A Computational Routine for Disaggregating Industry Margin Data to Estimate Product Margin Rates

Matthew Atkinson
Bureau of Economic Analysis
1441 L Street NW
Washington, DC 20230
matthew.atkinson@bea.gov

1 Introduction

The Industry Economics Division (IED) of the Bureau of Economic Analysis (BEA) encounters numerous disaggregation problems in compiling its benchmark input-output tables. Simulation is a powerful tool for dealing with disaggregation problems: Simulation facilitates systematic use of available information—namely, accounting rules (e.g., add-up conditions) and expert knowledge. This paper uses the example of estimating retail product margin rates to demonstrate how simulation can be used to disaggregate data.

The paper has three parts. Section 2 explains the estimation methodology and formalizes the estimation problem in terms of maximizing a joint conditional distribution. Section 3 explains how retail product margin rates are estimated using the methodology. Section 4 presents the results of testing the methodology using synthetic data, and discusses the product margin rates estimated using data from the 1997 Census of Retail Trade. The paper concludes with a comment on the general applications of simulation routines like the one developed herein.

2 Methodology

The Estimation Problem Retail product margin revenue is used to estimate the retail output proportion of final consumption commodities. However, the Census Bureau does not collect margin revenue data categorized by merchandise line. Margin revenue data are available for 73 retail industry groupings (see table 2 at the end of the paper for a list of the industry groupings). Industry product margin rates are estimated by disaggregating retail industry margin data by product line. In compiling the 1997 United States Benchmark Input-Output Tables,[1] IED used margin revenue data for 73 industry groupings to estimate 2,858 retail industry product margin rates.

Assumptions The methodology for estimating industry-by-product margin rates incorporates two Census Bureau data tabulations and one assumption that relates the tabulations to one another. The Census Bureau data tabulations used to produce the estimates are:

(1) The industry margin rates for 73 NAICS retail industry groupings (see table 2);[2]
(2) The proportion of industry product sales-to-total industry sales.[3]

Relating the data tabulations to one another requires a relational assumption: *The correlation between the industry margin rate and a product line's proportion of industry sales is intermediated by the (unknown) product margin rates.* This implicitly means that there is no direct relationship between the industry margin rate and a product line's proportion of industry sales. Most importantly, this means that, after conditioning on product margin revenue, industry product sales revenue and industry margin revenue are independent.

Evaluating the relationship between industry margin rates and the product proportions of industry sales is difficult for two reasons. First, product margin rates vary across industries, so a correlation based on observations from different industries cannot be directly related to individual industries.[4] Second, the revenue proportions for many products tend to covary with

[1] Ann Lawson, Kurt S. Bersani, Mahnaz Fahim-Nader, and Jiemin Guo, "Benchmark Input-Output Accounts of the United States, 1997," *Survey of Current Business* 82 (December 2002): 19-109.
[2] The Census Bureau's Annual Retail Trade Survey program provides BEA with retail industry margin rate data.
[3] The proportion of total industry sales comprised of sales of each merchandise line is calculated using data from the *Retail Merchandise Line Sales* subject series of the 1997 Economic Census.
[4] The variance of product margin rates across industries also precludes the use of regression for estimating product margin rates.

revenue proportions for other products; for example, industries that sell more cars than average tend to sell more tires than average as well. This covariance can bias product rate estimates if it is not controlled in the estimation process. These two complications are nearly impossible to overcome without formal computational methods.

Of course, no method would ever be able to precisely estimate product margin rates based solely on the two tabulations available. Therefore, any formal method that does not make systematic use of industry analyst expertise—a crucial supplement to the Census Bureau data tabulations—will likely be no better than a more informal disaggregation approach. The following model provides the structure needed for formally making use both of the data tabulations and of analyst expertise.

A General Model for Incorporating Expert Information and Accounting Rules This section outlines the disaggregation model in its most general form. It presents a general formula that can be used to incorporate different types of expert information and accounting rules.

