Regional Research Institute
West Virginia University

Working Paper Series

Aggregation Bias and Input-Output Regionalization

Detail or Accuracy?

Randall Jackson, Caroline Welter, and Gary Cornwall

Working Paper Number 2022-01

Website address: `rri.wvu.edu`

Aggregation Bias and Input-Output Regionalization

Randall Jackson*[1], Caroline Welter†[1] and Gary Cornwall‡[2]

[1]Regional Research Institute, West Virginia University
[2]U.S. Bureau of Economic Analysis

June 8, 2023

Abstract

Conventional wisdom holds that results from input-output (IO) models with greater sectoral detail are superior to those from models with less detail. However, there is an implicit assumption that the more detailed data are as accurate as their aggregated counterparts. In this paper, we explore the tradeoffs between sectoral detail and model accuracy in the context of IO regionalization, a practical context in which greater sectoral detail is commonly achieved via the imputation of missing values. This reality is especially apparent for increasingly smaller geographical regions where privacy concerns result in more suppressed and undisclosed data. As the number (or share) of disaggregated values that require imputation increases, the disaggregated model results will also deviate further from perfect accuracy. Is there a point at which using an aggregate model with greater certainty – relying on more reported and less imputed data – will provide results that are superior to a disaggregated model with greater potential imputation error and uncertainty? To address these questions, we design and implement simulation experiments founded on the concept of aggregation bias that enable us to evaluate the likelihoods that aggregate models would be superior to their disaggregated counterparts.

Keywords: Input-output, Regionalization, Aggregation bias
JEL Classification: R1

*rwjackson@mail.wvu.edu
†caroline.welter@mail.wvu.edu
‡gary.cornwall@bea.gov

1 Introduction

A generally accepted tenet in modeling system behaviors is that greater detail in classifying groups of actors and agents results in greater within-group homogeneity and between-group differentiation, hence superior model results. This is perhaps nowhere more true than in models of economic or industrial systems, where similar establishments are grouped into industries for analysis. Given a *ceteris paribus* choice between an economic model with a 400-industry classification scheme and another with only 70 industries, for example, few analysts would choose the 70-sector model. Nor would many analysts use a 20-sector model if a 70-sector model were available, and so on.

Economic input-output (IO) models provide a useful case in point, as a good bit of attention has centered around inaccuracies due to aggregation.[1] A great many of these studies appeared in the early IO literature and focused on sectoral aggregation, with a few extending the analysis to spatial aggregation. Lahr and Stevens (2002) addressed role of regionalization in the generation of aggregation error in regional input–output models, with emphasis on whether, in the pursuit of an aggregated regional IO model, the first step should be regionalization or aggregation.[2][3]

Missing from IO aggregation research, especially in the context of regionalization methods, is the recognition that for most of the variables that are used in regionalizing national IO accounts, increasing levels of sectoral detail also increase reliance on databases with greater levels of suppressed, undisclosed, masked, or imputed estimates. In practice, then, the operative question extends beyond aggregation bias to a combination of bias and uncertainty, or what we refer to here as reliability. Instead of asking a modeler whether she would prefer a model with more industrial detail to one with less, the question becomes one of whether a less detailed model based on more observed and reported underlying data might be preferred to a more detailed model based on less reliable data? Or more generally, would one

[1] See Lahr and Stevens (2002) for a thorough review of IO aggregation issues and the interplay between aggregation and regionalization.

[2] In IO modeling, regionalization refers to adapting national IO accounts to the regional level using of region-specific data.

[3] In the early days of IO modeling, the use of aggregated models was driven in part by computational constraints that no longer apply, perhaps with the exception of IO models used in computable general equilibrium (CGE) or hybrid econometric input-output models, where solution complexity can rise rapidly with increasing numbers of industries.

prefer a more reliable aggregated model to a less reliable but more detailed model, and what is the nature of the tradeoff between the two? In this paper, we confront these questions by an experimental design that enables an empirical assessment of the tradeoffs, and further characterizes the potential error that can arise from using less reliable detailed data.

2 Aggregation Bias in IO Models

In this section, we provide the foundation for the aggregation bias measure that we use first in our simulations. The justification for focusing first on Aggregation bias is that it is among the few analytical metrics that explicitly relate aggregated and disaggregated IO accounts. Consider a Leontief input-output model expressed as:

$$g = (I - A)^{-1} f \tag{1}$$

where $a_{ij} = z_{ij}/g$, $z_{ij} \in Z$ is a matrix of interindustry flows, $f_j \in f$ is a column vector of final demand values, and $g_j \in g$ is output from industry j.

