Exploring Noise Infusion for Disclosure Avoidance at BEA

Background report for June 9, 2023 presentation at Federal Economic Statistics Advisory Committee (FESAC) meeting

John Bockrath and Dan Yorgason

Abstract: *The Bureau of Economic Analysis (BEA) has traditionally used cell suppression for disclosure avoidance in the tables it publishes based on the business surveys it conducts. In part to allow for the possibility of publishing new statistics from these surveys, BEA is considering adopting EZS noise infusion as a replacement for cell suppression for at least some of its surveys—initially for the survey program on international trade in services. This background report provides an overview, in the context of the services trade data, of the EZS noise infusion approach and some of the likely outcomes of adopting such an approach.*

Disclosure avoidance for BEA surveys

The Bureau of Economic Analysis (BEA) is not primarily a producer of survey data; in large part, it is a user of survey data and other source data produced by other government agencies and other external entities. The exception to this rule is in its international economic accounts. An important portion of the source data for these accounts comes from surveys BEA collects on international trade in services, foreign direct investment, and the activities of multinational enterprises.[1] These BEA survey data are used in several BEA data products: data from quarterly surveys on international trade in services ("services trade") are used in the international transactions accounts (ITAs), the international services accounts, and a monthly joint release with the Census Bureau on international trade.[2] The trade in services data are also used in BEA's national economic accounts, where they feed into calculations of GDP.

The quarterly direct investment data are used in the ITAs, the international investment position (IIP) accounts, and statistics on direct investment by country and industry. They are used in the national economic accounts, as well, where they feed into calculations of corporate profits and gross national income. Through the national economic accounts and the IIP accounts, the trade in services and direct investment data ultimately feed into the Integrated Macroeconomic Accounts, which bring together data from BEA's National Income and Product Accounts and the Federal Reserve's Flow of Funds accounts, using consistent definitions, and present the information in a unified framework.

BEA's surveys of multinational enterprises are annual and cover the same enterprises as its surveys of direct investment, but the data collected there are primarily used for BEA's statistics on activities of multinational enterprises.

All of BEA's surveys are either censuses or deterministic (primarily cut-off) samples. Consequently, BEA does not publish measures of estimate variability arising from sampling. Estimates from the surveys are,

[1] For example, roughly 20 percent of current-account transactions in the international transactions accounts were collected on BEA's trade in services and direct investment surveys in 2022, and roughly 27 percent of the asset and liability positions in the international investment position accounts were collected on the direct investment surveys.

[2] The quarterly survey data are interpolated or extrapolated to generate monthly estimates in the monthly trade release.

of course, subject to other common sources of error in survey data such as misreporting, imputation error for non-sampled or non-responding units and/or items, and incomplete sampling frames. BEA does not estimate the variability arising from these sources.[3]

As with other statistical agencies, BEA is obligated by law to protect the confidentiality of the firms that respond to its surveys. BEA has historically used cell suppression for disclosure avoidance, and cell suppression continues to be used at present in tables based on BEA survey data. However, cell suppression comes with certain disadvantages—the most obvious being the missing values created in tables protected by suppression. In addition, suppression becomes frequent, and more difficult when calculations are made in a variety of domains. For instance, trade in services survey data are used to produce statistics on trade by service type and country; roughly 15 percent of cells are suppressed in those tables. Within the last several years, those data began to also be used to produce statistics on potentially "information and communications technology" (ICT)-enabled trade[4] by service type and country. To avoid increasing the share of suppressions in the main services trade tables, the suppression burden falls heavily on the statistics on trade in potentially ICT-enabled trade in services; in the most detailed such table approximately half the cells are suppressed.

Using EZS noise infusion

BEA is considering replacing cell suppression with EZS noise infusion for its surveys of trade in services.[5] The primary goal would be to publish additional statistics—either new tables (e.g., services trade by industry and country), or more detail in existing tables, or both. With cell suppression, ensuring adequate protection for the published statistics becomes increasingly complex as the number of published dimensions or complicated table hierarchies rises. These difficulties do not arise with noise infusion.