Estimation is the process of approximating the values of unknown variables using available data and well-defined assumptions. The optimal *estimate* is determined by finding the matrix of unknown variables that maximizes the joint probability of the observed data and the estimation matrix conditional on the analyst's methodological assumptions and subjective information. The model developed in this section formulates a computationally tractable maximization problem.

The joint probability of the industry margin rates, the industry product sales-to-total industry sales proportions, and an arbitrary matrix of industry-by-product margin rate estimates is represented as:

$$\text{Probability}\left(\vec{m},\, X,\, \widetilde{Y}_i \middle| \text{constraints}\right), \quad (1)$$

where $\vec{m}$ is a vector of industry group margin rates (table 2), X is an industry-by-product matrix of each product's proportion of total industry sales revenue (from 1997 Census of Retail Trade), $\widetilde{Y}_i$ is an arbitrary matrix of industry-by-product margin rate estimates, and 'constraints' includes the relational assumption stated above, accounting rules and expert knowledge.

Joint probability (1) can be reformulated into terms that make maximizing the joint probability function computationally tractable. The problem is rearranged using Bayes' Theorem:

$$\text{P}\left(\vec{m},\, X,\, \widetilde{Y}_i \middle| \text{constraints}\right) = \text{P}(\widetilde{Y}_i \,|\text{constraints}) * \text{P}\left(\vec{m},\, X \middle| \text{constraints},\, \widetilde{Y}_i\right), \quad (2)$$

$$\text{P}\left(\vec{m},\, X,\, \widetilde{Y}_i \middle| \text{constraints}\right) = \text{P}(\text{X} \,|\text{constraints}) * \text{P}\left(\vec{m},\, \widetilde{Y}_i \middle| \text{constraints},\, X\right), \quad (3)$$

and, because the right side of (2) is equal to the right side of (3):

$$P(\widetilde{Y}_i \middle| \vec{m},\, X, \text{constraints}) = \frac{P(\vec{m},\, X \middle| \widetilde{Y}_i, \text{constraints}) P(\widetilde{Y}_i \middle| \text{constraints})}{P(\vec{m},\, X \middle| \text{constraints})}. \quad (4)$$

Assuming k exclusive and exhaustive discrete values of $\widetilde{Y}_i$, the denominator in the right-hand side of (4) is written as a weighted sum of the conditional probabilities $\text{P}(\vec{m},\, \text{X} \middle| \widetilde{Y}_i)$ where the weights are $\text{P}(\widetilde{Y}_i)$:

$$P(\widetilde{Y}_i \middle| \vec{m},\, X, \text{constraints}) = \frac{P(\vec{m},\, X \middle| \widetilde{Y}_i, \text{constraints}) P(\widetilde{Y}_i \middle| \text{constraints})}{P(\vec{m},\, X \middle| \widetilde{Y}_1, \text{constraints}) P(\widetilde{Y}_1) + \ldots + P(\vec{m},\, X \middle| \widetilde{Y}_k, \text{constraints}) P(\widetilde{Y}_k)}. \quad (5)$$

To estimate the optimal $\widetilde{Y}_i$ matrix using equation (5), two terms need to be estimated for each $\widetilde{Y}_i$.

First, $\widetilde{Y}_i$ either satisfies the constraining conditions or it does not; therefore, $P(\vec{m},\, X \middle| \widetilde{Y}_i)$ is either 1 or 0.

The second term, $P\left(\widetilde{Y}_i \middle| \text{constraints}\right)$, is the prior distribution of industry product margin rates—a distribution the industry analyst sets before conditioning the estimates on the Census Bureau data tabulations. Well-considered prior distributions are crucial for tackling the enormous parameter space this problem involves in spite of the paltry data available for estimating the parameters. The prior confines the search for the optimal matrix $\widetilde{Y}_i$ within the relatively small area of the parameter space

where the solution is almost certain to exist. If no expert information is available regarding industry product rates, a well-dispersed distribution centered at the reported industry margin rate can be used to set the prior for each industry product rate. If expert information is available, the prior product rate distributions should be adjusted accordingly. Of course, the quality of all priors is not equal. The distributions of priors based on relatively good information should be more concentrated about the distribution center than priors based on sketchy information.

