procedure that reconciles the assembled state estimates to their row and column control totals while preserving the national technical structure, and we confirm that the final accounts balance to within numerical tolerance. The same procedure resolves the minor inconsistencies that arise because published source aggregates do not always reconcile across scales—for example, summed state QCEW totals need not equal the national total—by constraining the state accounts to sum to the national equivalent.

The regional purchase coefficients are fitted by beta regression on 1,723 industry-region observations across 51 regions and 36 commodities; the model reproduces the known coefficients with a squared correlation of $R^2 = 0.52$ and a pseudo-R^2 of 0.50, with a root-mean-square prediction error of 0.16 and a mean absolute error of 0.12 on the RPC scale and negligible mean bias (Table B2). The estimated coefficients carry signs consistent with economic intuition, and collinearity diagnostics confirm the specification is well identified: variance inflation factors are below conventional thresholds for all predictors except the deliberately retained supply and demand pair (Table B4). Fitted coefficients are bounded to the admissible (0, 1) interval, with values near the upper bound subject to the smooth compression described in Appendix B.

Finally, we apply a series of sanity checks to the assembled accounts: intermediate-use and final-demand entries are non-negative, value-added components reconcile to BEA state totals, and the tax rows aggregate to the federal, state, and local revenue figures reported by the IRS and the Census of Governments. Where these checks revealed discrepancies during construction, they were traced to source-data disclosure gaps or scale mismatches and resolved as described in the Methods; the published accounts pass all of them.

5. Usage Notes

We show an example of how to build a SAM for one U.S. state, New Jersey. It is a reusable foundation for subnational economic analysis—calibrating computable general equilibrium models, computing multipliers, and studying distributional questions at the state level—rather than as a finished answer to any one of them. Its accounts are fully balanced and internally consistent, but several elements remain deliberately simple: international trade enters through approximations rather than direct subnational trade data, regional purchase coefficients for services are set outside the freight-based regression, and households are represented by a single account. Each of these is a natural point of extension, whether by disaggregating households along income, age, or location, by drawing on Census trade series for a more direct treatment of trade, or by re-estimating coefficients as new data become available. We release the accounts and the code that builds them so that others can adapt the framework to their own regions and questions.

6. Data Availability

The SAMNJ22 dataset is available at TBD.

7. Code Availability

Code used to construct SAMNJ22 is available at TBD.

8. References

Álvarez-Martínez, M.T. & Polo, C. (2012). A general equilibrium assessment of external and domestic shocks in Spain. *Economic Modelling*, **29**, 2486–2493. https://doi.org/10.1016/j.econmod.2012.06.035

Álvarez-Martínez, M.T. & López-Cobo, M. (2018). WIOD SAMs adjusted with Eurostat data for the EU-27. *Economic Systems Research*, **30**, 521–544. https://doi.org/10.1080/09535314.2018.1448758

Bureau of Economic Analysis (BEA) (2024). *2017 Benchmark Input-Output Accounts of the United States*. U.S. Department of Commerce, Washington, DC. Available at: https://www.bea.gov/industry/benchmark-input-output-data

Cardenete, M.A., Boulanger, P., Delgado, M.C., Ferrari, E. & M'Barek, R. (2014). Agri-food and bio-based analysis in the Spanish economy using a key sector approach. *Review of Urban & Regional Development Studies*, **26**, 112–134. https://doi.org/10.1111/rurd.12022

Cazcarro, I., Amores, A.F., Arto, I. & Kratena, K. (2022). Linking multisectoral economic models and consumption surveys for the European Union. *Economic Systems Research*, **34**, 22–40. https://doi.org/10.1080/09535314.2020.1856044

Cribari-Neto, F. & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, **34**, 1–24. https://doi.org/10.18637/jss.v034.i02

de la Torre Cuevas, F., Lahr, M.L., Sargento, A.L.M. & Ferreira, J.P. (2026). Subnational MRIO building: the added value of household consumption and aggregate value-added data. *Economic Systems Research*, advance online publication. https://doi.org/10.1080/09535314.2025.2599092

European Commission, IMF, OECD, United Nations & World Bank (2009). *System of National Accounts 2008*. Series F, No. 2, Rev. 5, United Nations, New York. Available at: https://unstats.un.org/unsd/nationalaccount/sna2008.asp

European Commission, IMF, OECD, United Nations & World Bank (2025). *System of National Accounts 2025*. Adopted by the UN Statistical Commission, 56th Session, March 2025. Available at: https://unstats.un.org/unsd/nationalaccount/sna2025.asp

Felbermayr, G., & Gröschl, J. (2014). Within U.S. trade and the long shadow of the American secession. *Economic Inquiry*, **52**(1), 382–404. https://doi.org/10.1111/j.1465-7295.2012.00510.x

Foote, A. D., Graham, M. R. & Kutzbach, M. J. (2025). *LODES Design and Methodology Report: Methodology Version 7*. Center for Economic Studies Working Paper CES 25-52, U.S. Census Bureau. https://www.census.gov/library/working-papers/series/ces-wp.html

Gervais, A. & Jensen, J.B. (2019). The tradability of services: Geographic concentration and trade costs. *Journal of International Economics*, **118**, 331–350. https://doi.org/10.1016/j.jinteco.2019.03.003

Hewings, G.J.D., Yoon, S.G., Park, S., Kim, T.-J., Kim, K. & Kratena, K. (2017). Unraveling the household heterogeneity in regional economic models: Some important challenges. In Jackson, R. & Schaeffer, P. (eds.), *Regional Research Frontiers – Vol. 2: Methodological Advances, Regional Systems Modelling and Open Sciences*. Springer, Cham, pp. 23–47. https://doi.org/10.1007/978-3-319-50590-9_2

Hoover, E. M. (1948). *The Location of Economic Activity*. McGraw-Hill.

