ESTIMATING COUNTY TO COUNTY TRADE FLOW

FINAL PROJECT REPORT

by

Philip Watson, Michael Lowry, and Fauwial Khan
University of Idaho

Sponsorship
University of Idaho
Pacific Northwest Transportation Consortium (PacTrans)

for

Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology (OST-R)

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
	01764464	

4. Title and Subtitle	5. Report Date
Estimating County to County Trade Flow	June 2023
	6. Performing Organization Code

7. Author(s) and Affiliations	8. Performing Organization Report No.
Philip Watson, University of Idaho Mike Lowry, 0000-0001-8485-4799; University of Idaho Fauwial Khan, University of Idaho	2020-S-UI-1

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
PacTrans Pacific Northwest Transportation Consortium University Transportation Center for Federal Region 10 University of Washington More Hall 112 Seattle, WA 98195-2700	11. Contract or Grant No. 69A3551747110

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590	Final Report 14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded to: www.pactrans.org

16. Abstract

Estimating count to county commodity trade flows is important for understanding a multitude of transportation, regional planning, and economic problems. While no primary data track intranational trade flows, gravity models have often been used to estimate these flows. However, to properly specify a gravity model, data on total commodity supply and total commodity demand for a consistent set of commodities and for every county must be obtained. Because these data are not available through primary data sources, a systematic process for estimating these values must be generated.

This report details a process for starting with primary government data, specifically the quarterly census of employment and wages (QCEW) and national input-output accounts from the US Bureau of Economic Analysis (BEA), and generating employment, output, and commodity supplies and demands across 409 commodities in 3,142 counties. These commodity supplies and demands are then used in a gravity model to estimate bilateral trade for each commodity between every county in the United States. Employment, commodity trade data, and county to county trade totals are available at: https://tapestry.nkn.uidaho.edu/

17. Key Words	18. Distribution Statement
Commodity flow, Economic models, Input output models, Counties	

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	20	N/A

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This research project developed a proof-of-concept trade model to estimate foreign and domestic county-to-county and county-to-port commodity shipments. A gravity model was developed using Python code. The code for the gravity model is presented in the appendix of this document. The model runs on a standard desktop in 22 seconds per commodity per year. It runs on a University of Idaho network computer in 3.3 seconds per commodity per year. The model can successfully distribute commodities across the 3,142 counties of the United States.

The underlying goal of this project was to create a method to allow researchers to share data and results. Researchers can work together to develop more sophisticated and accurate methods for overcoming data suppression. By working collaboratively and sharing our knowledge and tools, we can move beyond the days of relying on crude estimates and instead develop more effective strategies for analyzing and utilizing suppressed data. This work will enable a better understanding of regional economic resilience in relation to supply chains and the movement of goods.

CHAPTER 1. INTRODUCTION

1.1. Problem Statement

Knowing what commodities are supplied and demanded within a region does not necessarily correspond to knowing where the commodities consumed within a region are produced. This difference relates primarily to interregional trade that occurs outside the region at the regional level but inside at the higher level of regional aggregation. This outside-inside the region final demand distinction makes aggregating regional results at the national level problematic. A solution to this problem lies in the derivation of a multi-regional input-output model (Round, 1978).

Gravity models have been widely applied in various fields, including international trade, regional economics, and transportation planning. In international trade analysis, gravity models are used to estimate trade patterns, evaluate the impacts of trade agreements, and forecast future trade flows. Additionally, the gravity model can be extended to incorporate variables such as gross domestic product (GDP), population, and trade costs to enhance its predictive power.

To estimate these trade flows, a systematic process is needed for estimating total commodity supply and total commodity demand for a consistent set of commodities for every county in the United States. Once these have been estimated, then a gravity model can be developed to spatially allocate these commodity supplies and demands both within each county and between each county.

1.2. Research Objectives

The desire for data has been growing at a rate that can be matched only by the advancements in computing power, more so the latter's ease of access and affordability. Urban planners and regional economists are utilizing new data sources, such as mobile phone and Global Positioning System (GPS) data, to analyze the location patterns and movements of individuals and firms, providing valuable insights into the spatial distribution of economic activity (Beige and Axhausen, 2017; Duranton and Overman, 2005). However, the "withholding" of undisclosed data has been detrimental to researchers in formulating analyses. This data suppression occurs more commonly because of the U.S. Bureau of Labor Statistics' (BLS) confidentiality terms and fair competition between establishments (Wise, 2022). Despite the U.S. being one of the world's largest and most developed economies, this data suppression causes gaps in understanding economic activity across all of the North American Industry

Classification System (NAICS) levels. Hertz and Zahniser (2013) noticed an example of this difference between reporting and "reality" when they found that in the Quarterly Census of Employment and Wages (QCEW) dataset, 47 percent of agricultural employment, which included 36 percent of farmworkers, was concealed because of suppression at the three-digit NAICS level.

This lack of holistic data on local economic activity is a challenge for policymakers and economists alike. Policy analysts and regional economists cast a wide data net to correctly estimate study location patterns and growth with new computing capabilities, trying to provide valuable insights into the spatial distribution of economic activity and the dynamics of regional growth (Peterson and Jessup, 2008). In addition, the use of advanced computational methods, such as the RAS algorithm and network evaluation, has enabled more sophisticated estimation and detailed spatial examination, providing new understanding of the relationships among economic-geographic activities (Trinh and Phong, 2013). The eventual need for more viable undisclosed economic employment data led to this undertaking: the development of an algorithmic data interface capable of estimating and assigning disclosed values of the QCEW dataset.

The goal of this research was to develop a trade model to estimate foreign and domestic county-to-county and county-to-port commodity shipments. It is important to understand the mobility of goods across counties and to ports to better understand how shocks to supply chains will disrupt regional economies. The results of this research will also enable a better understanding of regional economic resilience in relation to supply chains and the movement of goods.