Finally, the denominator in equation (5) is the summation of the k numerators associated with $\widetilde{Y}_1,...,\widetilde{Y}_k$.

Equation (5) is evaluated using simulation. Evaluating equation (5) using simulation begins with sampling a matrix $\widetilde{Y}_i$ from the prior probability function P(Y|constraints). Next, the equation $\widetilde{Y}_i$*X'=$\vec{m}$ is evaluated; if the equation is true $P(\vec{m}, X|\widetilde{Y}_i)$ is 1 (otherwise $P(\vec{m}, X|\widetilde{Y}_i)$ is 0). After thousands of $\widetilde{Y}_i$'s are sampled, the sample frequencies of $\widetilde{Y}_i$ are used to estimate the probability distribution of Y conditional on $\vec{m}$, X, and the constraints. The probability distribution's most probable matrix $\widetilde{Y}_i$ is approximately equal to the estimation matrix that maximizes joint probability (1). In other words, the simulation routine approximates the most probable set of estimates of the industry-by-product margin rates.

3 Estimating Retail Industry Product Margin Rates

Three accounting rules and two forms of subjective information are easily applied using simulation. Each of these data constraints is explained in turn.

Accounting Rules The first accounting rule imposes deterministic bounds.[5] All product margin rates are bound within 0 and 100 percent of the respective product's industry sales revenue. An important corollary of this constraint is that no product margin rate can assume a value that causes another product rate to breach the boundaries—binding many margin rates on intervals much smaller than [0, 100]. Formally, the deterministic bounds for an industry product margin rate are:

$$\max\left(0, \frac{(m|\gamma) - (1-x)}{x}\right) \le y \le \min\left(\frac{(m|\gamma)}{x}, 1\right),$$

where m is the industry margin rate, x is the product's proportion of industry sales, γ represents the industry's other product margin rates, and y is the (unknown) industry product margin rate. These bounds relate each industry product rate to every other product rate within the same industry.

The second accounting rule uses the known margin rate grand average. The margin rate grand average is the same regardless of whether it is calculated using industry margin rates and industry revenue proportions or product margin rates and product revenue proportions. Therefore, the second accounting rule requires that the product rate grand average implied by a candidate matrix of industry product margin rates $\widetilde{Y}_i$ equal the known industry rate grand average.

The third accounting rule uses the known average margin rate across industries. Industry margin rates are assumed to be directly related to industry product margin rates. Thus, *after conditioning on industry product margin revenue*, industry product sales proportions and industry margin revenue are independent. This provides a mechanism for comparing product margin rates across industries. The third accounting rule incorporates this information by requiring that the average product margin rates for a candidate estimation matrix $\widetilde{Y}_i$ accurately aggregates to the average industry margin rate. That is, the following equation must be true:

$$r = \vec{b}_i * \vec{p}',$$

where r is the average industry margin rate, $\vec{b}_i$ is the vector of average product margin rates implied by estimation matrix $\widetilde{Y}_i$, and $\vec{p}$ is a vector of each product's average proportion of industry sales. While imposing deterministic bounds relates product rates within each industry, this accounting constraint relates product rates across industries. Thus every industry

[5] The use of deterministic bounds in the routine is motivated by King's (1997) ecological inference method, which combines ecological regression (Goodman 1953) and the method of bounds (Duncan and Davis 1953).

product rate is directly or indirectly related to every other estimated industry product margin rate: When the accounting rules are imposed, each industry product rate estimate 'borrows strength' from every other estimate.

Expert Information Expert information helps define the search area by informing the prior distributions. Retail firms and industry associations publish voluminous data—providing a basis for subjectively approximating many industry product margin rates. Simulation helps leverage this information while taking into account the quality of the information. Simulation is especially useful for making use of industry information for two reasons:

(1) IED needs to estimate industry product margin rates that are consistent with Census Bureau data. Estimates using industry data must be sensitive to the potential for differences in sample representation and methodology.