The dimensions of equation (1) can be of order n or m, where $n > m$, and S is an m X n aggregation matrix that defines the relationship between n and m, such that $Sg = g^*$, and likewise, $SZS' = Z^*$, where the * denotes the m-dimensional variables.[4]

Miller and Blair (2022) present Morimoto's (1970) definition of total aggregation bias as $\tau = g^* - Sg$. Here, g^* is the result from the aggregated model and Sg is the aggregated result from the disaggregated model. Expanded, we have:

$$\tau = (I - A^*)^{-1} f^* - S(I - A)^{-1} f \tag{2}$$

or

$$\tau = \left[(I - A^*)^{-1} S - S(I - A)^{-1}\right] f \tag{3}$$

and in equivalent power series expansion,

[4] $f^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 + f_3 \\ f_4 \end{bmatrix}$ from Miller and Blair (2022).

$$\begin{aligned} \tau = & \left[\left(I + A^* + A^{*2} + \ldots\right) S - S\left(I - A + A^2 + \ldots\right)\right] f \\ \tau = & \left[\left(A^* S - SA\right) + \left(A^{*2} S - SA^2\right) + \ldots\right] f \end{aligned} \tag{4}$$

The first term in the series, $(A^*S - SA)f$, is defined as "first-order" aggregation bias ((Theil, 1957)):

$$\psi = \left(A^* S - SA\right) f \tag{5}$$

First-order aggregation bias disappears if $A^*S = SA$, and first-order bias is zero when the nonzero elements in the final demand vector are not aggregated, as would be the case in their example:

$$f = \begin{bmatrix} f_1 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

and

$$f^* = Sf = \begin{bmatrix} f_1 \\ 0 \end{bmatrix} \tag{7}$$

An implicit assumption underlying the conventional wisdom is that all underlying data at the different levels of aggregation and detail are known and accurate. It is clear that under this assumption, more detailed models will result in more accurate results and thus be preferred to more highly aggregated models, especially in modern times when computational power is no longer a binding constraint. In practice, however, this assumption does not hold because published data with greater detail generally have increasing numbers of undisclosed or imputed data (nondisclosure to protect privacy has long been a common practice used by government reporting agencies). The recent adoption of a newly developed deferentially private publication system by the U.S. Bureau of the Census amplifies levels of uncertainty in underlying detailed data (Abowd, 2018; Dwork, 2019). Whereas the application of disclosure rules has generally been viewed as the explanation for undisclosed and unreported data, even reported data can no longer be assumed to be accurate. The problem becomes more severe as sectoral detail increases and as region size (in terms of economic activity) decreases. Hence, the trade-off between more accurate aggregated data on the one hand and more detailed but less reliable data on the other is increasingly clear in the context of input-output modeling. How might the conventional wisdom change after explicitly recognizing this trade-off?

Assume that we have access to all of the *true*, perfectly accurate detailed data for the foundation of analysis and that the corresponding aggregate data are derived from the detailed data. In this extreme case, we have 100% reported and accurate detail-level and aggregated data. The aggregation bias theorem results apply without qualification and the first-order bias in the aggregated model results can be measured by ψ.

As we deviate from this *perfect and complete information* scenario, the number (or share) of inaccurate detailed values due to differential privacy and imputation increases and the disaggregated model results will deviate further from perfect accuracy. Is there a point at which the accuracy of an aggregate model with greater certainty will be preferred to a less reliable disaggregated model with greater detail?

We explore this question using a simulation approach that randomly perturbs the values in the detailed vector of output by industry, g. To maintain meaning and comparability, we require that our experimental simulation design must respect the constraint that the sum of the (perturbed) output values in the detailed industries (the child industries) must equal the true aggregate (parent sector) output, g_j^*, or:

$$\sum_{i \in j} g_i = g_j^*, \forall\ i,\ j \tag{8}$$

where industry $i \in j$ indicates that detailed industry i belongs to aggregate industry j. This constraint is met by applying proportional adjustments to the randomly perturbed values at each simulation step.

3 Experimental design in the context of input-output regionalization

Our primary interest in this paper is to explore these issues in the context of building regional IO accounts by combining region-specific data with national IO accounts. This is the standard approach to regional IO because IO accounts are rarely available for most sub-national regions. Employment and or compensation data by industry are typically used to estimate output by industry. Industry output data from the state of Illinois are then used as regional versions of corresponding national accounts counterpart variables and parameters that then approximate a region's industrial structure. Illinois

was selected for this initial analysis as a state with a sufficiently developed and highly integrated inter-industrial system to support meaningful and representative results. A large and integrated system is also less likely to be based on imputed values in place of reported accounts regionalization data.