1.3. Report Organization

Chapter 2 provides background information about the data generating process and about gravity models. Chapter 3 describes how the data were generated from this research. Chapter 4 provides the code that was developed to estimate county to county trade flows and provides a link to the data repository where the data can be accessed. Chapter 5 provides conclusions and suggestions for future work.

CHAPTER 2. BACKGROUND

2.1. Gravity Models

While gravity models have long been applied to international trade and have been shown to be good predictors of commodity trade flows (Anderson and Van Wincoop, 2003), less research has been done on applying the gravity model to domestic, interregional trade (Cai, 2023). Notable studies that have used the gravity model for domestic interregional trade include applications to Japan (Gabela, 2020) and the European Union (Alama-Sabater et al. 2015). One notable application of the gravity model to inter-county trade in the United States is the proprietary gravity trade model employed by IMPLAN (www.implan.com) in its commercial input-output models (Lindall, Olson, and Alward, 2006).

2.2. Gravity Models in International Trade

Gravity models were first applied to international trade. The intricate web of international trade has long captivated the attention of economists, policymakers, and scholars. Seeking to understand the factors that shape trade patterns, economists turned to the gravity model—an econometric framework inspired by Newton's law of gravity. In 1967, A.G. Wilson unveiled his groundbreaking paper, "A Statistical Theory of Spatial Distribution Models," published in *Transportation Research*. This report delves into Wilson's pivotal work, tracing the evolution of gravity models and exploring their applications, limitations, and ongoing refinements.

In the world of economic theory, Wilson's work marked a turning point. Drawing on the fundamental principle that trade flows are proportional to the economic sizes of countries and inversely proportional to the distance between them, Wilson postulated the first comprehensive framework for analyzing international trade patterns—the gravity model.

Wilson employed a mathematical equation borrowed from the discipline of physics that revolutionized the field. The model's core assumptions were simple yet powerful: bilateral trade between two countries hinges on their economic sizes and the distance separating them. By incorporating additional factors such as transportation costs, cultural ties, and trade barriers, the model sought to capture the intricate dynamics of international commerce.

Through his mathematical formulation, Wilson crystallized the essence of the gravity model. He developed an equation that unveiled hidden trade flows and facilitated insights into the underlying mechanisms driving economic interactions. The model's expression, Tij = Ai * Bj / Dij, depicts trade flows between countries i and j. Here, the economic sizes of the respective

countries, Ai and Bj, exert their gravitational pull, while the distance between them, Dij, acts as a force of resistance.

The application of gravity models extended far beyond the realm of economic theory. Policymakers embraced this tool as a means to estimate trade patterns, predict the consequences of trade agreements, and anticipate future trade flows. The simplicity and elegance of the model made it accessible to researchers and policymakers alike, enabling a deeper understanding of the dynamics of global trade.

However, the model was not without its limitations. Critics pointed out that the assumption of a constant elasticity of trade with respect to distance overlooked non-linear effects and neglected crucial determinants of trade, such as institutional quality and cultural similarities. Skeptics argued that gravity models painted an oversimplified picture of complex trade relationships, failing to capture the intricate interplay of evolving global dynamics.

Yet despite the criticisms, the gravity model has withstood the test of time and has remained a resilient tool in the economist's arsenal. Researchers and economists have recognized the model's potential and have embarked on a journey of refining and expanding its boundaries.

Over the years, developments and refinements of the gravity model have illuminated new pathways for understanding trade patterns. Economists have expanded the model to incorporate additional variables, recognizing the importance of cultural and linguistic factors, infrastructure quality, and trade policy variables. Advanced econometric techniques, such as panel data analysis, have been employed to tackle data limitations and address endogeneity issues, further strengthening the model's predictive power.

In an era of unprecedented computational power and vast trade databases, the gravity model has undergone a renaissance. Enhanced estimation techniques and the availability of extensive data sources have breathed new life into this decades-old framework. It continues to evolve, be adapted, and uncover insights into the global trade landscape.

For example, a consolidated region with three inter-related sub-regions can be modeled by separating the export-import-transfer flows among them from their respective totals. In this way, each sub-region becomes an exogenous institution to the other sub-regions, with a column vector of exports and transfers, and a row vector of imports and transfers to the other. At the same time these inter-sub-regional trade and transfers are endogenous to the consolidated region. We would expect three effects from a consolidated bi-regional model: those contained within

each sub-region, those that feedback in a loop among regions, and those that spill over (without feedback) from one to the other.

Defourny and Thorbecke (1984) asserted that embedded within each element of the inter-industry transactions matrix there exists an implicit set of supply chains connecting the flow of products and by-products from a sector of origin to a sector of destination. The challenge was to make these implicit supply chains explicit. To do this, they used a technique called structural path analysis and thereby proved their assertion.

Koks et al. (2015) used a hybrid (linear and non-linear), multi-regional input-output model of Europe to estimate the continent-wide impacts of simulated floods in the Netherlands. Their results showed that most regions were unaffected by the flood disaster. Those outside regions that were affected included those that benefited from increased demand for output for substitutes and construction, while those that suffered losses were those that were subjected to supply chain disruption. The net effect of these gains and losses depended on the size of the initial disaster. The indirect effects were positive for small disasters and negative for large ones.

CHAPTER 3. WORK COMPLETED

3.1. Unsuppressing Employment and Wage Data

The QCEW program publishes a quarterly count of employment and wages reported by employers covering more than 95 percent of U.S. jobs, available at the county, metropolitan statistical area (MSA), state, and national levels by industry. (Quarterly Census of Employment and Wages, n.d.). The Bureau of Labor Statistics (BLS) data on county employment and wages are derived from the monthly/quarterly administrative records of state unemployment insurance (UI) offices and, for the case of federal employees, from the Unemployment Compensation for Federal Employees (UCFE) program. The data have been readily available since 1989 on www.bls.gov/cew, and since that time states have also begun providing data to the QCEW at the business establishment level.

