(2024) 54, 109–118

The Review of Regional Studies

The Official Journal of the Southern Regional Science Association

Tapestry: Collaborative Tool for Regional Data and Modeling*

Philip Watson[a] and Greg Alward[b]

[a]Department of Agricultural Economics and Rural Sociology, University of Idaho, USA

[b]College of Natural Resounces, University of Idaho, USA

Abstract: This paper introduces and describes the Tapestry regional economic data and modeling project. Tapestry is an open source set of data generating algorithms, harmonization bridges, data organizing procedures and data models. It is also a database that provides access to Tapestry also represents a collaborative environment where researchers can use and improve data generating algorithms and methods. Tapestry represents many types of regional economic data and models, but ultimately it is designed to generate inter-regional social accounting matrices for every county in the United States.

Keywords: regional economic data, collaboration, inter-regional social accounting matrices

JEL Codes: E16, C82, R15

1. INTRODUCTION

Understanding the composition and interconnections of the constituent parts of regional economies has been described as the missing piece of understanding national macroeconomic trends and conditions (Martin and Sunley, 1996; Krugman, 2009). The regional economic data needed to achieve this is vital for economists and policy makers. While the federal government of the United States (as is the case with many federal governments) publishes copious amounts of economic data across many topics, geographies, and time-frames, the task of organizing the data into usable datasets, reconciling the various datasets, dealing with suppressed data, and creating economic models with the datasets remains a laborious and costly task.

The overarching goal of the Tapestry project is to create open-source methods, bridges, programming code, and datasets that can be vetted, critiqued, and improved by the regional

*Special thanks to the following people who have assisted on this project: Ken Bunzel, Tanner Varrelman, Steve Watson, Mike Lowry and Michele Vachon. Funding for the Tapestry Project was received from the USDA-NIFA, the University of Idaho Agricultural Experiment Station, and the U.S. Department of Transportation. Philip Watson is a Professor of Applied Economics at the University of Idaho, Moscow, ID 83844. Greg Alward is a Research Scientist at the University of Idaho, Moscow, ID 83844. *Corresponding Author:* P. Watson, E-mail: pwatson@uidaho.edu

ISSN 1553-0892, 0048-49X (online)
www.srsa.org/rrs

science community, and then reincorporated into the project to create ever better datasets that are available to all. Ultimately, Tapestry will generate unsupporessed employment, wages, output, and value-added estimates for every county in the United States at the 6-digit NAICS specificity level, county-level estimates of household income by source, county level personal consumption expenditures, county level government income and spending, county level savings and investment, and intercounty trade commodity trade flows which include international trade ports. These data will all be organized into an interregional social accounting matrix of every county in the United States. This interregional social accounting matrix can then be used to create an individual social accounting matrix (SAM) for any county or group of counties. Furthermore these SAMs can be incorporated into input-output or computable general equilibrium models for policy analysis, scenario analysis, and economic impact analysis (Watson et al., 2012, 2015).

While not a perfect spatial unit of analysis, counties represent a middle compromise between very high spatial resolutions of zip codes (Watson and Beleiciks, 2009) and more coarse resolution of states (Caron et al., 2015). Additionally, there is an abundance of data that is reported at the county level across many government agencies. Counties can also be aggregated together to produce regions such as Metropolitan Statistical Areas (MSA) and Commuting Zones (CZ). Because of these reasons, developing county-level economic datasets is a primary concern for regional economists.

2. DATA SOURCES AND UNSUPRESSION

Tapestry employs data from the Bureau of Labor Statistics (BLS), Bureau of Economic Analysis (BEA), United States Department of Agriculture (USDA), Department of Transportation (DOT), and the United States Census Bureau (USCB). The following section will outline how these datasets are compiled, unsuppressed, and harmonized to generate the interregional social accounts that make up the core of Tapestry.