Using EZS noise infusion instead of suppression would also eliminate the data "holes" in published tables. This is an important benefit on its own, but it also helps facilitate the goal of producing additional statistics. As new tables or more granular statistics are produced from the same underlying microdata, the share of cells suppressed tends to rise. Thus, even when technical complexities do not prevent the publication of additional statistics, the usability of the statistics becomes increasingly degraded.

[3] Because BEA produces no estimates of existing sources of variability, the estimates of variability discussed later in this paper refer solely to the variability induced by the application of noise infusion.
[4] ICT-enabled services are services with outputs delivered remotely over ICT networks. BEA publishes estimates on potentially ICT-enabled services because information is not available on which services are actually enabled by ICT.
[5] EZS noise infusion refers to a noise infusion method developed at the Census Bureau by Timothy Evans, Laura Zayatz, and John Slanta; see Evans, Zayatz, and Slanta (1998). See also Federal Committee on Statistical Methodology and pp. 14-15 in McKenna and Haubach (2019).

Finally, using noise infusion would ideally simplify, to a certain degree, the disclosure avoidance process. In applying EZS noise infusion, we'd aim for it to provide the same general level of protection provided by current suppression techniques.

The benefits of noise infusion do not come for free. The most obvious downside is that any published cell values based on survey data are distorted with noise infusion relative to the underlying reported data. For cells that would otherwise be suppressed, this is not a problem, as a distorted estimate is usually preferred to no estimate, but users might hesitate to accept the distortion to the rest of the cells. From the user perspective, the tradeoff is illustrated in tables 1 and 2 (which are purely hypothetical and include only an example hierarchical structure). In table 1, most cells are undistorted and perfectly reflect the underlying reported data. However, some of the cells show no data value. At best, data users can construct ranges for these suppressed values, though doing so requires a level of sophistication that not every data user possesses, and the ranges in some cases are very broad relative to the value of the suppressed estimate. In table 2, no cells are suppressed, but the cell values throughout the table only approximately reflect the underlying reported data. Moreover, users don't have complete insight into how imperfect any particular estimate is.[6]

Table 1: Suppression

	Total services	Computer services				Information services		
		Total	Computer software	Data storage services	Other computer services	Total	News agency services	Database services
All countries	19,466	11,625	4,223	2,710	4,692	7,841	3,412	4,429
Europe	10,893	6,578	2,893	861	2,824	4,315	1,993	2,322
France	3,408	1,457	(D)	(D)	665	1,951	954	997
Germany	2,243	1,978	936	187	855	265	137	128
Italy	3,054	2,102	(D)	(D)	962	952	171	781
United Kingdom	2,188	1,041	421	278	342	1,147	731	416
Africa	8,573	5,047	1,330	1,849	1,868	3,526	1,419	2,107
Morocco	2,517	1,398	348	(D)	(D)	1,119	264	855
Nigeria	3,237	1,397	49	610	738	1,840	(D)	(D)
South Africa	2,819	2,252	933	(D)	(D)	567	(D)	(D)

(D) Suppressed to avoid the disclosure of the data of individual companies

[6] For instance, data users don't know that the other computer services/Europe estimate is distorted −0.1 percent and the other computer services/Nigeria estimate is distorted 11.0 percent in Table 2.