This down-scaling process also carries implications for final demands. Industries that are concentrated within a given region, for example, will likely have greater surplus output available for exporting, and output from industries that are relatively less regionally well-represented might need to import relatively more than their national counterparts. Thus, while equation (5) can be used to estimate the combined effect of aggregation under conditions of disaggregated data uncertainty, we use equation (9), below, to capture these effects along with the effect of corresponding changes to estimates of final demand.

$$\psi^s = A^* S f - S A^s f^s \tag{9}$$

The final demand in the first RHS term of equation (9) is the true final demand vector used in equation (6) but f^s in the second RHS term is the final demand vector derived during the regionalization procedure. The first RHS term corresponds to the reference point for a perfectly aggregated regional model that was regionalized using true regional data while the second RHS term corresponds to the aggregated result from a model that was regionalized using simulated output values that abide by the constraints of equation (8). The difference between the two terms reflects bias due to a combination of aggregation *and* final demand effects, which accurately reflects most practical input-output applications based on currently available – and unavailable – published data.

3.1 Total Final Demand

The aggregation bias measure developed by Morimoto (1970) operates within an industry-by-industry accounting framework. Because modern accounts most commonly report final demand in commodity space, we need not only to reformulate the aggregation bias measure using the elements of the commodity-by-industry accounting frameworks, but we also need to formulate the appropriate final demand totals in commodity space.

Following Jackson and Járosi (2022), we see that export demand and

other final demand, which includes consumption, investment and inventory adjustment, and government expenditures, enter the accounting identity for open regional systems differently. We know that for open regional systems, substantial portions of other final demand are expected to be satisfied by imports. Hence, the conversion of other (commodity) final demand to industry output differs from the conversion from export final demand to industry output.

Let commodity demand for exports be f^{ex} and other commodity demand be f^D, and the total final demand be $f^t = f^D + f^{ex}$. Since our goal is a vector of final demand for regional industry output, the total final demand satisfied by regional industries after the transformation of commodity space to industry space is:

$$f^t = D(\widehat{Q}f^D + f^{ex}) \tag{10}$$

or

$$f^t = \widetilde{D}f^D + Df^{ex} \tag{11}$$

Where $D = V\widehat{q}^{\,-1}$ and $\widetilde{D} = D\widehat{Q}$. Q is a vector of proportions of own-region commodity demand satisfied by regional industries, otherwise known as regional supply proportions, calculated as $Q = ((q - \widehat{f^{ex}} + m)^{-1})(q - f^{ex})$. The effect of D is to reallocate commodities used by industries, to industries production of these commodities, regardless of their origin. Thus, the pre-multiplication of D by exports means all the commodities demanded for exports that the region satisfied by its domestic industry production.

Following the central point of Jackson's (1998) regionalization method, the make table (V) is standardized by total regional commodity supply rather than domestic commodity production (q). The standardized make table $\widetilde{D} = V/s^{-1}$, where the total supply is equal to the commodity output plus imports ($s = q+m$) can also be rewritten as $\widetilde{D} = V\widehat{q}^{-1}\widehat{Q}$ (see (see Jackson and Járosi, 2022).

$\widetilde{D}$ gives us not only the commodity demand supplied by domestic industries, but it also gives us the proportion of regional commodity supply that is imported by subtracting each column sum of $\widetilde{D}$ from 1.0. Thus, $\widetilde{D}$ captures the industry production of commodities from their domestic inputs (region) or from the rest of the world (imports). The pre-multiplication of $\widetilde{D}$ by the fixed components of final demand has the effect of removing the final demand imports since only the portion of regional final demand to be satisfied

by the region's own industries is the target. In other words, this identifies the final demand that the region's production can satisfy. Because of the relationships among industry output estimates and final demands, the final demand vector f^t will change with each output simulation draw. Therefore, to find the first-order aggregation bias we need:

$$\psi = (A^*S - SA)f^t \tag{12}$$

$$\psi = (D^*\widehat{Q}^*B^*S^I - S^ID\widehat{Q}B)D(\widehat{Q}f^D + f^{ex}) \tag{13}$$

In computing ψ note that for any given region the first term $(D^*\widehat{Q^*}B^*S^I)$, which is the aggregated technical coefficients matrix A^* will remain constant through subsequent simulations, and the regionalization method will yield the ψ for the $zero^{th}$ (no-change) iteration on all simulations, making this the benchmark ψ. Because the data for the $zero^{th}$ iteration is perfectly accurate by assumption, the benchmark ψ vector will represent the true aggregation bias values.