2.1. BLS Data

The first step in compiling Tapestry data is to download the Quarterly Census of Employment and Wages (QCEW) annual county-level employment and wage dataset for each year from the BLS website. Once these datasets are downloaded, they are subjected to an unsuppression algorithm that fills in the non-disclosed data in the dataset. The unsuppression algorithm is based on (Isserman and Westervelt, 2006) who developed an unsuppression algorithm for County Business Patterns (CBP) data from the USCB. We modified and altered their method to account for the differences between CBP and QCEW data. The algorithm is a recursive function that first generates an initial estimate for all suppressed values which then acts as a seed value for the recursive unsupression algorithm (equations 1 - 3).

Equation 1 first sums the disclosed employment totals by disclosed sector i in each county c. The subscript i indicates only the disclosed sectors while the subscript s will indicate all sectors. In the QCEW records, there is a "disclosure id" field for every record that is a 1 when the record is disclosed and a 0 when the sector exists in a county, but is suppressed. Using this information we can calculate the total of the disclosed employment in a county:

$$\text{sum\_disclos}_c = \sum_{i=1}^{n} (\text{annual\_avg\_emplvl}_{i,c}) \tag{1}$$

Equation 2 then estimates the difference between the sum of the disclosed employment and the total county employment. The total county employment is included in the downloaded QCEW datasets and it includes in the total both the sectors that were disclosed and the sectors that were suppressed. The employment total is also disclosed for all geographies so we therefore know the control total that the sectors must sum to at the county level:

$$\text{count\_suppres}_c = \text{COUNT}(\text{disclosure\_id} = 0 \text{ for each county}) \tag{2}$$

Equation 3 then calculates a simple initial estimate of the employment level (initial estimate emp) for each suppressed value in the QCEW dataset. This estimate becomes the initial seed value for the recursive unsupression algorithm:

$$\text{initial estimate emp} = \frac{\text{annual\_avg\_emplvl}_{\text{s,c}} - \text{sum\_disclos\_1}_{\text{s,c}}}{\text{count\_disclos\_0}_{\text{s,c}}} \tag{3}$$

The model then adjusts the initial estimate of employment levels by iteratively refining the estimates using aggregate data from higher geographical levels and recursively handling more detailed industry levels. The recursive algorithm continually refines the estimate of the suppressed employment values until the geographic and sectoral constraints are met.

The model then recursively calls itself with incremented industry length to handle more detailed industry levels. This recursion ensures that estimates at broader industry levels are refined using more granular data.

During each recursive call, the same process is repeated: aggregate data is collected, estimates are calculated and updated, and the model moves to more detailed levels.

The model iteratively and recursively refines the employment estimates using aggregate data from higher geographic and industry levels, progressively working down to more detailed levels. This multistep process ensures that the estimates are as accurate as possible given the available data. After the model converges, a final estimate of employment by 6-digit NAICS and for each county (annual_avg_emplvl) is stored in the database.

The recursive refinement process in the model ensures that estimates of employment levels are continually improved by leveraging data from higher levels of industry and geographical detail. By recursively refining estimates from broader to more specific levels, the model enhances accuracy and completeness, ultimately providing reliable estimates for undisclosed employment records.

A similar process is used to estimate the wages in the QCEW dataset except the the initial wage estimate (initial estimate wage) is calculated by multiplying the unsuppressed employment values calculated above by the average wage by sector at the national level. The intiial estimate for wages by sector and by county is as follows:

$$\text{initial estimate wage} = \left(\frac{\text{total\_annual\_wages\_US}}{\text{annual\_avg\_emplvl\_US}} \right) \times \text{annual\_avg\_emplvl} \tag{4}$$

In this way the initial wage estimate takes into account the level of employment in a given sector in a given county and the relative wage differences across sectors. This initial wage estimate is then recursively and iteratively updates using the same process as the employment estimates.

Ultimately, the Tapestry QCEW algorithms yield a fully unsupressed dataset of employment and wages at the county and 6-digit NIACS levels that conform to high level (state and 2-digit NIACS) QCEW reported totals. However, it is recognized that QCEW does not cover all employment, there is employment that QCEW does not cover. To account for this we use QCEW data as an allocator of BEA regional employment and wage totals.