The QCEW dataset has a number of suitable qualities. The data provide more detailed industry classification because of their availability at the six-digit NAICS level. It features coverage of all employers, i.e., employers with paid employees, employers with unpaid employees (small businesses), and even self-employed workers. This allows the QCEW to provide a better picture of the labor market in a given time period. Finally, it is consistent and reliable over time across varying geographic areas in the U.S. because the data are recorded quarterly.

As mentioned previously, the North American Industry Classification System (NAICS) is used to classify establishments and industries in North America. It is organized into hierarchical levels, with the highest level consisting of 21 broad sectors, such as agriculture, mining, manufacturing, and services (North American Industry Classification System (NAICS) U.S. Census Bureau, 2022). These sectors are then divided into subsectors and further divided into industries, which are identified by a six-digit code. The first two digits of the code represent the sector, the third digit represents the subsector, and the fourth through sixth digits represent the industry. Take, for example, NAICS 311351 – Chocolate and Confectionery Manufacturing from Cacao Beans. Each six-digit industry is part of a five-digit industry and iterates back to a two-digit industry. So NAICS 311351 – Chocolate and Confectionery Manufacturing from Cacao Beans is part of five-digit 31135 – Chocolate and Confectionery Manufacturing, which is part of four-digit 3113 – Sugar and Confectionery Product Manufacturing, which is part of three-digit 311 – Food Manufacturing, which is part of two-digit 31 – Manufacturing. In addition,

employment in an industry summed across all counties in a state must equal the state's employment in that industry; employment summed across states and the District of Columbia must equal national employment (Isserman and Westervelt, 2006).

Within the QCEW employment tables, some categories are marked with "ND" or an asterisk (*) to indicate suppressed information. The withholding of complete employment information is due to the BLS' disclosure rules, which are in place to protect the confidentiality of specific employers (Why Certain Employment Data Are Suppressed, 2022). According to the Department of Labor and Workforce Development in Alaska, "Data are typically suppressed in small geographic areas, an industry dominated by a single employer, or where one segment of government dominates (but information on federal employees is fully disclosable). This is because if the pool is small enough, it may be possible to distinguish the results of a single or handful of entities" (2022).

There are two ways in which data may be omitted, primary and secondary suppression. Primary suppression is necessary when the identity of an employer or data can be deduced from the numbers. Primary suppression in a particular category is determined by a formula of the BLS that considers the number of establishments, total employment, the number of employers, and the contribution of the largest employers to total wages and jobs (Justis, 2008).

Secondary suppression refers to the act of withholding certain data because they can be easily calculated by using other data that have already been released. For instance, if data for one industry group in a particular county are suppressed, then data for another industry group in that same county, which has the smallest non-zero employment, must also be suppressed. Still, the concept of "total covered employment" is a combination of four distinct types of ownership: private, local government, state government, and federal government. The disclosure rules that govern data from the BLS apply to individual ownership levels rather than the overall total covered employment level. This is done to prevent anyone from using disclosed information to calculate missing values (Justis, 2008). This kind of suppression increases significantly as data move from two- to six-digit NAICS categories for a finer inference. Therefore, to take a holistic view with less uncertainty about the labor mix, suppression at both the primary and secondary levels must be tackled simultaneously.

Development of this project's algorithm was deeply rooted within the RAS method. The nomenclature of the RAS method is believed to refer to Richard Stone, who co-authored the

System of National Accounts (SNA) paper with Abraham Aidenhoff . They devised a method to balance estimated values of the input-output or supply-use tables iteratively across employment levels. The procedure runs the two-stage process outlined in the previous section in an amalgamated fashion.

3.2. Data Processing

Before the procedure is executed, the data must be preprocessed. This ensures that the requirements for a successful run are satisfied. To begin, the QCEW data for a user-specified year are downloaded directly from the BLS website in CSV format. All estimations of suppressed values are made directly to the CSV table dynamically. It is within the preprocessing that the totals for employment and wage data are fixed. For both the employment and wages fields, additional fields are created to store the fixed totals, the minimum and maximum, the estimated value, any error that occurred during the suppression estimation process, and a final value resulting from adjustments for one-digit NAICS and 999 county and industry values.

To understand more clearly, wording from familial relationships has been adopted to describe the working of the algorithm. Table 3.1 presents the hierarchical properties of the QCEW dataset, where the NAICS data for all six-digit industries must add up to their five-digit classification, all five-digit data must add up to their four-digit classification, and so on.

Table 3.1 Excerpt of North American Industrial Classification System (NAICS) (2017)

NAICS		***Employment***
3-Digit industry		
311	Food Manufacturing	**1590229**
4-Digit industries of food manufacturing		
3111	Animal Food	60276
3112	Grain and Oilseed Milling	61384
3113	Sugar & confectionery products	75112
3114	Fruit & vegetable preserving & specialty food	172630
3115	Dairy products	144779
3116	Animal slaughtering & processing	510965
3117	Seafood products preparation & packaging	35579
3118	Bakeries & tortilla	311680
3119	Other food	217824
Sum		**1590229**
5-Digit industries of dairy products		

NAICS		*Employment*
31151	Dairy products (except frozen)	123521
31152	Ice cream & frozen dessert	21258
Sum		**144779**
6-Digit industries of dairy products		
311511	Fluid milk	55089
311512	Creamery butter	2798
311513	Cheese	48269
311514	Dry, condensed, evaporated dairy products	17365
311520	Ice cream & frozen dessert	21258
Sum		**144779**

In Table 3.1, food manufacturing is the "parent" of nine three-digit industries. They are its "children," tagged hereditarily or numerically by their parent's complete NAICS code before their own code. To one another, the nine three-digit industries are industrial "siblings." The sibling Dairy products (NAICS 3115) has two "children," Dairy Products (Except Frozen) and Ice cream (NAICS 31151) and Frozen Dessert (NAICS 31152). NAICS 31151 has four "children" at the six-digit level whereas NAICS 31152 has only one "child," Ice Cream and Frozen Dessert (NAICS 311520).