Hoover, E. M. & Giarratani, F. (1999). *An Introduction to Regional Economics*. Web Book of Regional Science, Regional Research Institute, West Virginia University (originally published 1984). https://researchrepository.wvu.edu/rri-web-book/4/

Jones, S. (2010). The economic contribution of tourism in Mozambique: Insights from a Social Accounting Matrix. *Development Southern Africa*, **27**, 679–696. https://doi.org/10.1080/0376835X.2010.522831

Kim, K., Kratena, K. & Hewings, G.J.D. (2015). The extended econometric input–output model with heterogeneous household demand system. *Economic Systems Research*, **27**, 257–285. https://doi.org/10.1080/09535314.2014.991778

Kim, K., Hewings, G.J.D. & Kratena, K. (2016). Household disaggregation and forecasting in a regional econometric input–output model. *Letters in Spatial and Resource Sciences*, **9**, 73–91. https://doi.org/10.1007/s12076-015-0140-8

Kratena, K. (2021). Employment, wage setting, and the multiplier. Working Paper 0022, CESAR (Centre of Economic Scenario Analysis and Research), Vienna. https://www.cesarecon.at/wp-content/uploads/2021/10/StateDependent_MultiplierWP.pdf

Kratena, K. (2024). Effective demand, wages and prices, and the multiplier. *Economic Systems Research*, **36**, 226–248. https://doi.org/10.1080/09535314.2023.2204393

Lahr, M.L. (2001). Reconciling domestication techniques, the notion of re-exports and some comments on regional accounting. *Economic Systems Research*, **13**, 165–179. https://doi.org/10.1080/09537320120052443

Lahr, M.L., Ferreira, J.P. & Többen, J.R. (2020). Intraregional trade shares for goods-producing industries: RPC estimates using EU data. *Papers in Regional Science*, **99**, 1583–1605. https://doi.org/10.1111/pirs.12541

Lahr, M.L. & de Mesnard, L. (2004). Biproportional techniques in input-output analysis: Table updating and structural analysis. *Economic Systems Research*, **16**, 115–134. https://doi.org/10.1080/0953531042000219259

Lahr, M.L., Yao, Y., Fei, D. & Lee, A. (2021). The Total Economic Impacts of Philadelphia's Beverage Tax. New Brunswick, NJ: Rutgers Economic Advisory Service (R/ECON) & The National Institute for Early Education Research. https://nieer.org/sites/default/files/2023-08/Beverage-Tax2020.11.06Final_edits4.14.21.pdf

Mainar-Causapé, A.J., Philippidis, G. & Sanjuán-López, A.I. (2021). Constructing an open access economy-wide database for bioeconomy impact assessment in the European Union member states. *Economic Systems Research*, **33**, 133–156. https://doi.org/10.1080/09535314.2020.1785848

Miller, R.E. & Blair, P.D. (2022). *Input-Output Analysis: Foundations and Extensions* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781108676212

Persky, J., Doussard, M. & Wiewel, W. (2009). Export orientation and the limits to local sovereignty. *Urban Studies*, **46**, 519–536. https://doi.org/10.1177/0042098008100993

Pyatt, G. & Round, J.I. (eds.) (1985). *Social Accounting Matrices: A Basis for Planning*. World Bank, Washington, DC. Available at: http://documents.worldbank.org/curated/en/919371468765880931/Social-accounting-matrices-a-basis-for-planning

Rodrigues, J.F.D. (2014). A Bayesian approach to the balancing of statistical economic data. *Entropy*, **16**, 1243–1271. https://doi.org/10.3390/e16031243

Rodrigues, J.F.D. & Lahr, M.L. (2018). The reconciliation of multiple conflicting estimates: Entropy-based and axiomatic approaches. *Entropy*, **20**, 815. https://doi.org/10.3390/e20110815

Round, J.I. (2003). Social Accounting Matrices and SAM-based multiplier analysis. Chapter 14 in F. Bourguignon & L.A. Pereira da Silva (eds.), *Techniques for Evaluating the Poverty Impact of Economic Policies*. World Bank and Oxford University Press, Washington, DC. Available at: https://www.un.org/en/development/desa/policy/mdg_workshops/eclac_training_mdgs/round_2003_sams_chapter14.pdf

Sargento, A.L.M., Lahr, M.L., Ferreira, J.P. & de la Torre Cuevas, F. (2024). Revisiting methods for estimating interregional input-output accounts: It's not just about trade flows. *Structural Change and Economic Dynamics*, **69**, 664–677. https://doi.org/10.1016/j.strueco.2024.04.004

St. Louis, L.V. (1989). Empirical tests of some semi-survey update procedures applied to rectangular input-output tables. *Journal of Regional Science*, **29**, 373–385. https://doi.org/10.1111/j.1467-9787.1989.tb01384.x

Stevens, B.H., Treyz, G.I., Ehrlich, D.J. & Bower, J.R. (1983). A new technique for the construction of non-survey regional input-output models. *International Regional Science Review*, **8**, 271–286. https://doi.org/10.1177/016001768300800306

Stevens, B.H., Treyz, G.I. & Lahr, M.L. (1989). On the comparative accuracy of RPC estimating techniques. In R.E. Miller, K.R. Polenske & A.Z. Rose (eds.), *Frontiers of Input-Output Analysis*. Oxford University Press, New York, pp. 245–257.