(2) There is no information for informing many industry product margin rate estimates. Therefore it is critical that the estimation process use the available information to confine the problem in a manner that sheds light on the likely values of the product rates for which little information is available.

Two forms of expert information inform the estimation of industry product margin rates. The most basic form of expert information involves assumptions about the relative magnitude of product margin rates. For example, it is fairly safe to assume that the margin rate for optical goods is higher than the margin rate for automotive fuels. All relationships of this form can be included as constraints.

The second type of expert information is approximate knowledge of an industry product margin rate. Approximate knowledge of an industry product rate is used to set the center and dispersion of the rate's prior distribution. For some products, industry product margin rates can be reasonably well approximated; where these product rates are concerned, the prior distribution is dispersed relatively tightly about the distribution center. For other products, there may be good reason to believe that industry product rates are very different than the margin rates reported by the industries selling them; where these product rates are concerned, even if exact rates cannot be accurately approximated, rate estimates can be improved by using a best guess to set the center of a widely-dispersed prior distribution.

4 Disaggregate Estimates

The methodology explained in section 2 was used to estimate product margin rates based on data from the 1997 Census of Retail Trade. But, first, the methodology was tested using synthetic data.

Synthetic data was generated following the structure of the data tabulations available for the 73 industry groups. First, for each of the 42 product lines, the center of the product rate distribution was randomly generated. Next, the product rate centers were used to set product line margin rate distributions, and industry product margin rates were randomly drawn from these distributions. Finally, industry margin rates were calculated using the industry-by-product sales proportions matrix and the synthetic industry product margin rates.

The simulation program was used to recover the product margin rates based on the product sales proportions and the industry margin rates. The test assumed that the ordinal relationship between product margin rates was accurately known and that reasonably good estimates of the industry margin rates were available for setting half of the prior distributions. That is, the test was designed to assess whether—given reasonably good knowledge of half of the margin rates and the ordinal relationships among the product rates—the simulation program could accurately estimate product margin rates. The results of the test are reported in table 1. As the results show, 28 of the 42 product margin rates were accurately recovered within 2.5 percentage points. Only four of the rates were inaccurately estimated by more than five percentage points.

The results of disaggregating the industry margin rates to obtain product margin rate estimates are reported in table 3 for Census Bureau broad merchandise line categories. These estimates are, in general, very consistent with the estimates IED has produced using a clinical approach. The estimates are also consistent with industry publications—including those that were not used to inform the prior distributions.

The margin rates reported by a large grocery store chain were used to evaluate some of the margin rates estimated using the disaggregation routine. The Marketing Group at the University of Chicago School of Business complied a database of sales revenue and cost of goods sold records for 29 product categories sold by 96 stores run by Dominick's Finer Food—one of the

largest grocery store chains in the Chicago metropolitan area.[6] Of course, the product margin rates of one grocery store chain operating in a single metropolitan area are imperfect proxies for national grocery store industry product margin rates, but, because the grocery store industry is competitive, Dominick's Finer Food's product margin rates serve as a useful proxy for six Census merchandise line categories where comparable University of Chicago product categories exist. In general, the product margin rate data for Dominick's corroborate the accuracy of the simulation routine. Comparisons for the six comparable (detailed) merchandise line categories are presented in Table A. For all six merchandise lines, the difference between Dominick's rate and the estimated margin rate can be reasonably explained by difference in the specific product sales composition of the merchandise line category (between the national average and Dominick's) and by the magnitude by which Dominick's unique product margin rate can be expected to deviate from the national average product margin rate.