Figure 3.1 graphically represents the parent, child, and sibling relationships. It also shows the relationships among the geographic siblings at the county level, as well as siblings at the industry level. This shows how places add an extra layer of relationships into the "family." So, to consider the data in Table 1 again, Ice Cream and Frozen Dessert are geographic siblings if they are in the same state.

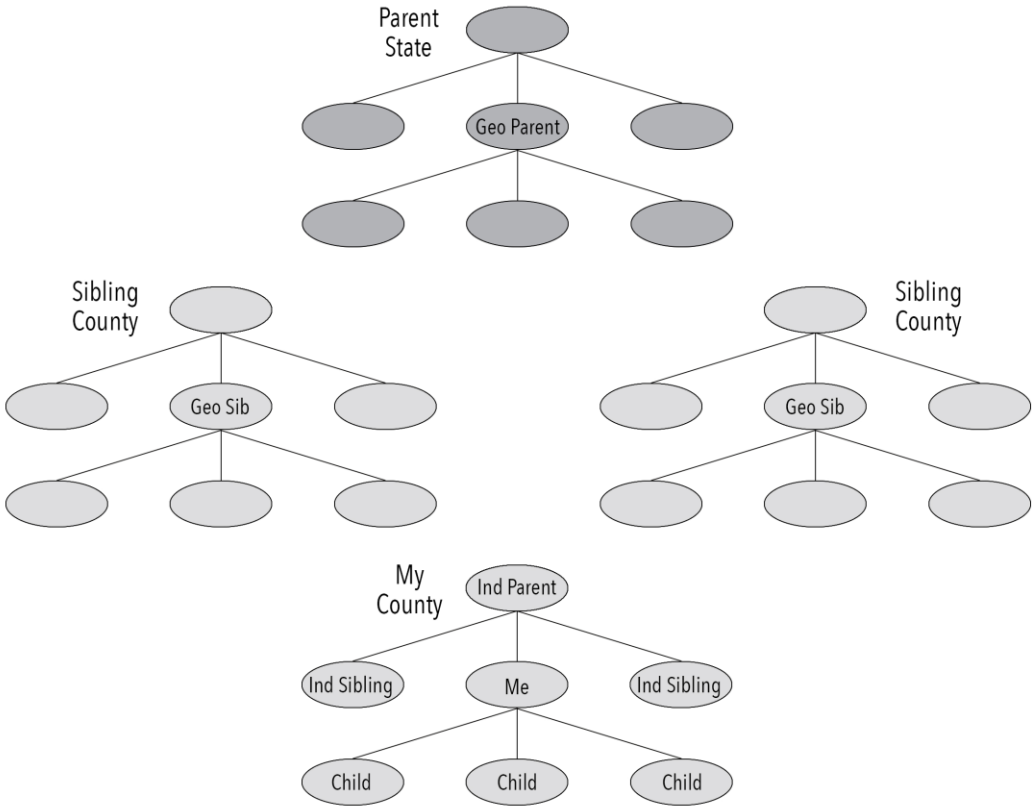

Figure 3.1 Entity Relationship Diagram for the NAICS

This balancing is a great way to verify whether the data are accurate. More importantly, it can help inform decisions based on where to open a business or seek investment. The algorithm begins by downloading the QCEW dataset for a user-specified year and then converts the disclosure column values from "N" and blank to 0 and 1, respectively.

The algorithm begins with calculating the minimum and maximum employment and wage for undisclosed values. This is based on disclosed values in lower and higher levels of the hierarchy (NAICS and spatial). These are factual values that are used as constraints during the assigned estimation stage. These constraints may prevent the estimation process from becoming infeasible. The following calculations can be made. The minimum employment of an industry is its parent industry's minimum minus the summed maximum of its sibling industries. Its maximum employment is its parent's maximum minus the summed minimum of all its siblings.

For industries that are not suppressed, the minimum and maximum are both set to be equal to actual disclosed employment.

The minimum employment of a suppressed industry is the sum of its children's minimum values, and its maximum employment is the sum of its children's maximum values. The above two calculations can also be done for the geographical hierarchies (county, state, national). The calculations occur in the following order. First, at the county-level industry hierarchy. Second, by county-state, and third by state-level industry hierarchy. The fourth calculation is by the state-county and state-national levels. Fifth is by national level industry-wide hierarchy, and then finally the sixth calculation is for the national-state relationship. The calculations must be done in this order, i.e., iteratively, as they are interdependent. Because the algorithm calculates new minimum and maximum values and these are then used to calculate the minimum and maximum values of other industries and at other levels, these calculations are inter-reliant and need to be performed repeatedly in multiple iterations. Each iteration uses the largest minimum and the smallest maximum identified in all previous steps. The above calculations are repeated iteratively until no further adjustments can be made.

From here, the next step is to calculate the actual point estimates for suppressed cells using the input-output table compilation method (Tapestry's version of the RAS algorithm), incorporating the results of the Range Finder. To ensure accuracy, the estimation process takes a top-down approach—starting from the NAICS two-digit codes and working down to the six-digit codes. This is to simplify the algorithm and also to ensure that all undisclosed values have a minimum and maximum specified. To perform this process, the initial minimum for all undisclosed values is set to zero, and the initial maximum is set to a value guaranteed to be greater than any that would be calculated. To maximize the processing performance of the algorithm, it is defined to process one two-digit NAICS hierarchy at a time. Each two-digit NAICS hierarchy is independent of the other two-digit NAICS hierarchies. This limits the number of records that need to be processed within each run of the algorithm.