Stone, R. (1947). *Measurement of National Income and the Construction of Social Accounts*. League of Nations, Committee of Statistical Experts, Geneva.

Stone, R. (1956). *Quantity and Price Indexes in National Accounts*. Organisation for European Economic Co-operation, Paris.

Stone, R. (1961). *Input-Output and National Accounts*. Organisation for European Economic Co-operation, Paris.

Timmer, M. (ed.) (2012). *The World Input-Output Database (WIOD): Contents, Sources and Methods*. WIOD Working Paper No. 10. Available at: http://www.wiod.org/publications/papers/wiod10.pdf

Treyz, G.I., Friedlaender, A.F. & Stevens, B.H. (1980). The employment sector of a regional policy simulation model. *Review of Economics and Statistics*, **62**, 63–73. https://doi.org/10.2307/1924273

Treyz, G.I. & Stevens, B.H. (1985). The TFS regional modelling methodology. *Regional Studies*, **19**, 547–562. https://doi.org/10.1080/09595238500185531

U.S. Bureau of Economic Analysis (2023). *Regional Economic Accounts: Personal Income and Gross Domestic Product by State, 2022*. U.S. Department of Commerce, Washington, DC. Available at: https://www.bea.gov/data/by-place-county-metro-state

U.S. Bureau of Labor Statistics (2022a). *Consumer Expenditure Survey, 2022*. U.S. Department of Labor, Washington, DC. Available at: https://www.bls.gov/cex/

U.S. Bureau of Labor Statistics (2022b). *Quarterly Census of Employment and Wages (QCEW), 2022*. U.S. Department of Labor, Washington, DC. Available at: https://www.bls.gov/cew/

U.S. Bureau of Labor Statistics (2024). *Industry-occupation matrix data, by industry, 2024–34 National Employment Matrix*. Employment Projections program, U.S. Department of Labor, Washington, DC. Available at: https://www.bls.gov/emp/tables/industry-occupation-matrix-industry.htm

U.S. Bureau of Transportation Statistics & Federal Highway Administration (2022). *Freight Analysis Framework Version 5 (FAF5)*. U.S. Department of Transportation, Washington, DC. Available at: https://www.bts.gov/faf

U.S. Census Bureau (2020). *2020 Census Gazetteer Files: States*. U.S. Department of Commerce, Washington, DC. Available at: https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html

U.S. Census Bureau (2022a). *Annual Survey of State and Local Government Finances, Fiscal Year 2022: New Jersey*. U.S. Department of Commerce, Washington, DC. Available at: https://www.census.gov/programs-surveys/gov-finances.html

U.S. Census Bureau (2022b). *State Trade Data: U.S. Imports of Merchandise by State of Destination, 2022*. Foreign Trade Division, U.S. Department of Commerce, Washington, DC. Available at: https://www.census.gov/foreign-trade/statistics/state/

U.S. Census Bureau (2024). *2022 Census of Governments, State and Local Government Finance Tables*. U.S. Department of Commerce, Washington, DC. Available at: https://www.census.gov/programs-surveys/cog.html

U.S. Department of the Treasury (2022). *USAspending.gov: Federal Spending Data, Fiscal Year 2022*. Bureau of the Fiscal Service, Washington, DC. Available at: https://www.usaspending.gov/

U.S. Internal Revenue Service (2023a). *Data Book, Fiscal Year 2022*. Publication 55B, Washington, DC. Available at: https://www.irs.gov/pub/irs-prior/p55b--2023.pdf

U.S. Internal Revenue Service (2023b). *Statistics of Income—Individual Income Tax State Data, Tax Year 2022 (Historic Table 2)*. Statistics of Income Division, Washington, DC. Available at: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-state-data

United Nations (1993). *System of National Accounts 1993*. United Nations, New York. Available at: https://unstats.un.org/unsd/nationalaccount/sna1993.asp

Yang, X. (2024). Bulkiness of goods and the gravity of international trade: Differential impact of trade barriers. *Research in Economics*, **78**, 101013. https://doi.org/10.1016/j.rie.2024.101013

9. Author Contributions

TBD.

10. Competing Interests

The authors declare no competing interests.

11. Funding

TBD.

Appendix B. Estimating Regional Purchase Coefficients

This appendix gives the full specification, estimation, and post-processing of the regional purchase coefficients summarised in the main text: the beta-regression model and the rationale for each variable, the fitted results (Tables B2–B4), and the treatment of sectors outside the goods regression, including services and the labor row.

B.1 Fitting the model

Our data set relies heavily on the 2022 Freight Analysis Framework (FAF5; U.S. Bureau of Transportation Statistics and Federal Highway Administration, 2022), by commodity and state. The values of the dependent variable *rpc* are the ratio of intrastate shipments to total shipments originating from in-state establishments and thus lie strictly within the (0, 1) interval.

Two issues nonetheless attach when these flows are mapped to industry RPCs. First, FAF5 classifies shipments by the Standard Classification of Transported Goods (SCTG), and the SCTG categories are coarser than the benchmark IO sectors the RPCs are applied to (Table B1): one SCTG code can span several IO industries, which then share its commodity coefficient. Second, freight is recorded as it physically moves, so a shipment routed through an in-state wholesaler or distribution center is credited to that intermediary rather than to the producing or consuming establishment; this pulls measured flows toward distribution hubs and mixes genuine local self-supply with goods that merely pass through. Both work in the direction of measurement error, not bias in the estimator.

We estimate RPCs at the state level by beta regression, the natural choice for a response bounded on (0, 1); we set out the rationale for that functional form below, with the variable rationale.