2.2. BEA Data

BEA data is used to generate county level aggregated sectoral level control totals for employment, wages, and household income. The BEA uses what they term "line codes" that consist of aggregations of sectors. Therefore a bridge was built that mapped the NAICS codes used in the QCEW data to the line codes used in the BEA data.

The main tables that we use are CAINC5 (Personal income by major component and earnings by NAICS industry), CAINC6 (Compensation of employees by NAICS industry), and CAEMP25 (Total full-time and part-time employment by NAICS industry). Like the QCEW data discussed above, these BEA tables are subject to non-disclosure and suppression. We therefore employ a similar unsuppression algorithm as was discussed in section 2.1.

After being subjected to the unsuppression algorithm, an estimate of a fully disclosed CAINC5, CAINC6, and CAEMP25 table is generated for each county in the United States. These tables are much less detailed in their sectoral specificity than are the QCEW data, but they are more inclusive in terms of what jobs are included. Therefore, we uses the BEA totals by aggregated sector as the control totals for the more disaggregated QCEW sectors. Operationally, this means that we developed a table that bridges the line codes in the BEA data to the QCEW NAICS sectors for which they consist and then calculate the percent of NAICS employment in each county that is in each line code. We then multiply the respective BEA line code by the percent represented by each NAICS code to generate an employment and wage estimate that has QCEW 6 digit NAICS specificity but that maintains BEA line code control totals.

In the BEA table CAINC45, gross receipts are provided at the county level for both "livestock and products" and for "crops". The county level "livestock and products" value is used as the control total for county level output for the aggregate NAICS codes in 112 and the "crops" value is used as the control total for the aggregate NAICS codes 111. These aggregate values are then proportionally disaggragted into the more specific sectors using QCEW data for the livestock sector and USDA's crop-scape land cover data for the crop data.

The BEA is also the source for state-specific personal consumption expenditure (PCE) profiles that quantify the commodity-specific consumption of households. These PCE data are subsequently used to generate household consumption vectors in the social accounting matrices.

Lastly, the BEA is the source of the national make and use tables that are part of the national income and product accounts (NIPA). These two tables form the starting point for the production functions (Use Table) and byproducts table (Make Table) that will also be used in the generation of the social accounting matrices.

2.3. USDA Data

The primary USDA datasets that are used in Tapestry are the Cropland Data Layer (CDL) that has spatial land cover data which can identify what specific crops are being grown down to the field level. These data are then summed by county so we know what crops are being grown in each county. We also use national level data from the USDA National Agricultural Statistics Service (NASS) on production values measured in dollars per acre by crop. This gives us a way to weight the acres of the respective crops by the relative value per acre they generate. This gives us a relative distribution of the value of crop production by sector that we then benchmark to the BEA gross receipts for crops value from BEA table CAINC45.

2.4. Census Trade Data

The U.S. Census Bureau publishes international trade data by port and by commodity for both imports into the nation and for exports out of the nation. This data can be accessed through the Census Bureau's USA Trade Online's website (https://www.census.gov/foreign-trade/reference/products/catalog/usatradeonline.html). The Census Bureau collects import and export data for about 400 unique ports at the 6-digit "Harmonized System" (HS) sector level. Because this HS is slightly different from the NAICS sectoring used in other parts of Tapestry, the data needed to be bridged to reconcile the different sectoring schemes. Tapestry uses this trade data to model international commodity trade by assigning each port a pseudo FIPS code and then treating each port as if it were a county with its own commodity demand (exports) and commodity supply (imports). These port level trade data are then used in the gravity trade model described in section 4.