Table 2: Noise infusion								
		Computer services				Information services		
	Total services	Total	Computer software	Data storage services	Other computer services	Total	News agency services	Database services
All countries	19,519	11,660	4,236	2,674	4,750	7,859	3,405	4,454
Europe	11,003	6,577	2,860	896	2,821	4,426	2,009	2,417
France	3,364	1,457	589	202	666	1,906	888	1,018
Germany	2,257	2,002	925	176	901	255	124	131
Italy	3,076	2,055	952	208	895	1,021	187	834
United Kingdom	2,307	1,064	394	310	359	1,244	811	433
Africa	8,516	5,083	1,375	1,778	1,929	3,433	1,396	2,038
Morocco	2,468	1,414	366	725	323	1,054	251	803
Nigeria	3,276	1,443	50	573	819	1,833	868	965
South Africa	2,773	2,226	960	480	786	546	276	270

EZS noise infusion could only be applied for new or newly revised statistics. They could not be published for statistics that have been previously published because the combination of noise infused statistics and suppressed statistics could, in certain cases, allow attackers to infer more about protected data of individual survey respondents than would be possible with either one on its own.[7] As a consequence, BEA published tables, which present time series views of trade in services estimates, would be protected by both cell suppression (older periods) and noise infusion (newer periods). In addition, some ITA tables include both cells based on BEA's trade in services surveys and cells based on BEA's direct investment surveys. At present, we are only considering using noise infusion for the cells based on the trade in services surveys, so even within the same year, certain tables might use both cell suppression (direct investment cells) and noise infusion (trade in services cells) to protect confidentiality.[8] Over the longer term, we plan to investigate ways to extend the use of noise infusion to other BEA surveys.

EZS noise infusion overview

The basic mechanics of EZS noise infusion are simple and well known. Each reported data value is assigned a random multiplicative noise factor. The (unperturbed) reported data value is then multiplied by the noise factor to yield a perturbed data value. The perturbed data value is used instead of the

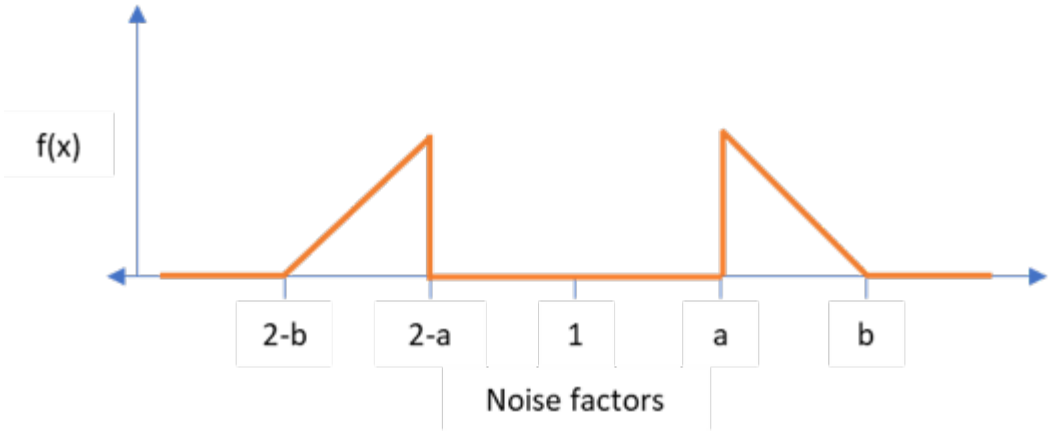
[7] An "attacker" refers to a data user whose focus is using the data product to infer confidential information about one or more survey respondents.

[8] Direct investment survey data present more complexities for the use of noise infusion than trade in services data—partly because of the interrelationships between variables and partly because of the more extensive focus on net values in the former.

unperturbed value as the basis of published table; published cell values are the sums of the perturbed data values over the specified domains.[9]

Under EZS noise infusion, the randomness of the noise factors is typically circumscribed in some manner. Often factors too close to, or too far from, 1 are disallowed. The noise should be significant enough to provide adequate protection for a record, but not so significant as to excessively distort cell estimates. Figure 1 shows a noise factor distribution described in some of the EZS literature; a split symmetric triangular distribution is used with no relative perturbation smaller in magnitude than 100 × (a − 1) percent and no relative perturbation larger than 100 × (b − 1) percent.[10] In principle, a variety of different distributions could be used. We have investigated several distributions, both bounded and unbounded (but all split and all symmetric).