3.2 Simulation Design

The simulation framework is designed so as to perturb child industries in a way that respects the adding up constraint of equation (8). Simulated output vectors, g^s, drive the regionalization process and subsequent outcome assessment. Five of the 20 parents have only a single child industry, so these parents and their children retain their true values throughout all simulation draws. Parent sectors and industry aggregation schemes are shown in Table 1. All children's industries are listed in the Appendix.

We implement two kinds of simulations: one parent sector at a time and all parent sectors at once. For the case of one parent at a time, children of only one parent are perturbed to generate one set of data supporting FAB and multiplier analysis for each perturbed parent. We anticipate fewer substantial results at in this first case but evaluating one parent at a time allows us to identify any parent whose perturbations might dominate the sensitivity of outcomes. The case of all parents at once is the most relevant case, as reliability issues will not be limited to subsets of industries in practice. In this case, the children of all parents with more than one child are perturbed, supporting the empirical basis for a single FAB and multiplier analysis.

Table 1: Parent Sectors

Children	Parents	Sector Name
1, 2	1	Agriculture, forestry, fishing, and hunting
3 - 5	2	Mining
6	3	Utilities
7	4	Construction
8, 17, 20 - 23	5	Wood leather, textile, and paper products
9 - 12	6	Mineral and Metal Products
13, 14, 18	7	Electronic and Electrical products
15, 16	8	Transportation products
24-26	9	Chemical products
19	10	Manufacturing
27	11	Wholesale trade
28 - 31	12	Retail trade
32 - 39	13	Transportation and warehousing
40-43	14	Information
44-49	15	Finance, insurance, real estate, rental, and leasing
50 - 55	16	Professional and business services
56 -60	17	Educational svcs, health care, and social assistance
60- 64	18	Arts, entertainment, accommodation, & food svcs
65	19	Other services, except government
66 - 67	20	Government

For every perturbed parent sector, we add random shocks to the true child-industry output values. The size of each shock is benchmarked to the magnitudes of corresponding true industry output values. Benchmarking the shocks to true output values helps ensure against extreme changes to the true distributions of industry output values for a given parent. Post-shock child output values are then re-scaled to sum to parent sector totals. Perturbed output vectors then drive a regionalization process, resulting in a unique disaggregated regional model that supports the calculation of interest variables for assessment. More formally, each simulation draw follows these steps:

1. Simulate a trial regional output vector.

    a. Add a random disturbance to every disaggregated output value,

g_i from $\sim N(0, \frac{g_i}{4})$.

2. Impose the adding-up constraint.
3. Invoke the regionalization method to generate A^s and f^s.
4. Compute and store ψ^s and industry multipliers.

4 Analytics

Our analysis falls into three kinds of approaches. Two of these are based on probability and the third relies on descriptive statistics.

4.1 Statistical Measures

Our first approach is to apply a standard statistical method to determine whether the average simulation value would be expected conditional on the null (no difference in the FAB) being true. This is the traditional approach to assessing the outcomes of our simulations and is achieved using standard t-tests. With known true values and distribution of observed values, a significant t-value will reject the null hypothesis $H_0 : \psi = S^{-1} \sum_{i=1}^{S} \hat{\psi}^s$ versus an alternative $H_1 : \psi \neq S^{-1} \sum_{i=1}^{S} \hat{\psi}^s$.

The second approach compares the likelihoods of two specified competing values: the true value and the modal simulated value. We use the pdf of the simulation results to identify these likelihoods and then use them to assess which is more likely and by how much. These odds ratios then complement the traditional t-test by estimating how much more likely, conditional on the simulated distribution of outcomes, that a specific value will be observed than the null value. We would interpret an odds ratio of 3, for example, as H_1, the modal value of $f(\hat{\psi})$, being three times more likely to be observed than H_0. More plainly, the odds of observing H_0 as compared to H_1 are 25% (1 in 4) against, and the odds are 75% (3 in 4) in favor of H_1.

Odds ratios also can be converted to approximate probabilities (p-values) that indicate the statistical significance of the odds ratios. Odds ratios are typically less statistically significant than corresponding t-tests because where t-tests evaluate H0 vs. all other possible values, the odds ratios evaluate H_0 against a specific H_1 value. The interpretation of a 2 to 1 odds ratio

holds, however, irrespective of its p-value and thus provides information that is not provided by the t-test.