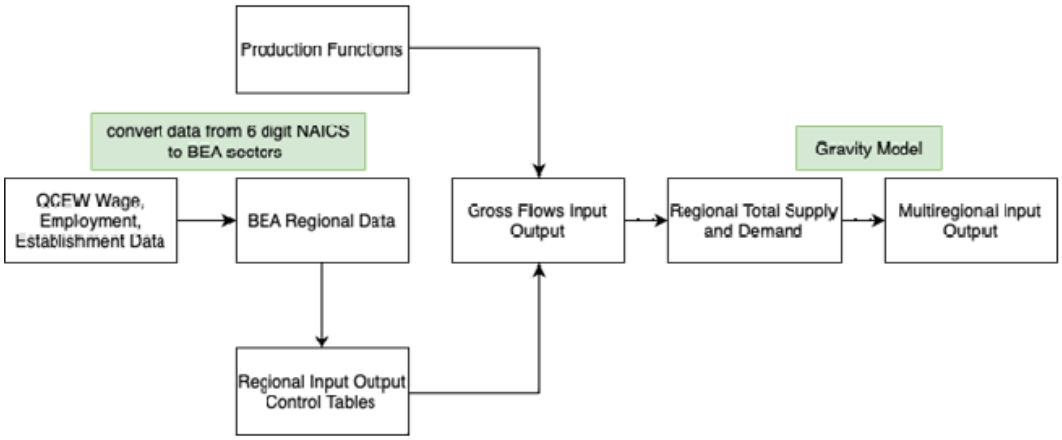
3. DATA COMPILATION AND MODELING

The data sources described in section 2 are then used to generate trade flow and, ultimately, an inter-regional social accounting matrix that describes the intra-regional and inter-regional transactions between the various sectors and institutions within and between the various counties in the United States. A simplified visual description of the data generating process is presented in Figure 1.

3.1. Tapestry Sectors

Version 1.0 of Tapestry employs 475 unique sectors in each of the 3,142 counties in the United States. These sectors are originally based on the BEA sectors which are in the national benchmark tables, but are expanded to include some additional sectors of interest. These additional sectors are dissagregated from the BEA sector for which they were originally

Figure 1: General Flow of the Data Compelation and Modeling Processes Employed by the Tapestry Project



included. We then generate new Tapestry benchmark make and use tables which dissagregate the national make and use tables into the 475 Tapestry sectors.

The bridges used to map from QCEW sectors to BEA annual data sectors, to BEA benchmark sectors, to Tapestry sectors are all available on the Tapestry website. As more collaborators develop additional production it is our goal to increase the number of sectors that are modeled within Tapestry.

3.2. Output, Value-Added, and Employment by Sector

The data sources described in section 2 are first used to generate the control total values of total industry and commodity output, value-added, and employment for each of the 475 Tapestry sector in each county.

The preliminary estimate of sectoral output for sectors other than the agricultural sectors is generated using national level employment to output ratios. These ratios are calculated by taking the output of each sector from the Tapestry benchmark make table and dividing it by the Tapestry employment in that benchmark year. These ratios, however, do not do a good job of estimating output for agricultural sectors, therefore we employ a different method for estimating those.

For the agricultural sectors, the top level output of "crop farming" and "animal farming" is taken from the BEA table CAINC45 which provides county level totals for both "livestock and products" and for "crops". These high level control totals for output are then disaggraged into the various agricultural subcategories using county level production data from the USDA.

3.3. Production Functions and Consumption Functions, and Commodity Supply and Demand

The initial production functions used in Tapestry come from the BEA's national use table and are scaled to the level of employee compensation, proprietor income, and total value-added

observed at the county level. In other words, the production functions for each sector begin with the column vector from the national use matrix and those values are then converted to technical coefficients by dividing even vector element by the vector sum. This vector of technical coefficients is then multiplied by the total industry output for each sector and for each county (see Section 3.2) to give the initial estimate of county specific intermediate input demand and value-added demand for each sector. These inital estimates are then scaled using county specific value-added data for each sector. The intermediate inputs are then scaled up or down proportionally to maintain the sector total.

This way of generating county specific production functions allows the model to account for regional differences in production regimes. For example, as of 2024, Idaho is the third largest producer of milk in the nation. The state also produces a great deal of cheese. However, Idaho has tended to produce large amounts of unbranded commodity bagged cheese that is sold to wholesale. While it still takes the same amount of milk to produce a pound of commodity cheese, it takes more milk to produce a dollar's worth of commodity cheese than a dollar's worth of branded cheese. Therefore, the production function in Idaho should show more intermediate milk purchased per dollar of output than in, say, Wisconsin, and the value-added to output ratio for cheese in Idaho should be lower than in Wisconsin.