Table A: Comparison of Estimated Grocery Stores (NAICS 4451) Margin Rates and Dominick's Finer Food Average Product Margin Rates for 6 June 1996 to 14 May 1997

Census Merchandise Line	Estimated Margins for NAICS 4451	Dominick's Product Category	Dominick's Average Margin Rate
Bottled, canned, or packaged soft drinks	30	Bottled, canned, or packaged soft drinks	31
Candy	34	Candies, gums, and bars displayed at check-out registers	43
Nonprescription medicines	32	Pain relievers and related products	38
Other hygiene needs (including deodorants etc)	31	Grooming products (deodorants, razors, etc.)	35
Paper & related products (including paper towels, toilet tissue, etc)	16	Paper towels	21
		Bathroom tissue	20
Soaps, detergents, & household cleaners	18	Liquid and powder laundry detergents	23
		Liquid and powder dish detergent	24

5 Conclusions and Future Extensions

The routine presented in this paper demonstrates how estimation based on clinical intuition can be successfully implemented quantitatively. Indeed, the clinical approach to estimation is basically a process of making inferences using intuition to evaluate quantitative data. Thus when the premises of intuitive inference are explicit, effective quantitative methods can be designed by mathematically formalizing the intuitive relationships.[7]

In fact, the estimation of retail industry product margin rates is merely a specific example of the disaggregation problem often encountered in producing economic estimates. The routine described in this paper can be applied to many other estimation processes in which the disaggregation problem arises.

References

Goodman, Leo. 1953. Ecological regressions and the behavior of individuals. *American Sociological Review* 18, 663-664.

Duncan, Otis Dudley, and Beverly Davis. 1953. An alternative to ecological correlation. *Am Sociological Rev* 18, 665-66.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton UP.

[6] The University of Chicago Graduate School of Business Marketing Group's Dominick's Finer Food database is available at the following web site: http://gsbwww.uchicago.edu/research/mkt/.

[7] Inference is the process of drawing the conclusion from the premises. In economic estimation, most inference rules are easily mathematically formalized. Computationally evaluating the data with respect to the inference rules is more precise than clinical inference.

Table 1: Synthetic Data Product Margins

Industry Code	Estimated Product Margin Rate	True Product Margin Rate	Abs Dif b/t Estimate and True Value
29	56.67	56.72	0.04
33	30.80	30.88	0.08
27	54.57	54.74	0.17
36	46.80	46.97	0.17
21	66.59	66.41	0.18
32	61.35	61.13	0.23
10	35.87	36.13	0.26
2	18.52	18.17	0.35
24	36.52	36.96	0.44
6	14.38	13.89	0.48
23	20.88	21.41	0.53
25	68.70	69.28	0.59
42	39.54	38.82	0.72
41	49.25	50.15	0.90
20	40.84	41.74	0.90
30	70.12	71.13	1.01
17	39.10	38.06	1.04
22	54.94	53.81	1.13
9	43.22	44.42	1.20
19	43.42	42.15	1.27
28	47.50	48.87	1.37
11	41.92	43.30	1.38
39	50.44	51.98	1.54
18	42.78	44.34	1.56
12	22.69	21.03	1.67
15	40.46	42.45	1.99
13	46.36	48.73	2.37
37	32.59	30.20	2.39
34	42.83	45.44	2.61
35	54.07	56.93	2.85
16	51.00	54.21	3.21
31	53.57	49.81	3.76
14	26.48	22.20	4.28
38	62.55	66.92	4.38
26	30.10	25.69	4.41
8	45.88	50.46	4.59
7	72.35	67.76	4.60
1	16.23	11.30	4.93
3	22.77	17.75	5.02
40	26.16	20.14	6.02
5	48.93	55.14	6.21
4	36.21	42.59	6.38