To further clarify the odds ratio, consider the example in Figure 1. Suppose by perturbing a parent sector's child-industry output values we obtain the FAB distribution identified as *pdf A*. We would like to know: what are the odds against the null hypothesis of no bias, that is $H_0 : \mathbb{E}(fab) = 0$? Following Mills (2018) and Cornwall et al. (2019), we can calculate the height of the distribution (F) in some ϵ window around C, the most likely value observed by our simulations, and repeat the process (G) around our null value, 0. We can then form the posterior odds by simply taking the ratio, F/G, which represents the "odds against the null hypothesis".[5]

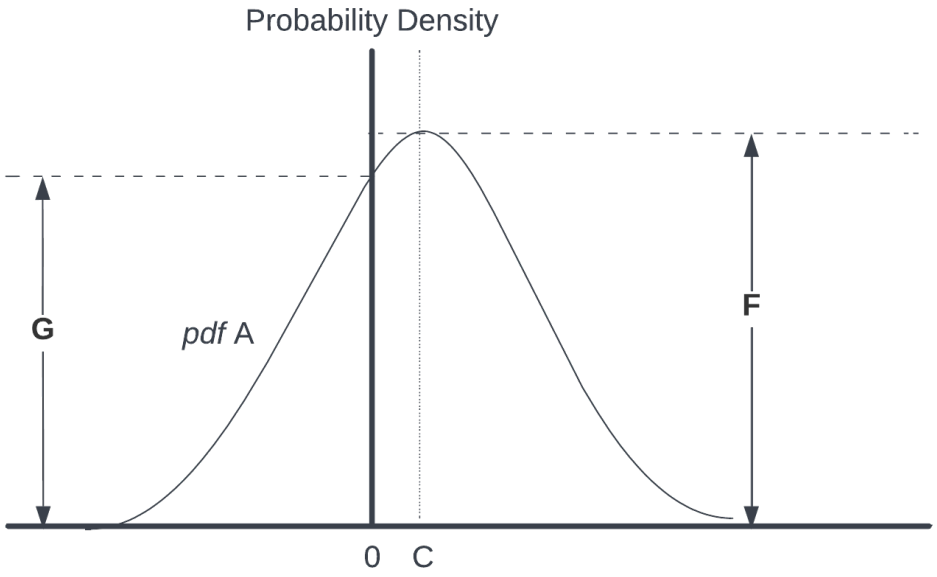

Figure 1: Odds Ratio

[5]Shifting the distribution to the right or left to modify the hypothesized null value would be functionally equivalent.

4.2 Descriptive Measures

Neither the t-tests nor the odds ratios, however, measure the magnitudes of the deviations of simulated outcomes from true values. The third dimension of our analysis adds three more measures, namely the symmetric mean absolute percentage error (SMAPE), the root mean squared error (RMSE), and the root mean squared percentage error (RMSPE). The SMAPE identifies the average deviation from the true value, giving equal weight to all observations. The RMSE is similar but adds weight to outliers, which is relevant for guarding against observing extreme values, and the RMSPE presents the RMSE relative to the size of the variable in question.

MAPE is widely used and consists of a relative measure that expresses the absolute error between the actual value and the forecast value. It is good for when the outcome variables depend upon the proportional size of the errors relative to the true data, which is the point we are making here. However, the disadvantage of using MAPE is that lacks statistical theory and is influenced by outliers (Makridakis (1993)). In order to avoid the latter, we corrected the formula, using what is called the symmetric MAPE:

$$SMAPE = \frac{100}{n} \sum \frac{|\psi^s - \psi^t|}{(|\psi^t + \psi^s|)/2} \tag{14}$$

where ψ^t is the "true" FAB, ψ^s is the FAB resulted from the simulations design and n is the number of simulation trials.

The formula for multiplier analysis follows the same form:

$$SMAPE = \frac{100}{n} \sum \frac{|M^s - M^t|}{(|M^t + M^s|)/2} \tag{15}$$

where M denotes multiplier variables.

The root mean square error measure is shown below.

$$RMSE = 100(\sqrt{\frac{1}{n} \sum\nolimits_{n=1}^{N} \mu^2}) \tag{16}$$

where μ is the percentage difference between simulated and benchmark values for the variables of interest.

SMAPE and RMSPE were calculated as percentages of the multiplier value less 1 rather than as a percentage of the standard multiplier value. The adjustment to the denominator reflects the recognition that 1.0 is an invariant

multiplier floor in this context, while the multiplier effects are captured in the remainder. I.e., a multiplier of 1.0 indicates no multiplier effect.

4.3 Results

In this section, we apply these analytics to FAB and output multipliers. The FAB metric is a vector whose elements indicate the bias in each respective parent sector. We evaluate each FAB element independently to provide insight into the implications of data unreliability on bias for each parent sector. Output multipliers are Leontief inverse column sums derived from interindustry coefficient matrices on each simulation draw. We focus on all parents at once results given that they are the most relevant case. One parent at a time results are discussed in the sensitivity analysis.