4. TRADE MODEL

The total commodity demand and total commodity supply estimates discussed in section 3 do not have a geographic source associated. In other words, we have an estimate of how much of each commodity is produced in each county and how much of each commodity is demanded in each county, but we do not have an estimate for where the commodities demanded are sourced or where the commodities produced are sold. Traditionally these estimates were made using supply/demand pooling, location quotients, econometrically estimated regional purchase coefficients, or a formal trade model (Lazarus et al., 2002). Tapestry uses a formal trade model to allocate county and international port level commodity supply to county level and international port level demand. Specifically, Tapestry uses a gravity trade model to allocate and balance commodity trade across counties and international trade ports (Anderson and Van Wincoop, 2003).

The gravity model is designed to calculate the trade between different regions for a single commodity. The necessary inputs for this calculation include the commodity's supply and demand in each region, the distances between the regions, and an impedance factor (which weights the distances). The impedance factor used in Tapestry is based on Oak Ridge National Laboratories estimates of commodity specific transportation costs between counties across transportation modes (ORNL, 2008).

The Port level commodity import and export data (see Section 2.4) that is collected from the U.S. Census Bureau and each port is assigned a geographic location equal to the county FIPS code where the port is physically located. However, while the distances between the ports and the counties which are used in the gravity model are equal to the county where the port is located, we model the ports as having a near infinite distance between the other ports so that they do not trade between themselves directly. The ports are then incorporated into the gravity model with their own specific commodity supplies (imports) and commodity

demands (exports).

The form of the gravity model used here is based on models which were originally suggested by Wilson (1967) and which have been improved upon for decades (Yotov, 2022). The equations of the gravity model used in Tapestry are as follows:

$$\mathbf{S}_D = \text{diag}(\text{sup}) \tag{5}$$

$$\mathbf{D}_D = \text{diag}(\text{dem}) \tag{6}$$

where S_D is the matrix of the supply of a given commodity in each county along the diagonal and D_D is the demand for a given commodity in each county along the diagonal.

The transportation cost matrix (cost_mat) is given as follows:

$$\text{cost\_mat} = \alpha \cdot (\text{imped})^{\beta} \cdot \exp(\gamma \cdot \text{dist}) \tag{7}$$

where the impedance matrix is the multimodal average impedance matrix obtained from the Oak Ridge National Laboratory, and α, β, and γ are all distance decay parameters.

The attraction equation for the origin is given as:

$$\text{att\_org} = \mathbf{S}_D \cdot \text{cost\_mat} \tag{8}$$

The attraction equation for the destination is:

$$\text{att\_des} = \text{cost\_mat} \cdot \mathbf{D}_D \tag{9}$$

The allocation of commodities between where there are produced and where they are sold is then calculated using an iterative algorithm described in the following equations:

$$\mathbf{MS}_{\text{new}} = \text{att\_org} \cdot \mathbf{A} \tag{10}$$

$$\mathbf{MD}_{\text{new}} = \text{att\_des} \cdot \mathbf{B} \tag{11}$$

$$\mathbf{A}_{\text{new}} = \frac{1}{\sum(\mathbf{MD}, \text{axis} = 1)}, \quad \text{where } \sum(\mathbf{MD}, \text{axis} = 1) \neq 0 \tag{12}$$

$$\mathbf{B}_{\text{new}} = \frac{1}{\sum(\mathbf{MS}, \text{axis} = 0)}, \quad \text{where } \sum(\mathbf{MS}, \text{axis} = 0) \neq 0 \tag{13}$$

where the initial A matrix is defined as an array of 1s and the inital B matrix is defined as 1/(att_org).

The probability matrix is calculated as follows:

$$\text{int\_prob} = \mathbf{A}_{\text{diag}} \cdot \text{att\_des} \tag{14}$$

$$\text{prob} = \text{int\_prob} \cdot \mathbf{B}_{\text{diag}} \tag{15}$$

The final shipping matrix is then calculated as follows:

$$S = \mathbf{S}_d \cdot \text{prob} \tag{16}$$

The final shipping matrix, S, then represents the gravity modeled allocation of the total commodity supply and total commodity demand between domestic users and demanders from every other county in the nation. Likewise, international trade is allocated to specific county production and demand due to international trade ports being incorporated as geograpies within the gravity model.