The Tapestry algorithm is one type of "iterative proportional fitting" technique and is a version of an RAS algorithm. In a scenario in which the sums of data entries (on an input table) are not equal to their margins, which are known as true values, one needs to adjust the values of the entries to make their sums as close to the margins as possible. The adjustment is done

iteratively between rows and columns until the sums of both rows and columns converge to their corresponding margins.

These data are then used to calculate total commodity supply and total commodity demand for each of the 409 BEA commodities. These commodity supplies and demands are then used as an input into the gravity model to estimate county by county commodity trade. The code for the gravity model is given in section 4.2. The employment and wage data are downloadable at: https://tapestry.nkn.uidaho.edu/.

CHAPTER 4. FINDINGS

4.1. Model Performance

The gravity model performed well, and the results of the gravity model are available on the Tapestry data website at https://tapestry.nkn.uidaho.edu/. The code for the gravity model is presented in the appendix of this document. The model ran on a standard desktop in 22 seconds per commodity per year. It ran on a University of Idaho network computer in 3.3 seconds per commodity per year. The data were downloadable by year (Figure 4.1) and by commodity (Figure 4.2).

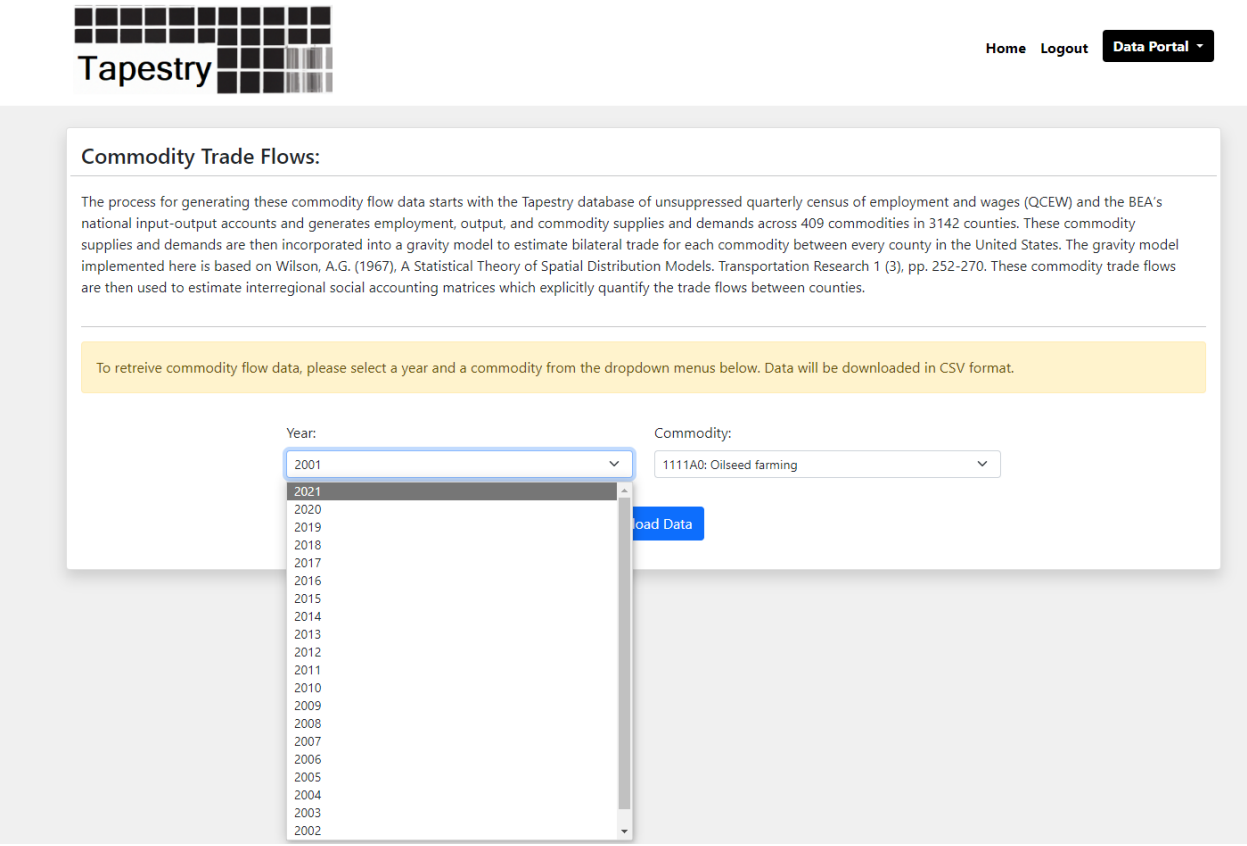

Figure 4.1 Commodity Trade Flows Website Showing the Years of Data Available.