Figure 1. Example Distribution of Noise Factors with EZS Noise Infusion



We have also investigated tailoring noise factor distributions—that is, using more diffuse distribution for records we anticipate being more vulnerable to disclosure given the cells to which they contribute (for example, records in cells dominated by one or two records would be assigned a more diffuse distribution). Our present view is that the benefits conveyed by such tailoring do not outweigh the complications it introduces. Among other things, records contribute both to relatively aggregated quarterly (and monthly) cells and to more detailed annual cells. Thus, the annual cells tend to be more vulnerable to disclosure. However, noise factors need to be applied to, for example, first-quarter records many months before the vulnerability of the annual cell can be measured. Consequently, tailoring the assignment of noise factor distributions to the vulnerability of a cell cannot be done

[9] Some of BEA surveys are censuses and some are samples. Even the sample surveys do not use sample weights, though. Instead, the unsampled (or unreported) records are imputed, so the cell total equals the unweighted sum of reported and imputed cell values. For the purposes of this overview (and for the purposes of disclosure avoidance methods at BEA), imputed values are treated as though they are actually reported values.
[10] For example, see Massell and Funk (2007) and Martin (2013).

precisely; BEA would have to project vulnerability months in advance based on very little data, giving us little ability to target vulnerability in practice.

Not surprisingly, different record-level noise factor distributions result in different cell-level distributions.

In general, disclosure avoidance using EZS noise infusion exhibits the following characteristics:

- Because noise factors can be either above or below one, and because most cells consist of more than one record, the distortions from noise infusion for two or more records can sometimes offset each other. For vulnerable cells, where one or two records contribute disproportionately to the cell total, we intend for such offsetting to be infrequent and incomplete. For non-vulnerable cells, we intend for such offsetting to be frequent and more complete.
- If noise factor distributions are bounded, relative distortions of cells composed exclusively of non-negative records are also bounded. For a measure with both positive and negative contributions (such as a trade balance, which is measured as exports minus imports), however, relative distortions are potentially unlimited, given that the measure itself can be arbitrarily close to zero, regardless of the magnitude of the individual contributions.
- Protection is provided in an expectation sense. Not every vulnerable cell is materially distorted; most are, though, and the expected absolute distortion for vulnerable cells can be tuned to provide adequate protection. This protection-in-expectation is in contrast to cell suppression, where protection is seemingly a discrete phenomenon; either a cell is suppressed or shown without (disclosure avoidance-related) distortion.
- Protection is applied at the level of the individual record, not the level of the cell.
- No formal privacy guarantee is provided.[11]
- No privacy budget is used.[12] As a consequence, there is no clear limit on the number of statistics that can be produced from an underlying set of microdata.

The final three characteristics (particularly the last two) in the list distinguish EZS noise infusion from differentially private disclosure avoidance methods. BEA has not yet investigated formally private methods for its own surveys.

[11] A formal privacy guarantee is a mathematically provable limit on the degree to which any potential attacker, regardless of the amount of outside information they possess, could use the published data product(s) to increase the precision with which they estimate the reported data value of a survey respondent.

[12] Because each piece of information provided in a data product (such as a published table) reveals something about the underlying microdata, a privacy budget is used in formally private methods to ensure that the privacy guarantee is upheld in all statistical products that are based on the microdata. The stronger the privacy guarantee, the smaller the privacy budget, and the less information the statistical office can provide (either less precision, or fewer statistics, or both).