5. CONCLUSION

Once we have the estimate of commodity trade flows from the gravity trade model (Section 4) that conforms to the control total of total commodity supply and total commodity demand estimated for each commodity and for each county, we can then generate an industry by commodity (IxC) interregional social accounting matrix (SAM) for any set of counties in the United States. These SAMs will be freely available on the tapestry website (www.uidaho.edu/tapestry) along with links to the code repository on GitHub. Additionally, the unsuppressed estimates of QCEW employment, wages, and establishment counts are also available on the website along with commodity trade flow estimates between counties and between counties and ports.

While we are working diligently to make what we think are the best possible assumptions and employ the best available evidence based science to build these datasets, we fully recognize that other people may have improved methods or different opinions on what the best available science is in building these datasets. We welcome and encourage people to engage with the data and methods and we are fully open to vetting and incorporating suggested improvements from users while giving them credit for any parts of the model that they improve. Ultimately, it is our hope that Tapestry facilitates a collaboration of regional economists looking to improve and streamline data and methods as much as it is simply a dataset or collection of models.

For example, one obvious area for improvement is the handling of government institutions. In the initial version of Tapestry, government (and other fourth quadrant) accounts are not very sophisticated and are basically just spatial disaggregations of the National Income and Product Accounts. This could certainly be improved and some additional datasets and analysis could be incorporated that would not only improve the SAMs, but also provide a useful standalone dataset for people working in local public finance.

We hope that this introduction to the Tapestry data tool will encourage people to explore, use, improve, and collaborate on regional economic datasets and social accounts. Again, links to the data and source code are available on the Tapestry website: www.uidaho.edu/tapestry.

REFERENCES

Anderson, James E and Eric Van Wincoop. (2003) "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93(1), 170–192.

Caron, Justin, Sebastian Rausch, and Niven Winchester. (2015) "Leakage from Sub-National Climate Policy: The Case of California's Cap-and-Trade Program," *The Energy Journal*, 36(2), 167–190.

Isserman, Andrew M and James Westervelt. (2006) "1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data," *International Regional Science Review*, 29(3), 311–335.

Krugman, Paul. (2009) "The Increasing Returns Revolution in Trade and Geography," *American Economic Review*, 99(3), 561–571.

Lazarus, William F, Diego E Platas, and George W Morse. (2002) "IMPLAN's Weakest Link: Production Functions or Regional Purchase Coefficients?," *Journal of Regional Analysis and Policy*, 32(1).

Martin, Ron and Peter Sunley. (1996) "Paul Krugman's Geographical Economics and Its Implications for Regional Development Theory: A Critical Assessment," *Economic Geography*, 72(3), 259–292.

ORNL. (2008) *Transportation Energy Data Book*. Oak Ridge National Laboratory, Department of Energy, Office of Vehicle and Engine Research.

Watson, Philip and Nick Beleiciks. (2009) "Small Community Level Social Accounting Matrices and Their Application to Determining Marine Resource Dependency," *Marine Resource Economics*, 24(3), 253–270.

Watson, Phillip, Stephen Cooke, David Kay, and Greg Alward. (2015) "A Method for Improving Economic Contribution Studies for Regional Analysis," *Journal of Regional Analysis & Policy*, 45(1), 1–15.

Watson, Philip S, Kimberly Castelin, Priscilla Salant, and J. D. Wulfhorst. (2012) "Estimating the Impacts of a Reduction in the Foreign-Born Labor Supply on a State Economy: A Nested CGE Analysis of the Idaho Economy," *The Review of Regional Studies*, 42(1), 51.

Wilson, A. G.. (1967) "A Statistical Theory of Spatial Distribution Models," *Transportation Research*, 1(3), 253–269.

Yotov, Yoto V. (2022) "On the Role of Domestic Trade Flows for Estimating the Gravity Model of Trade," *Contemporary Economic Policy*, 40(3), 526–540.