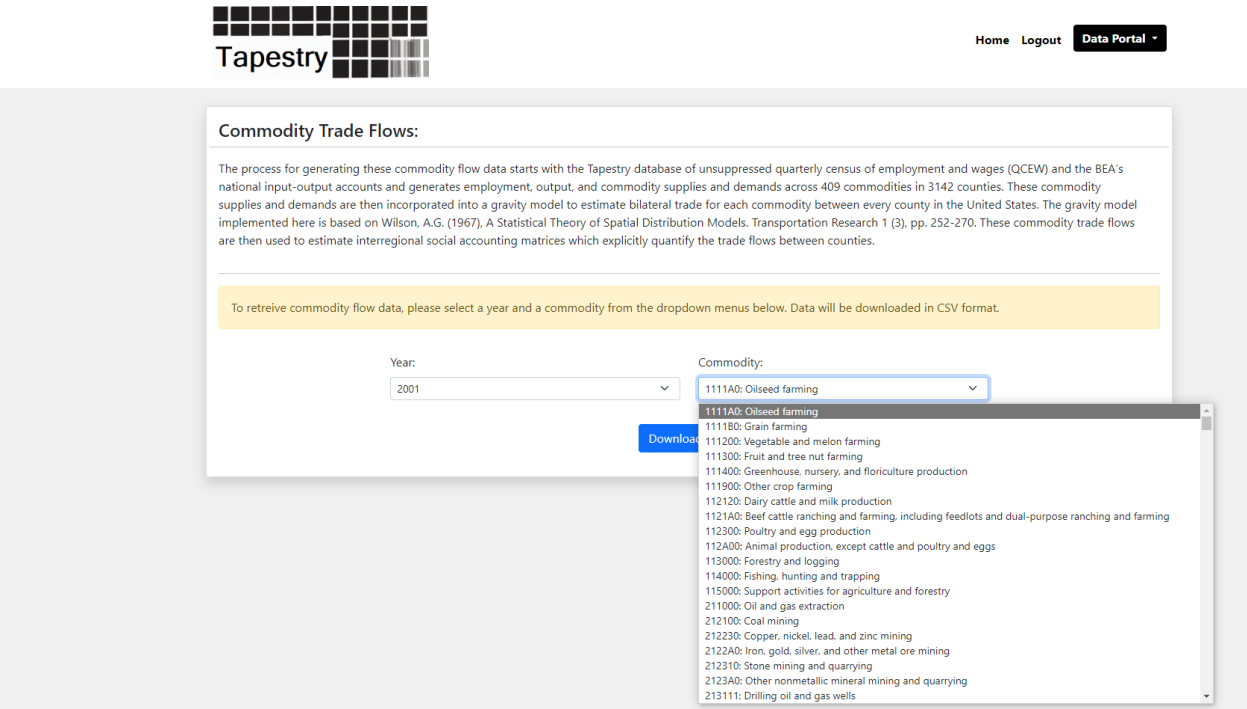

Figure 4.2 Commodity Trade Flows Data Website Showing NAICS Available for Download.

Table 4.1 presents an example of the output of the data model. The actual data were calculated by commodity and by year. Each data set was then 3,142 counties by 3,142 counties in dimension for each commodity. Table 4.1 presents a truncated example for a specific commodity (1119, crop farming) and for six counties. The first five counties are the first five county Federal Information Processing System (FIPS) codes (see https://www.census.gov/library/reference/code-lists/ansi.html for more information on FIPS codes), and are all counties in Alabama. The last county is the last county FIPS code, a county in Wyoming.

The values in Table 4.1 represent the value of trade of the respective commodity between counties. The on-diagonal elements (where the column heading and row heading are identical) represent the value of a commodity that was both produced and consumed in that county.

Table 4.1 Trade Between Selected Counties for Commodity 1119, "Other Crop Farming"

FIPS	01001	01003	01005	01007	01009	01011	…	56045
01001	19.40764	0.36018	0.282788	0.770301	0.698348	0.86339	…	0
01003	0	10.38624	0	0	0	0	…	0
01005	0.411035	0.405897	22.72166	0.317954	0.521723	2.525074	…	0
01007	0.660041	0.328254	0.187437	27.11752	1.063778	0.485254	…	0
01009	0	0	0	0	2.981775	0	…	0
01011	0.378714	0.25936	0.762009	0.248406	0.375946	27.54882	…	0
…	…	…	…	…	…	…	…	…
56045	0	0	0	0	0	0		4.93

CHAPTER 5. CONCLUSION

Creating algorithms to overcome data suppression is likely the most effective solution within a federal data system that must maintain the confidentiality of employers' operational information. However, there is room for improvement, and one way to achieve improvement is to promote the cooperative dissemination of cleaned data and openly discuss strategies for overcoming data suppression. That approach was employed in this algorithm's proof of concept.

In the present day, data can be easily shared within the research community, so there is no need for individual researchers to create their own estimates that are not documented or shared with others. Instead, researchers can work together to develop more sophisticated and accurate methods to overcome data suppression. By working collaboratively and sharing our knowledge and tools, we can move beyond the days of relying on crude estimates and instead develop more effective strategies for analyzing and utilizing suppressed data.

However, this algorithm is not the end-all solution to non-disclosure. This method has a limitation in that the columns in later iterations always converge to their true margins, while the rows in earlier iterations may deviate from their true margins to some extent. The choice of whether to bound the rows or columns is up to the researcher's discretion. In this case, the county total is ensured to be bounded, as it is reported more accurately. Another limitation is that the algorithm requires an input table that accurately represents the reality being studied. The cells are adjusted on the basis of the ratio of "margin/sum," so if the input table is significantly different from reality, the results may not be close to it either. Additionally, cells with zero values will remain at zero. Therefore, the challenge is to create an input table with realistic initial values for point estimates.

To determine the initial values for the algorithm, we used the results from the Range Finder stage, which provided a lower and upper bound value for each suppressed cell. One might assume that calculating the midpoint of the range would be a simple way to generate an initial value, but this approach can lead to poor estimates, especially with higher-digit NAICS codes, whose range tends to be wider. Therefore, relying solely on the midpoint of the range may not be sufficient, and other methods may need to be considered to obtain more accurate initial values for the algorithm. Some future improvements to enhance this dataset could the amalgamation of the BEA's employment-wage data with those of the QCEW to form a well-rounded, queryable, and most complete version of the U.S. economic health. The federal agencies that handle data

security could also grant access to researchers to compare their estimates with actual hard statistics thereby not only improving forecasting models but also informing future policy.