4.3.1 FAB

The FAB analytics for FAB for the case of perturbing all parents at once are revealing. Table 2 presents the salient statistics. The benchmark FAB values per parent are presented in column 2 for reference. P-values for the t-statistics in column 3 were all significant at $p < 0.01$, so the column was omitted. None of the odds ratios exceed 1.5, and none of them were shown to be statistically significant. Note, however, that an odds ratio of 1.5 would correspond to the odds of H_1 being observed 3 times for every two observations of the benchmark FAB.

The SMAPE values range from a low of around 2% to a high of 169%. There is an inverse relationship between SMAPE and the absolute value of FAB, with a correlation of -0.38; smaller FAB values are subject to a larger percentage error. The RMPSE is shown in column 7. As expected from its heavier weighting of extreme values, and with the exception of one parent sector (FIRE), RMSPE values consistently exceed their corresponding SMAPE values. In the case of parent 2 (Mining) – which has a very small FAB – the RMSPE is more than an order of magnitude larger than SMAPE.

Together, these statistics support the existence of a trade-off between greater model detail and model accuracy. The t-statistics reject the null hypothesis H_0 = Benchmark FAB conditional on the associated density function. The odds ratios suggest that in nearly all cases, there is a greater likelihood of observing the most frequently observed simulated value than the true FAB, though the odds ratios are not generally large and none are

statistically significant. Of greater concern might be the scale of bias potentially introduced by unreliable data. The average SMAPE over all parents exceeds 60%, and the average RMSPE is nearly three times as large. Hence, there is a clear risk that greater detail can introduce substantial changes to model outcomes.

Given these results, a closer evaluation of results at the disaggregated level is clearly warranted. The next section presents the results of the multiplier analysis.

Parent	Fab	t	Odds	SMAPE	RMSE	RMSPE
1	-68.9	-29.0	1.03	42%	48.7	71%
2	-2.6	-88.7	1.30	169%	46.0	1775%
3	372.9	-49.2	1.03	94%	422.5	113%
4	-118.3	-20.5	1.00	136%	507.0	429%
5	-804.9	4.5	1.06	21%	207.1	26%
6	-379.4	32.2	1.00	115%	731.1	193%
7	-708.2	95.4	1.08	25%	193.6	27%
8	-394.0	-25.7	1.04	17%	86.0	22%
9	-1376.7	101.0	1.01	69%	1125.5	82%
10	-421.6	-22.1	1.00	30%	159.2	38%
11	-209.0	-140.6	1.49	78%	357.2	171%
12	337.0	64.9	1.12	57%	314.4	93%
13	2470.0	47.6	1.01	26%	893.0	36%
14	-696.0	-2.8	1.01	64%	557.1	80%
15	-4003.3	145.9	1.47	95%	3674.3	92%
16	-32888.0	76.9	1.03	9%	3486.8	11%
17	13543.9	59.9	1.00	2%	323.1	2%
18	375.7	35.5	1.00	87%	516.3	137%
19	-249.5	-65.3	1.06	58%	217.8	87%
20	-906.0	-48.5	1.05	15%	186.3	21%

Table 2: FAB Analytics

4.3.2 Multipliers

Perturbing all eligible parent sectors at once – those that have multiple children – reflects the more likely scenario where data unreliability is not re-

stricted to small subsets of industries. These simulations provide the foundation for evaluating the impacts of detailed data unreliability on detailed outcomes, namely impacts on output multipliers.

As was the case with one parent at a time, t-statistics were almost uniformly statistically significant. The average odds ratio over all industries was 2.85, indicating that observing the H_1 value was 65% more likely than observing the true multiplier. Forty-two of the 67 multiplier odds ratios were greater than 2 to 1, and the odds ratios for 21 industry output multipliers were greater than 3 to 1. Twelve multiplier odds ratios exceeded 4 to 1. Twenty-nine of 67 odds ratios were statistically significant at $p < 0.10$, with 12 significant at $p < 0.05$. All SMAPE values exceeded 2.35%, 22 were between 4.0% and 5.0%, and 42 were greater than 5%. RMSPE showed a maximum value of 13.78%, and for all but eight industries, the RMPSE exceeded 5%. The average RMSPE value is 6.39%.

Table 3 shows the salient output multiplier statistics when all parents are perturbed at once, for the industries with the top 20 RMSPE, in sort order. T-statistics and their p-values, virtually all of which were significant, have been omitted for clarity. To summarize the multiplier results, the t-statistics virtually all reject the null hypothesis that the true multiplier would be observed conditional on the corresponding distribution of simulated values. Likewise, the odds ratios suggest that for 42 industries, the specific alternative value was at least 50% more likely to be observed than the true benchmark multiplier. SMAPE values indicate that that nearly 60% of the multipliers are subject to greater than 5% error, and that number grows to 88% when extreme outliers are weighted more heavily by RMSPE.