The specification combines three groups of explanatory variables. The first group is eight continuous predictors, each an economic-geography proxy for regional self-sufficiency:

1. *emp_estab* — employment per establishment (average firm size) in the industry-region. Establishment and employment counts are from the 2022 QCEW.
2. *emp_estab_ratio* — that firm size relative to the national figure for the same industry, also from the QCEW.
3. *l_supply* and *l_demand* — natural logarithms of regional supply (output minus exports) and demand (supply plus imports). Supply (output) is taken from the regional output estimates described in the preceding section; demand is intermediate demand — the row sums of the intra-national interindustry matrix Znj — plus final demand.
4. *empLQ* — an employment location quotient: the region's employment share in the industry divided by the nation's share, the classic measure of regional industrial concentration, computed from QCEW employment.
5. *lodging_empLQ* — the location quotient of the lodging industry (NAICS 721000) in the same region, a proxy for tourism intensity, also from the QCEW.
6. *lodging_wageshare* — lodging's share of regional wages, a further tourism proxy, computed from QCEW wages.
7. *landshare* — the region's land area as a fraction of the national total, from the 2020 decennial census (U.S. Census Bureau, 2020); state land shares do not change appreciably over time.

The second group consists of binary variables that shift the regression intercept: four bundle related SCTG commodity categories (*animals*, *aggood*, *woodpaper*, and *metalequip*), seven flag individual high-volume categories (*sctg10*, *sctg17*, *sctg30*, *sctg31*, *sctg37*, *sctg38*, and *sctg39*), and two group states by census division (*esc* and *pac*); their membership and rationale are detailed in Sections B.2.5 and B.2.6 and in Table B1. The third group is a single continuous term, *us_s_wv_ratio*—the commodity's state-level weight-to-value ratio relative to the national equivalent, from the 2022 Freight Analysis Framework (FAF5)—whose role is discussed in Section B.2.3.

B.2 Rationale for the explanatory variables

Each variable proxies a determinate identified in the econometric RPC literature. The term "RPC" itself is due to Treyz, Friedlaender, and Stevens (1980). And Stevens et al. (1983) and Treyz and Stevens (1985) were the first to emphasize estimates of it via regression analysis. They articulate the RPC as substitution between intra- and extra-regional sources in response to relative delivered costs and identify it via an estimating equation rather via simply ratios. Stevens, Treyz, and Lahr (1989) followed by showing a regression approach is superior to such ratio approaches, which ignore crosshauling and, hence, are necessarily upwardly biased. We carry on this line of research in a subnational setting, encouraged by the findings of Sargento et al. (2024), who show that the general regionalization approach is mildly superior for constructing MRIOs. The following paragraphs describe the roles played by the various sets of variables.

11.1.1 *B.2.1 Local supply capacity relative to local absorption*

The pair *l_supply* and *l_demand* is our analogue of the supply/demand pool. The pool logic is mechanical: a region cannot meet locally more of a commodity than it produces, so the ratio of local supply to local demand bounds the RPC from above. Entering supply and demand as separate logged terms, rather than a single censored ratio, relaxes that ceiling into an estimated elasticity and lets the data set how fast added local production is absorbed locally and added local demand is met by imports. Substitution predicts opposite signs for the two terms—demand positive, supply negative. Holding production fixed, a larger local market is more likely to support a local supplier; holding demand fixed, a region whose output far exceeds its needs must ship the surplus out and so retains less of its own demand.

11.1.2 *B.2.2 Regional scale*

The size proxy *landshare* has been a standard regressor in RPC modeling since Stevens et al. (1983). A region's share of national land area proxies its physical and economic size: larger regions source more of their demand internally, since more of any commodity's production and consumption fall inside their borders. We prefer land area to output- or population-based measures because it is exogenous to the trade flows being explained, is free of the disclosure and revision problems that afflict economic aggregates, and is effectively time-invariant, so it adds little to no spurious year-to-year variation.

11.1.3 *B.2.3 Transport cost and shippability*

Because the RPC reflects substitution in response to relative delivered cost, the weight-to-value ratio informs how costly a commodity is to ship per dollar of value—a regressor used in this literature since Stevens et al. (1983), with the same value-density effect on shipping distance and mode documented as recently as Yang (2024). Its link to local sourcing, however, runs in two directions. For low-value commodities that are also produced almost everywhere—sand, gravel, and building stone—the high cost of shipping relative to value makes it non-economical to haul them far when a local source is at hand, so demand is met near the point of production and the RPC is high. But low value per unit of weight does not by itself keep a commodity at home: other bulk goods whose production is geographically concentrated, such as grain and coal, are shipped over long distances precisely because their low value lets them move by slow, inexpensive modes such as rail and water rather than by the faster but costlier road and air modes that high-value goods require. What matters for the RPC is therefore the combination of shipping cost and the local availability of supply. Shipping cost, in turn, is not governed by weight and value alone. Some commodities are costly to move for reasons unrelated to their bulk: perishables require refrigeration and fast transit, fragile or flammable goods need special handling and protective packaging, hazardous materials face routing and mode restrictions, and high-value or sensitive items demand secure carriage. These commodity-specific determinants of transfer cost have been long established (Hoover, 1948; Hoover and Giarratani, 1999). Each of these raises delivered cost and pushes demand toward nearby sources, and because such requirements are properties of the commodity itself, they are absorbed by the commodity fixed effects discussed below rather than by the weight-to-value ratio. Our version of the ratio enters in relative form: measured against the national average for the same commodity, a high value indicates that the commodity is comparatively heavy or low-valued at its point of origin in that state, raising the share of local demand that nearby production can meet cheaply. We use the relative rather than the absolute ratio because the commodity binary variables already absorb the cross-commodity differences in bulkiness—there is an indicator for nearly every SCTG category—so the level of each commodity's weight-to-value ratio is captured by those fixed effects.