CHAPTER 6. REFERENCES

Alamá-Sabater, L., Márquez-Ramos, L., Navarro-Azorín, J. M., & Suárez-Burguet, C. (2015). A two-methodology comparison study of a spatial gravity model in the context of interregional trade flows. *Applied economics*, 47(14), 1481-1493.

Anderson, J. E., & Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American economic review*, *93*(1), 170-192.

Beige, Sigrun & Axhausen, Kay W., 2017. "The dynamics of commuting over the life course: Swiss experiences," Transportation Research Part A: Policy and Practice, Elsevier, vol. 104(C), pages 179-194.

Cai, M. (2023). A calibrated gravity model of interregional trade. Spatial Economic Analysis, 18(1), 89-107.

Defourny, J., & Thorbecke, E. (1984). Structural path analysis and multiplier decomposition within a social accounting matrix framework. *The Economic Journal*, 94(373), 111-136.

Duranton, G., & Overman, H. G. (2005). Testing for Localization Using Micro-Geographic Data. *The Review of Economic Studies*, 72(4), 1077–1106. https://doi.org/10.1111/0034-6527.00362

Hertz, T., & Zahniser, S. (2013). Is There a Farm Labor Shortage? *American Journal of Agricultural Economics*, 95(2), 476–481.

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

Justis, R. (2008). What Do You Mean the Data Are Suppressed? Understanding the Ins and Outs of QCEW Disclosure Rules. INcontext, 9(7). https://www.incontext.indiana.edu/2008/july-august/2.asp#f2

Koks, E. E., Carrera, L., Jonkeren, O., Aerts, J. C. J. H., Husby, T. G., Thissen, M., ... & Mysiak, J. (2015). Regional disaster impact analysis: comparing Input-Output and Computable General Equilibrium models. Nat. Hazards Earth Syst. Sci. Discuss, 3, 7053-7088.

Lindall, S. A., Olson, D. C., & Alward, G. S. (2006). Deriving multi-regional models using the IMPLAN national trade flows model. Journal of Regional Analysis and Policy, 36(1100-2016-89756). 83–499.

North American Industry Classification System (NAICS) U.S. Census Bureau. (n.d.). Retrieved May 11, 2023, from https://www.census.gov/naics/

Peterson, S. K., & Jessup, E. L. (Eds.). (2008). Evaluating the Relationship Between Transportation Infrastructure and Economic Activity: Evidence from Washington State. Journal of the Transportation Research Forum. https://doi.org/10.22004/ag.econ.206909

Quarterly Census of Employment and Wages: U.S. Bureau of Labor Statistics. (n.d.). Retrieved May 11, 2023, from https://www.bls.gov/cew/

Round, J. I. (1978). On estimating trade flows in interregional input output models. *Regional Science and Urban Economics*, 8(3), 289-302.

Trinh, B., & Phong, N. (2013). A Short Note on RAS Method. *Advances in Management & Applied Economics*, 3, 133–137.

Why certain employment data are suppressed. (n.d.). Retrieved May 11, 2023, from https://live.laborstats.alaska.gov/qcew/empnumsuppressed.html

Wise, R. (n.d.). Questions and Answers (Q&A): U.S. Bureau of Labor Statistics. Retrieved May 10, 2023, from https://www.bls.gov/cew/questions-and-answers.htm#Q14

CHAPTER 7. BIBLIOGRAPHY

Fournier Gabela, J. G. (2020). On the accuracy of gravity-RAS approaches used for inter-regional trade estimation: Evidence using the 2005 inter-regional input–output table of Japan. *Economic Systems Research*, 32(4), 521-539.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., & Shleifer, A. (1992). Growth in Cities. *Journal of Political Economy*, 100(6), 1126–1152.

History: Handbook of Methods: U.S. Bureau of Labor Statistics. (n.d.). Retrieved May 11, 2023, from https://www.bls.gov/opub/hom/cew/history.htm

Hörl, S., Ruch, C., Becker, F., Frazzoli, E., & Axhausen, K. W. (2019). Fleet operational policies for automated mobility: A simulation assessment for Zurich. Transportation Research Part C: Emerging Technologies, 102, 20–31. https://doi.org/10.1016/j.trc.2019.02.020

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, 99(3), 4

McMillen, D. P., & Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics*, 53(3), 321–338. https://doi.org/10.1016/S0094-1190(03)00026-3

Orr, B., & Buongiorno, J. (1989). Improving Estimates of Employment in Small Geographic Areas—IOS Press. *Journal of Economic and Social Measurement*, 15. https://doi.org/10.3233/JEM-1989-153-402

Partridge, M. D. (2005). Does Income Distribution Affect U.S. State Economic Growth?. *Journal of Regional Science*, 45(2), 363–394. https://doi.org/10.1111/j.0022-4146.2005.00375.x

Porter, M. (2003). The Economic Performance of Regions. *Regional Studies*, 37(6–7), 549–578. https://doi.org/10.1080/0034340032000108688