Table 2: 1997 ARTS Industry Margin Rates

NAICS	Industry Description	Industry Margin Rate
441	Motor Vehicle and Parts Dealers	19.81
44111	New Car Dealers	16.96
44112	Used Car Dealers	*
44121	Recreational Vehicle Dealers	*
441221	Motorcycle Dealers	*
441222	Boat Dealers	*
441229	All Other Motor Vehicle Dealers	*
4413	Automotive Parts & Tire Stores	36.89
442	Furniture and Home Furnishings S	42.23
44211	Furniture Stores	*
44221	Floor Covering Stores	*
44229	Other Home Furnishings Stores	*
443	Electronics and Appliance Stores	25.16
443111	Household Appliance Stores	*
443112	Radio, TV & Other Electronics	*
44312	Computer and Software Stores	*
44313	Camera and Photo Supplies	*
444	Bldg Material and Garden Equip	26.71
4441	Building Material and Supplies	25.89
44413	Hardware Stores	*
4442	Lawn and Garden Equip	*
445	Food and Beverage Stores	26.04
4451	Grocery Stores	25.39
4452	Specialty Food Stores	34.43
44523	Fruit and Vegetable Markets	*
445291	Baked Goods Stores	*
445292	Confectionery and Nut Stores	*
445299	All Other Specialty Food Stores	*
44531	Beer, Wine, and Liquor Stores	26.93
446	Health and Personal Care Stores	31.74
44611	Pharmacies and Drug Stores	26.81
44613	Optical Goods Stores	*
44619	Other Health and Personal Care	*
447	Gasoline Stations	22.05
448	Clothing & Clothing Accessories	41.47
44811	Men's Clothing Stores	44.01
44812	Women's Clothing Stores	39.36
44813	Children's and Infants' Clothing	*
44814	Family Clothing Stores	39.84
4481	Clothing Stores	46.66
44821	Shoe Stores	41.53
44831	Jewelry Stores	*
44832	Luggage and Leather Goods	*
451	Sporting Goods, Hobby & Book	37.79
45111	Sporting Goods Stores	*
45112	Hobby, Toy, and Game Stores	*
45113	Sewing & Needlework	*
45114	Musical Instrument and Supplies	*
451211	Book Stores	*
451212	News Dealers and Newsstands	*
45122	Tape, CD, & Record Stores	*
452	General Merchandise Stores	26.73
4521	Department Stores	28.30
4521101	Conventional Depart Stores	*
4521102	Discount or Mass Merchandising	22.13
4521103	Chain Department Stores	*
45291	Warehouse Clubs & Superstores	19.64
45299	All Other General Merchandise	34.80
453	Miscellaneous Store Retailers	43.33
45311	Florists	*
45321	Office Supplies and Stationery	*
45322	Gift, Novelty, and Souvenir	*
45331	Used Merchandise Stores	*
4539	Other Misc Store Retailers	*
453991	Tobacco Stores	*
45393	Manufactured Home Dealers	*
454	Nonstore Retailers	42.81
45411	Electronic & Mail-Order Houses	42.82
45421	Vending Machine Operators	*
454311	Heating Oil Dealers	*
454312	Liquefied Petroleum Gas	*
454319	Other Fuel Dealers	*
45439	Other Direct Selling Est	*

* denotes public datafile rate that is not published

Table 3: Estimated Product Rates

Product Description	Product Margin Rate
Groceries & other foods	27.78
Meals, unpackaged snacks	58.18
Packaged liquor, wine, & beer	31.59
Cigars, cigarettes, tobacco	34.90
Drugs, health aids, & beauty aids	29.60
Soaps, & household cleaners	19.25
Paper & related products	19.31
Men's wear	38.82
Women's, juniors' wear	36.39
Children's wear	35.59
Footwear	39.21
Sewing & knitting goods	34.67
Curtains, draperies, blinds, etc.	34.41
Major household appliances	22.26
Televisions, video recorders, etc.	26.13
Audio equipment & musical inst	34.87
Furniture & sleep equipment	40.67
Flooring & floor coverings	43.08
Computer hardware, software	27.63
Kitchenware & home furnishings	36.86
Jewelry	39.07
Books	41.67
Photographic equipment	31.88
Toys, hobby goods, & games	36.48
Optical goods	64.34
Sporting goods	35.66
Recreational vehicles & parts	23.16
Hardware, tools, etc.	30.43
Lawn, garden, & farm equip	26.28
Dimensional lumber	25.20
Paint & sundries	24.99
Manufactured (mobile) homes	27.62
Automobiles, vans, trucks, etc.	11.06
Automotive fuels	8.74
Automotive lubricants (oil, etc.)	22.64
Automotive tires, tubes, etc.	36.48
Household fuels	42.32
Pets, pet foods, & pet supplies	43.12
Line 850 All other merchandise	41.41
Line 9810 All other merchandise	22.97
Nonmerchandise receipts	61.89