11.1.4 *B.2.4 Industrial structure and concentration*

Four firm- and concentration-based terms—*emp_estab*, *emp_estab_ratio*, *empLQ*, and the lodging measures—take the size logic to the industry-region level, where the location quotient has long served as the workhorse self-sufficiency proxy. *empLQ* measures regional industrial concentration: a region whose employment share in an industry exceeds the national share is specialized in it, is more likely to host a local supplier, and so retains more of its own demand, which we expect to enter with a positive sign. *emp_estab* and *emp_estab_ratio* add what the location quotient misses, namely average establishment size relative to the national figure; scale economies and minimum-efficient-plant size bear on whether a local producer can serve local demand at competitive cost.

The two lodging terms are tourism proxies. They are necessary because tourism causes economies to rethink the origins of the demand. In a tourism-intensive region, visitors buy local output but—like residents—buy it on-site. Thus, the region retains and consumes a larger share of its own production than resident demand alone might suggest. Lodging intensity captures this visitor demand, which resident-only derived goods consumption alone would not.

11.1.5 *B.2.5 Commodity fixed effects*

The commodity binary variables—the bundled categories (*animals*, *aggood*, *woodpaper*, *metalequip*) and the stand-alone categories (*sctg10*, *sctg17*, *sctg30*, *sctg31*, *sctg37*, *sctg38*, *sctg39*)—are commodity-group fixed effects, commodity binary variables of a kind used throughout this line of RPC regressions—Stevens et al. (1983) through de la Torre Cuevas et al. (2026)—though the research teams differ in how many they include; Table B1 lists every SCTG code with its commodity description and the approximate NAICS sectors to which we assign it when mapping fitted RPCs to benchmark industries. They absorb commodity-specific determinants of tradability the continuous regressors miss—perishability, sanitary or regulatory barriers to interstate movement, product differentiation, and the structure of distribution channels. A separate intercept by group lets the model fit a different baseline RPC for each commodity type, net of the continuous predictors.

11.1.6 *B.2.6 Regional fixed effects*

The two census-division binary variables (*esc* and *pac*) flag regions whose transportation networks and position in the national space economy lead them to internalize an atypical share of their trade. The Pacific division includes not only the noncontiguous states of Alaska and Hawaii, whose separation from the national rail and highway network raises the cost of importing from other states, but also the contiguous western coastal states—California, Oregon, and Washington—whose populations and economic activity are concentrated on the seaward side of the Sierra Nevada and Cascade ranges. These mountains form a substantial barrier between the coastal production centers and the interior of the country, raising the cost of hauling goods overland to and from the rest of the nation, so these states likewise meet an atypically large share of their demand from local production, which leads us to expect a positive *pac* coefficient. The East South Central binary variable absorbs a common deviation in the other direction, reflecting the persistent cultural and institutional legacy of the North–South divide, which continues to dampen interstate trade across the former Union–Confederacy border to the present day (Felbermayr and Gröschl, 2014). As fixed effects, both keep division-level idiosyncrasies from confounding the economic-geography coefficients.

B.3 Functional form

The functional form follows from the dependent variable. Because the RPC is a proportion confined to the open unit interval and is heteroskedastic near its bounds, ordinary least squares can predict values outside [0, 1] and misstates precision. Stevens et al. (1983) fitted the original RPC equation as a linear-probability model and ran into exactly this problem—the most concentrated industries returned predictions above one. Treyz and Stevens (1985) fixed it by tightening the specification into a quasi-binomial form: a logit link on a binomial variance, fit by quasi-likelihood, which holds predictions inside the unit interval.

We opted instead for a beta regression, following Sargento et al. (2024) and de la Torre Cuevas et al. (2026).[19] The reason is the shape of the data: as Cribari-Neto and Zeileis (2010) note, rates and proportions are typically both heteroskedastic and skewed, and the beta model corrects for both at once. Empirical RPCs cluster near zero and one. The logit link keeps fitted values in (0, 1); the mean-dependent variance $\mu(1-\mu)/(1+\varphi)$ shrinks toward the bounds, matching the boundary heteroskedasticity; and the beta density itself shifts from right- to left-skewed as the fitted mean moves across the interval, so the asymmetry near each bound is built in rather than assumed away. A quasi-binomial model, by contrast, specifies only the mean and a single dispersion scalar on a binomial variance—it posits no distribution for the response and so cannot represent that skewness or estimate the extra dispersion the RPCs show. We follow Cribari-Neto

[19] We also ran a quasi-binomial regression, which generated rather similar, though not identical, results to those of our beta regression.

and Zeileis (2010) in the parameterization and hold the precision φ constant across observations, though their framework allows φ to depend on covariates as well.