APPENDIX A: GRAVITY MODEL CODE

```
"""
@author: Philip Watson: pwatson@uidaho.edu

This calculates trade between regions for a single given commodity
using a fully constrained gravity model

Commodity supply and demand for each region as well as a distances between
regions and impedance factor (weights on distance) are the required inputs

Gravity model Based on Wilson, A.G. (1967), A Statistical Theory of
Spatial Distribution Models. Transportation Research 1 (3), pp. 252-270
"""
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import integrate

# ============================================================================
# INPUT DATA Folder Paths and Parameters
# ============================================================================

#sys.path.append('C:/Users/pwatson/Dropbox/Tapestry/')
# new_dir = "C:/Users/pwatson/Dropbox/Tapestry"
# os.chdir(new_dir)

# This is the commodity supply and demands file:
sup_dem_file = 'comm_sup_dem.csv'

#INPUT DISTANCE DATA HERE (INDEXED ON REGION BY REGION)
# This is the distance/impedance file :
dist_file = 'fips_dist.csv'
m_dist = pd.read_csv(dist_file)
m_dist.set_index('fips', inplace=True)
dist = m_dist.to_numpy()

#INPUT GIVEN COMMODITY SUPPLY (sup) AND COMMODITY DEMAND (dem) DATA HERE (INDEXED
ON REGION)

comm_sup_dem = pd.read_csv(sup_dem_file)
sup = comm_sup_dem['comm_sup']
dem = comm_sup_dem['comm_dem']
tot_sup = sum(sup)
tot_dem = sum(dem)

print('original total supply', tot_sup)
print('original total demand', tot_dem)

#INPUT DISTANCE IMPEDENCE DATA HERE (INDEXED ON COMMODITY)
alpha = 1
beta = -1.1
gamma = 0
# ============================================================================
# CALCULATED VALUES BELOW
# ============================================================================
```

```
s_d = np.diag(sup)
d_d = np.diag(dem)
cost_mat = alpha * np.power(dist,beta) * np.exp(gamma * dist)
att_org = np.matmul(s_d, cost_mat)
att_des = np.matmul(cost_mat, d_d)

# ============================================================================
# Calculate A and B through iteration (about 100 iterations works best
# and CANNOT be an odd number)
# ============================================================================

def calculate_MS_MD_A_B(MS, MD, A, B, iteration):
   if iteration == 100:
      return MS, MD, A, B,
   else:
      MS_new = att_org * A[:, np.newaxis]
      MD_new = att_des * B
      A_new = np.reciprocal(np.sum(MD, axis=1), where=np.sum(MD, axis=1)!=0)
      B_new = np.reciprocal(np.sum(MS, axis=0), where=np.sum(MS, axis=0)!=0)

      return calculate_MS_MD_A_B(MS_new, MD_new, A_new, B_new, iteration+1)

# Initial values of A and B
initial_MS = att_org
initial_B = 1/(np.sum(att_org, axis=0))
initial_MD = initial_B * att_des
initial_A = 1
initial_A = np.array([1.0])

# Calculate A and B recursively
final_MS, final_MD, final_A, final_B = calculate_MS_MD_A_B(initial_MS,
                              initial_MS, initial_A, initial_B, 0)

A = final_A
B = final_B

A_diag = np.diag(A)
B_diag = np.diag(B)

int_prob = np.matmul(A_diag, att_des)
prob = np.matmul(int_prob, B_diag)

# ============================================================================
# Results of the model is the shipping matrix S
# ============================================================================
S = np.matmul(s_d, prob)
S_orig = S

# Set any values less than this value to zero
S[S < 0.001] = 0
# Calculate row and column totals
row_totals = np.sum(S, axis=1)
col_totals = np.sum(S, axis=0)

# Adjust row and column totals to match supply and demand totals
row_factor = tot_sup / np.sum(row_totals)
```

```
col_factor = tot_dem / np.sum(col_totals)

S *= col_factor

S_rowsum = np.sum(S, axis=1)
S_colsum = np.sum(S, axis=0)

# Verify that row and column totals match supply and demand totals
final_sup_total = sum(S_rowsum)
final_dem_total = sum(S_colsum)

print("Total shipped suppy:", final_sup_total)
print("Total shipped demand:", final_dem_total)
Final_S = S

# ========================================================================
# Create plots of trip lengths
# ========================================================================
# create long list of all distances
trips_rounded = np.round(Final_S).astype(int)
trips_rounded = trips_rounded.reshape(1, -1)[0]
dist_rounded = np.round(dist).astype(int)
dist_rounded = dist_rounded.reshape(1, -1)[0]
all_distances = []
for i in range(0, len(dist_rounded)):
   distance = dist_rounded[i]
   count = trips_rounded[i]
   all_distances.extend([distance] * count)

mean = sum(all_distances)/len(all_distances)
print("average trip length", mean)

# Create histogram of trip lengths.
plt.figure(1)
bin_number = min(int((len(set(all_distances))*0.6)), 20)
print("bin size", bin_number)
hist_output = plt.hist(all_distances, bins=bin_number, density=True)
plt.axvline(mean, color='k', linestyle='dashed', linewidth=1)
plt.title("Trip Length Histogram")
plt.xlabel("Impedance (Distance or Travel Time)")
plt.ylabel("Percent of Shipments")
plt.show()

# Create xy plot of trip lengths.
y = hist_output[0]
binEdges = hist_output[1]
bincenters = 0.5 * (binEdges[1:] + binEdges[:-1])
x = bincenters
y_predicted = alpha * np.power(x, beta) * np.exp(gamma * x)
area = integrate.simps(y_predicted, x)
y_predicted = y_predicted/area

plt.figure(2)
plt.plot(x, y_predicted)
plt.plot(x, y)
plt.legend(["Impedance Equation", "OD Distribution"], loc='upper right')
```

```
plt.title("Trip Length Distribution")
plt.xlabel("Impedance (Distance or Travel Time)")
plt.ylabel("Percent of Trips")
plt.show

# ============================================================================
# Write output
# ============================================================================

# Turn cost_mat into a dataframe with index and headers.
cost_mat_df = pd.DataFrame(data=cost_mat)
f_df[zone_id_field] = zone_ids
f_df = f_df.set_index(zone_id_field)
f_df.columns = zone_ids
cost_mat_df.to_csv(r'\F_Cost_Matrix.csv')

# Turn shipments matrix (S) into a dataframe with index and headers.
S_shipments_matrix = pd.DataFrame(data=S)
OD_trip_matrix[zone_id_field] = zone_ids
OD_trip_matrix = OD_trip_matrix.set_index(zone_id_field)
OD_trip_matrix.columns = zone_ids
S_shipments_matrix.to_csv('S_shipments_trip_matrix_1.csv')
```