4.3.3 Sensitivity Analysis

Perturbing each parent's child industry values one parent at a time provides the foundation for assessing whether any unreliability attributed to any one parent dominates impacts on model outcomes at the detailed level of analysis via assessing the FAB results and the detailed output multipliers. In the FAB results, virtually all of the 300 t-statistics generated (15 multiple-child parents X 20 FAB elements) were statistically significant. Only sixteen of the 300 odds ratios had values exceeding 2.0, but eleven of those odds ratios exceeded 100. FABs for parents 4, 10, 15, and 19 were among the most susceptible to large FAB impacts in these simulations.

Perturbing each parent's industry children one parent at a time generates

Industry	Multiplier	Odds	p-value	SMAPE	RMSE	RMSPE
47	2.275	1.69	0.10	10.97%	0.176	13.78%
61	1.398	1.33	0.20	7.18%	0.034	8.67%
45	1.652	4.77	0.04	7.76%	0.056	8.52%
41	1.516	1.47	0.43	7.01%	0.043	8.25%
30	1.496	1.56	0.20	6.91%	0.041	8.21%
10	1.705	1.06	0.77	6.87%	0.058	8.18%
44	1.413	5.59	0.04	7.39%	0.033	8.03%
34	1.846	2.71	0.10	6.90%	0.068	8.01%
36	1.635	4.89	0.04	7.02%	0.050	7.84%
13	1.210	1.11	0.38	6.14%	0.016	7.69%
39	1.623	1.77	0.11	6.20%	0.047	7.55%
11	1.646	1.13	0.66	6.30%	0.049	7.52%
32	1.422	2.66	0.16	6.56%	0.031	7.40%
50	1.332	2.47	0.10	6.32%	0.024	7.36%
56	1.365	1.68	0.16	5.66%	0.026	7.16%
53	1.481	4.03	0.05	6.38%	0.034	7.14%
31	1.558	2.37	0.11	6.00%	0.039	7.06%
49	1.419	5.03	0.04	6.30%	0.029	6.95%
33	1.520	7.46	0.03	6.34%	0.036	6.91%

Table 3: Multiplier Analytics.

a very large number of outcomes for output multipliers, but was motivated by the same rationale as that for FAB analytics summarized above. Nearly all generated t-statistics were highly significant for the multiples, with odds ratio values nearly all between 1.0 and 2.0. Therefore, the case of one parent at a time confirms the patterns of the results found in the all parents at once simulation design, meaning that the results are robust and are not sensitive to different ways of perturbing the regional output.

5 Considerations and Conclusion

This paper began by recognizing that the conventional wisdom regarding the advantages of models with greater detail might be offset by poorer reliability of detailed model data and that in the context of input-output modeling, especially at the regional level, greater detail in published data is almost always less reliable than data published at higher levels of aggregation. This is easily confirmed by an examination of virtually any economic data series published at varying levels of industrial classification and geographical extents, and it is quite clearly evident in published regional employment, wages, and output data series.

In light of this recognition, our objective was to begin to identify the nature of the trade-offs between model detail and model accuracy by focusing on the use of data in the context of generating regional IO accounts. The preference for detailed regional IO models over those with less sectoral detail has long been implicit in the literature, but the question of the effect of increased unreliability has never been explicitly assessed. Our research set out to fill this void in the literature.

We designed a simulation framework that could be used to evaluate the impacts on unreliable detailed data used in generating regional IO models derived from national accounts. We focused first on aggregation bias because it is among the few well-known metrics that explicitly include both aggregate and disaggregated versions of the same IO model. First-order aggregation bias (FAB) thus provided a convenient mechanism for the initial evaluation of the extent of the detail-reliability tradeoff. From a traditional statistical standpoint, the dominance of statistically significant t-statistics indicated that a more detailed assessment was warranted, which led to a more restricted approach using odds ratios to quantify the relative likelihood of observing the true values vs. the modal value from the simulated distributions of corresponding values. Odds ratios confirmed for some parent sectors, the odds of observing the true values conditional on the simulated distributions were quite low.

While FAB provides a mechanism for assessing the relationship between aggregated data and disaggregated data, we are also interested in the implications of using unreliable data at the detailed level, in terms of the potential error in detailed model outcomes. To assess these relationships, we added an analysis of detailed industry multipliers. To this end, using the same simulation framework, we derived distributions of detailed industry multipliers

and assessed them using t-tests and odds ratios, finding again that nearly all t-statistics were highly significant and many odds ratios were large enough to be statistically significant.