Table B1. Standard Classification of Transported Goods (SCTG) Codes, Commodity Descriptions, and Approximate NAICS Correspondence

SCTG	Commodity description	Approx. NAICS
01	Live animals and live fish	112, 1141
02	Cereal grains	1111
03	Other agricultural products	111 (exc. 1111)
04	Animal feed, eggs, honey, and other products of animal origin	1112–1119, 311119
05	Meat, poultry, fish, seafood, and their preparations	3116, 3117
06	Milled grain products and preparations, and bakery products	3118
07	Other prepared foodstuffs, fats, and oils	3114, 3115
08	Alcoholic beverages and denatured alcohol	3121
09	Tobacco products	3122
10	Monumental or building stone *	2123 (in part)
11	Natural sands	2123 (in part)
12	Gravel and crushed stone	2123 (in part)
13	Nonmetallic minerals, n.e.c.	2123, 2129
14	Metallic ores and concentrates	2122
15	Coal	2121
16	Crude petroleum	2111 (in part)
17	Gasoline, aviation turbine fuel, and ethanol *	324110 (in part)
18	Fuel oils	324110 (in part)
19	Other coal and petroleum products	324 (other), 2111
20	Basic chemicals	3251
21	Pharmaceutical products	3254
22	Fertilizers	3253
23	Other chemical products and preparations	3252, 3255, 3256, 3259
24	Plastics and rubber	326
25	Logs and other wood in the rough	1133
26	Wood products	321
27	Pulp, newsprint, paper, and paperboard	3221
28	Paper or paperboard articles	3222
29	Printed products	323
30	Textiles, leather, and articles of textile or leather *	313–316
31	Nonmetallic mineral products *	327
32	Base metal in primary or semi-finished forms	331
33	Articles of base metal	332
34	Machinery	333
35	Electronic and other electrical equipment, and office equipment	334, 335
36	Motorized and other vehicles, including parts	3361–3363
37	Other transportation equipment *	3364–3369
38	Precision instruments and apparatus *	334, 339 (in part)
39	Furniture, mattresses, lamps, lighting, and illuminated signs *	337
40	Miscellaneous manufactured products	339
41	Waste and scrap	562, 4239 (in part)
42	Mixed freight	—
43	Mail and courier parcels	—

Note: Two-digit SCTG commodity codes carried in the 2022 Freight Analysis Framework, with their descriptions and the approximate NAICS sectors to which we assign them when mapping fitted RPCs to benchmark industries. The SCTG was designed for comparability with NAICS at the broad level only, so the correspondences are approximate; "in part" marks codes that map to a portion of the listed NAICS sector, and codes 42 and 43 are mixed shipments with no single industry of origin. An asterisk (*) marks the seven codes that enter the beta regression as their own fixed effect; the remaining goods codes enter through the bundled commodity groups described above.

One consequence matters for reading the table. Under the logit link each coefficient is a shift in the latent index, not a direct effect on the RPC, so the marginal effect on the RPC is smaller and depends on the fitted mean (see Table B2 and the accompanying discussion). We estimate the model by maximum likelihood, with a logit link on the mean and a constant precision parameter (φ); standard errors are the nonrobust observed-information estimates. Because the observations cluster by state, we re-estimated with standard errors clustered by state: several standard errors roughly double, but every coefficient remains statistically significantly different from zero at conventional levels with a two-tailed test.

B.4 Fitted results

The estimated coefficients are reported in Table B2; all are statistically significantly different from zero at conventional levels with a two-tailed test.

Table B2. Beta-Regression Estimates of FAF5-based State-level Regional Purchase Coefficients for 2022

Variable	Coef.	Std. err.	P>|z|	1-SD effect (logit)
const	-2.3376	0.201	0.000	
emp_estab	-0.0040	0.001	0.000	−0.1672
emp_estab_ratio	0.1775	0.040	0.000	0.1134
l_demand	0.8073	0.195	0.000	1.5800
l_supply	-0.7007	0.196	0.000	−1.3610
lodging_wageshare	74.0144	11.719	0.000	1.0066
lodging_empLQ	-0.8380	0.135	0.000	−1.0190
empLQ	0.1032	0.014	0.000	0.2674
us_s_wv_ratio	0.0462	0.003	0.000	0.3145
landshare	9.0566	1.100	0.000	0.2074
esc	-0.2546	0.068	0.000	—
pac	0.4846	0.070	0.000	—
animals	1.2748	0.118	0.000	—
aggood	0.7523	0.054	0.000	—
woodpaper	0.4855	0.067	0.000	—
metalequip	0.2004	0.083	0.015	—
sctg10	1.6002	0.122	0.000	—
sctg17	1.9914	0.137	0.000	—
sctg30	-0.5319	0.121	0.000	—
sctg31	0.8079	0.111	0.000	—
sctg37	-0.8098	0.130	0.000	—
sctg38	-0.7287	0.122	0.000	—
sctg39	0.8335	0.111	0.000	—
precision	1.8530	0.032	0.000	—

Note: N = 1,723 industry-region observations across 51 regions and 36 commodities (rows with incomplete predictors dropped); df model = 22, df residuals = 1,699; log-likelihood = 721.17; AIC = −1,394; BIC = −1,264. Goodness of fit: fitted RPCs reproduce the known training values with a squared correlation of $R^2 = 0.52$; the pseudo-R^2 is 0.50 (Ferrari and Cribari-Neto measure, 0.47); root-mean-square prediction error is 0.16 and mean absolute error 0.12 on the RPC scale, with negligible mean bias. The 1-SD effect is the change in the logit of the RPC produced by a one-standard-deviation increase in the regressor, computed as the coefficient times the regressor's standard deviation (Table B3). It places the regressors on a common scale; the binary indicators are omitted because a one-standard-deviation change is not meaningful for a 0–1 variable.

The signs of regression parameters line up with economic intuition. Land area as a share of the nation (*landshare*) and the lodging wage share (*lodging_wageshare*) carry large positive coefficients, consistent with larger and more tourism-oriented regions retaining more of their own demand. The supply and demand logs enter with opposite signs—*l_supply* negative and *l_demand* positive—capturing the balance between local production capacity and local absorption. The commodity binary variables shift the intercept sharply by product type, raising the fitted RPC for some categories, such as *sctg10* (monumental or building stone) and *sctg17* (gasoline, aviation turbine fuel, and ethanol), and lowering it for others, such as *sctg37* (other transportation equipment) and *sctg38* (precision instruments and apparatus); the underlying transport-cost and availability logic is set out in Section B.2.3. Signs of the Census-division binary variables, *pac* and *esc*, match hypothesized signs as well.

The final column of Table B2 places the continuous regressors on a common footing. Because they are measured on very different scales, their raw coefficients are not directly comparable; the column instead reports the standardized effect of a one-standard-deviation change in each regressor—the coefficient times the regressor's standard deviation—expressed on the logit scale. The supply and demand logs are the dominant influences, each moving the logit by roughly 1.4 to 1.6 over a one-standard-deviation change, followed by the two lodging measures at about one logit unit each; the remaining continuous regressors—*us_s_wv_ratio*, *empLQ*, *landshare*, and the two establishment-size terms—move it by only about 0.1 to 0.3. The ranking confirms that local supply and demand, and tourism intensity, are what chiefly determine self-sufficiency in the fitted model. The binary commodity and region indicators have no standardized effect of this kind and are marked —; average marginal effects on the RPC, if wanted, follow from the coefficients via the mean of $\mu(1-\mu)$, here 0.219, so that the effect on the RPC is about 0.22 times the reported coefficient.

Table B3. Descriptive Statistics of Regression Variables

Variable	Mean	Std. dev.	Min.	Max.
rpc	0.4384	0.2353	0.0001	1.0000
emp_estab	36.4594	41.7886	1.0000	395.6500
emp_estab_ratio	0.9411	0.6387	0.0219	8.6873
l_demand	14.0573	1.9572	5.2832	19.8172
l_supply	13.9032	1.9424	5.2040	19.4864
lodging_wageshare	0.0097	0.0136	0.0027	0.0880
lodging_empLQ	1.2943	1.2160	0.4791	8.1898
empLQ	1.2920	2.5915	0.0015	59.8849
us_s_wv_ratio	2.7862	6.8064	0.0178	35.5416
landshare	0.0200	0.0229	0.0003	0.1616
esc	0.0818	0.2742	0.0000	1.0000
pac	0.0963	0.2951	0.0000	1.0000
animals	0.0290	0.1679	0.0000	1.0000
aggood	0.2281	0.4197	0.0000	1.0000
woodpaper	0.1126	0.3162	0.0000	1.0000
metalequip	0.0592	0.2361	0.0000	1.0000
sctg10	0.0284	0.1663	0.0000	1.0000
sctg17	0.0238	0.1525	0.0000	1.0000
sctg30	0.0296	0.1695	0.0000	1.0000
sctg31	0.0296	0.1695	0.0000	1.0000
sctg37	0.0290	0.1679	0.0000	1.0000
sctg38	0.0296	0.1695	0.0000	1.0000
sctg39	0.0296	0.1695	0.0000	1.0000

Note: N = 1,723 industry-region observations, the beta-regression estimation sample. The dependent variable *rpc* and the continuous predictors are summarised together with the binary commodity- and region-group indicators, for which the mean is the share of observations in the group.

B.5 Dispersion, collinearity, and robustness

The final row of Table B2, *precision*, requires separate comment because it is not a regressor and should not be read as one. The beta regression we adopt, following Cribari-Neto and Zeileis (2010), reparameterizes the beta distribution into two quantities: a mean μ, the expected RPC for a given region-commodity observation, and a precision φ, which measures how tightly the realized RPC concentrates around that mean. The mean for region-commodity observation ι is linked to the regressors through the logistic link,

$$\mathrm{logit}(\mu_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} \tag{B.1}$$

The coefficients above—from *const* through *sctg39*—model this mean RPC, while φ governs the dispersion of the realized RPCs around the value the model fits for each observation. In this parameterization the conditional variance of the RPC for industry i, denoted ρ_i—the single-industry counterpart of the vector $\boldsymbol{\rho}$ in equation (10)—is

$$\mathrm{Var}(\rho_i) = \mu_i(1 - \mu_i) \,/\, (1 + \varphi) \tag{B.2}$$

In this expression, φ is the beta-family counterpart of the residual variance in an ordinary regression: a larger φ implies that realized RPCs cluster more tightly around their predicted values, a smaller φ that they scatter more widely. A single dispersion parameter is reported because we hold φ constant across observations rather than letting it vary with covariates, which is the standard default and matches the beta-binomial set-up of de la Torre Cuevas et al. (2026). Two points of interpretation follow. First, the variance expression is mechanically compressed toward zero as the fitted mean approaches either bound, which is exactly the boundary heteroskedasticity that motivates the beta specification over ordinary least squares; carrying φ explicitly is what lets the model accommodate it rather than ignore it. Second, the test statistic on the *precision* row carries no economic content. The reported *p*-value is for the hypothesis φ = 0, which is not a meaningful null because φ is positive by construction; the value simply reflects that the parameter is precisely estimated ($\hat{\varphi}$ = 1.853, standard error 0.032), and unlike the slope rows its significance should not be read as evidence for or against any substantive effect.

Two pairs of regressors are collinear by design, and Table B4 reports the resulting variance inflation factors: above 400 for the supply and demand logs and near 75 for the two lodging measures, against values below two for every other predictor. We retain both pairs deliberately. The supply and demand logs are almost perfectly correlated—a region that produces much of a commodity also absorbs much of it—and enter together precisely to capture the small net balance between them, which is what bears on self-sufficiency; the two lodging variables are alternative proxies for the same tourism intensity. Collinearity inflates the standard errors of the affected coefficients and blurs the split of explanatory power between them, but it does not bias the coefficients, distort the fitted values, or compromise the well-identified linear combinations—here the net supply–demand contribution—on which our prediction depends. Every coefficient nonetheless remains statistically significantly different from zero at conventional levels with a two-tailed test, including under state-clustered standard errors, and the goodness-of-fit and out-of-sample error are unaffected because prediction turns on the fitted values rather than the individual coefficients. We therefore report the inflation factors for transparency, not as evidence that the specification fails.