Combined, these analyses indicate a strong potential for statistically significant differences between model outcomes based on perfectly accurate and less reliable data. However, the identified statistical differences alone do not reveal the magnitude of potential errors that might arise. Hence, we added descriptive measures to quantify the sensitivity of the model to unreliable detailed data. The results for FAB revealed the potential for an average of 60% error in individual elements of the FAB vector, and for a more intuitive and practical assessment, revealed that differences in excess of 5% might be expected for the majority of industry multipliers.

Our findings are tempered by several considerations. First, these results are based on empirical findings for a single region, and while selected for having a representative industrial structure, there might be peculiarities specific to our test region that influence the outcomes. Second, our analysis is conducted at a level of aggregation that is already somewhat high. In a practical setting, U.S. IO analysts more commonly decide between using the 71-industry classification of published annual IO accounts and the 405-industry classification of the quinquennial accounts. Future research will leave regionalization aside and focus solely on comparisons between these two levels of aggregation. And third, the results presented here are a function of the definition of random shocks that we used to simulate inherent data uncertainty. Sensitivity analyses using alternative definitions in future research will reveal much about whether the identified risks of trading model accuracy for greater industrial detail are justified.

References

Abowd, J. M. (2018). The u.s. census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, page 2867, New York, NY, USA. Association for Computing Machinery.

Cornwall, G. J., Mills, J. A., Sauley, B. A., and Weng, H. (2019). Predictive testing for granger causality via posterior simulation and cross-validation.

In *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A*. Emerald Publishing Limited.

Dwork, C. (2019). Differential privacy and the us census. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '19, page 1, New York, NY, USA. Association for Computing Machinery.

Jackson, R. (1998). Regionalizing national commodity-by-industry accounts. *Economics Systems Research*, 10(3):223–238.

Jackson, R. and Járosi, P. (2022). Io-snap regionalization 2.0. *Technical Document*.

Lahr, M. L. and Stevens, B. H. (2002). A study of the role of regionalization in the generation of aggregation error in regional input–output models. *Journal of Regional Science*, 42(3):477–507.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, 9(4):527–529.

Miller, R. E. and Blair, P. D. (2022). *Input-Output Analysis: Foundations and Extensions*. Cambridge University Press, 3 edition.

Mills, J. A. (2018). Objective bayesian precise hypothesis testing.

Morimoto, Y. (1970). On aggregation problems in input-output analysis. *The Review of Economic Studies*, 37(1):119–126.

Theil, H. (1957). Linear aggregation in input-output analysis. *Econometrica*, 25:939–954.

A Appendix

Industry	Industry Name
1	Farms
2	Forestry, fishing, and related activities
3	Oil and gas extraction
4	Mining, except oil and gas
5	Support activities for mining
6	Utilities
7	Construction
8	Wood products
9	Nonmetallic mineral products
10	Primary metals
11	Fabricated metal products
12	Machinery
13	Computer and electronic products
14	Electrical equipment, appliances, and components
15	Motor vehicles, bodies and trailers, and parts
16	Other transportation equipment
17	Furniture and related products
18	Miscellaneous manufacturing
19	Food and beverage and tobacco products
20	Textile mills and textile product mills
21	Apparel and leather and allied products
22	Paper products
23	Printing and related support activities
24	Petroleum and coal products
25	Chemical products
26	Plastics and rubber products
27	Wholesale trade
28	Motor vehicle and parts dealers
29	Food and beverage stores
30	General merchandise stores
31	Other retail
32	Air transportation
33	Rail transportation
34	Water transportation

35	Truck transportation
36	Transit and ground passenger transportation
37	Pipeline transportation
38	Other transportation and support activities
39	Warehousing and storage
40	Publishing industries, except internet (includes software)
41	Motion picture and sound recording industries
42	Broadcasting and telecommunications
43	Data processing, internet publishing, and other information services
44	Federal Reserve banks, credit intermediation, and related activities
45	Securities, commodity contracts, and investments
46	Insurance carriers and related activities
47	Funds, trusts, and other financial vehicles
48	Real estate
49	Rental and leasing services and lessors of intangible assets
50	Legal services
51	Computer systems design and related services
52	Miscellaneous professional, scientific, and technical services
53	Management of companies and enterprises
54	Administrative and support services
55	Waste management and remediation services
56	Educational services
57	Ambulatory health care services
58	Hospitals
59	Nursing and residential care facilities
60	Social assistance
61	Performing arts, spectator sports, museums, and related activities
62	Amusements, gambling, and recreation industries
63	Accommodation
64	Food services and drinking places
65	Other services, except government
66	Federal government defense
67	Total government nondefense

Table 4: Industries Names.