Table B4. Variance Inflation Factors

Variable	VIF
emp_estab	2.02
emp_estab_ratio	1.89
l_demand	442.98
l_supply	441.18
lodging_wageshare	72.39
lodging_empLQ	76.00
empLQ	1.33
us_s_wv_ratio	1.35
landshare	1.88
esc	1.03
pac	1.27
animals	1.17
aggood	1.55
woodpaper	1.36
metalequip	1.15
sctg10	1.19
sctg17	1.14
sctg30	1.11
sctg31	1.08
sctg37	1.14
sctg38	1.11
sctg39	1.08

Note: Variance inflation factors from an auxiliary linear regression of each predictor on the others (constant included).

B.6 Sectors outside the regression and the upper bound

The regression above covers tradable goods, but not all sectors are goods-producing. Services and the other non-goods sectors are not represented in the freight-shipment data on which the beta regression is trained—they have no commodity weight, no shipment mode, and no entry in the Freight Analysis Framework—so the fitted equation cannot generate an RPC for them. For these excluded sectors we follow the standard practice in this literature, which dates at least to Stevens et al. (1983) and is retained by Lahr, Ferreira, and Többen (2020), Sargento et al. (2024), de la Torre Cuevas et al. (2026): we set each service RPC equal to the region's supply/demand ratio for that sector, censored from above at 1:

$$\rho_i = \min(1, s_i / d_i) \tag{B.3}$$

We right-censor it at 1 because RPCs—ratios of local regional supplies to regional demands by industry cannot exceed 1. If they did, it would imply the impossibility that a region consumes more of an industry's production than it needs. Thus, a region produces for its own needs and exports any excess. The region therefore also imports from other nearby states any services for which local production falls short of local demand. This treatment is appropriate for most services precisely because they generally are at best only weakly tradable; they are consumed where they are produced and cross state lines far less readily than goods. Moreover, they lack characteristics (e.g., weight and transportability) that guide goods-oriented trade. Thus, the delivered-cost substitution mechanism that underlies the econometric model for goods is basically invalid in this setting. Where finer information exists, the supply/demand ratio can be scaled by a tradability factor below 1 to admit a modest degree of cross-state sourcing, as in de la Torre Cuevas et al. (2026); absent such information for U.S. states we apply the censored supply/demand ratio directly. The lodging industry (721000) is an exception. In this one case, we apply a quasi-arbitrary rule rather than the generic supply/demand rule. It is keyed on whether its location quotient *lodging_empLQ* (denoted LQ_i in the rule below) is at or below 1.42 (RPC set to 0.5 times *landshare* divided by 0.015) or above it (RPC set to *lodging_empLQ* divided by 1.42), reflecting that a region's ability to serve visitors depends on its size and concentration of lodging activity rather than on freight-style variables. Specifically:

$$\rho_i = 0.5 \cdot landshare_i / 0.015 \text{ if } LQ_i \leq 1.42; \quad \rho_i = (LQ_i / 1.42) \cdot \rho_i \text{ if } LQ_i > 1.42 \qquad \text{(B.4)}$$

RPCs initially above 0.95. An RPC is bounded above by 1 by definition, but the regression is an unconstrained predictor and can return values pressed against that ceiling. Rather than truncate such values to a hard 1.0 — which would erase the distinction between a very self-sufficient region and an essentially closed one — Equation (B.5) applies a smooth compression: for every RPC greater than 0.95 the transformed value is

$$\rho_i' = 1 - \exp(-3.154 \cdot \rho_i), \quad \rho_i > 0.95. \qquad \text{(B.5)}$$

This asymptotic function approaches but never reaches 1, so transformed values stay strictly below the ceiling while preserving their ordering; because the formula returns approximately 0.95 at an input of 0.95, it joins the untouched values smoothly at the threshold. Values at or below 0.95 are left exactly as produced.

The order of these steps matters, since each operates on the result of the last: the regression predicts first; excluded-commodity rows are overwritten with the supply-to-demand ratio; non-finite values are zeroed; lodging is overwritten with its quasi-arbitrary rule (equation B.4); and the upper bound finally compresses anything still above 0.95.

B.7 RPC for the labor row

Finally, the Households row receives its own self-sufficiency coefficient — not a tax, but computed in the same function and analogous to an industry RPC. It is the ratio of earnings by place of residence to earnings by place of work, taken from the Bureau of Economic Analysis table CAINC5N, and it captures commuting: how much of a region's labor income is earned by people who live there versus by people who work there. This ratio passes through the same upper-bound compression applied to industry coefficients.

A natural refinement would let this labor-income coefficient vary by industry, since commuting patterns differ systematically across sectors and wage levels: higher-paid workers tend to commute farther, and sectors differ in commute length—manufacturing workers, for instance, draw from wider labor sheds than many local-serving industries, and construction workers may travel to job sites that shift between projects. The BEA residence/workplace earnings ratio used here is economy-wide for each region. The Census Bureau's LEHD Origin–Destination Employment Statistics (LODES; Foote, Graham, and Kutzbach, 2025) offer a path to this refinement, resolving home-to-work flows by industry and earnings down to the census block—data already used to document the wage–distance gradient—though they resolve industry only to 20 NAICS sectors and currently omit a few states.