Fine-tuning the noise infusion approach

If individual data records were unrelated to other data records, the noise infusion process described in the previous section would (mostly) suffice to protect respondent confidentiality. However, a record in BEA's business surveys can be related to other records by coming from the same survey respondent on different aspects of the respondent's business at a given point in time or by coming from the same respondent on the same aspect of the business at different points in time. These connections potentially pose problems on different sides of the distortion spectrum. The first, as when company X reports exports to both France and Germany, can result in no *apparent* protection for company X in the "exports to Europe" cell if the infusions to the two individual records happen to offset.[13] The second can result in excessive volatility in the time series of a given cell value if distortions flip back and forth from moderately negative to moderately positive. For instance, if a distortion for a cell flips from a negative three percent to a positive three percent, the apparent rate of growth will be roughly 6 percentage points different than the actual rate.

Researchers at the Census Bureau have addressed similar issues, and we propose to adopt solutions that they have already explored.[14] First, we intend to constrain every noise factor assigned to the records of a given survey respondent to be either all greater than one or all less than one, which is equivalent to requiring every record-level distortion from the respondent be either all positive or all negative. Thus, the distortions for company X cannot offset in the "exports to Europe" cell because neither two positives nor two negatives can sum to zero.

Second, we intend to constrain the noise factors assigned to the records of a given survey respondent to either be above one in all time periods or be below one in all time periods. This limits, but does not eliminate, the degree of time series distortion for the dataset. In addition, another approach we've investigated is assigning some probability of holding constant from one period to the next the noise factor assigned to a data item (e.g., respondent X's exports of service type Y, to country Z) with some probability, but we haven't made a final determination of whether to adopt this approach.

As previously noted, using EZS noise infusion would allow BEA to discontinue cell suppression, and, because protection is applied at the record/respondent level, to publish new tabulations with additional detail (with appropriate level of aggregation) or in new dimensions. In practice, every published table cell that is based on survey data would be distorted to some degree, except for very rare cases where record-level distortions happen to exactly offset. In general, the higher aggregates in tables would be

[13] There would still be protection in the sense of expectations, in line with the discussion above. However, it is possible, if quite unlikely, that company X could be the only contributor to some cell. The company might be concerned to see the sum of its contributions published as the cell total without material distortion, even though the prior likelihood of a non-material distortion was small.

[14] See Martin (2013) and Evans, Zayatz, and Slanta (1996).

little distorted, but distortion would increase, on average, as within-cell concentration increased so that the most vulnerable cells would tend to be the most heavily distorted.

We acknowledge that not all users will agree that small distortions to key aggregates are an acceptable price to pay for the elimination of cell suppression. One option, also foreseen by Census Bureau researchers, is to reset key aggregates in one or more tables to their undistorted values and then "rake" the remaining cells to ensure additivity. For instance, the table margins could be reset to the undistorted totals and the interior cells all made subject to raking. This approach results in a total distortion that consists of the sum of the "pure" noise-infusion distortion and any further adjustment from raking.[15]

One variant of this approach would, in addition, hold the most vulnerable cells at their distorted values to ensure that raking doesn't bring the total distortion in these vulnerable cells too close to zero.

We anticipate that if BEA were to adopt noise infusion, raking would be used to allow non-vulnerable aggregates that feed into BEA's gross domestic product estimates (and potentially other estimates in BEA's National Income and Product Accounts and Industry Accounts) to be published without noise infusion-related distortion. The impact on cells that remain subject to distortion would generally be to increase the diffusion of their distortions.[16] Raking could also partially undo some of the intended impact of keeping noise factors all above one or all below one for multiple records from the same survey respondent or across time periods for the same record.

Potential outcomes

In this section of the report, we present some illustrative (not definitive) examples of how noise infusion for BEA's microdata from surveys on trade in services might affect its published statistics. The examples are based on distortions in International Services table 2.2, which is BEA's most detailed trade in services table and is published annually.[17] This is a relatively large 2-dimensional table. One dimension is geography; this dimension includes 90 countries, regional groupings, or international organizations, along with 4 rows on affiliation status (whether the transacting parties are in a direct investment relationship). The other dimension is service type; this dimension includes 103 service types for each of

[15] Raking also goes by other names, including iterative proportional fitting.

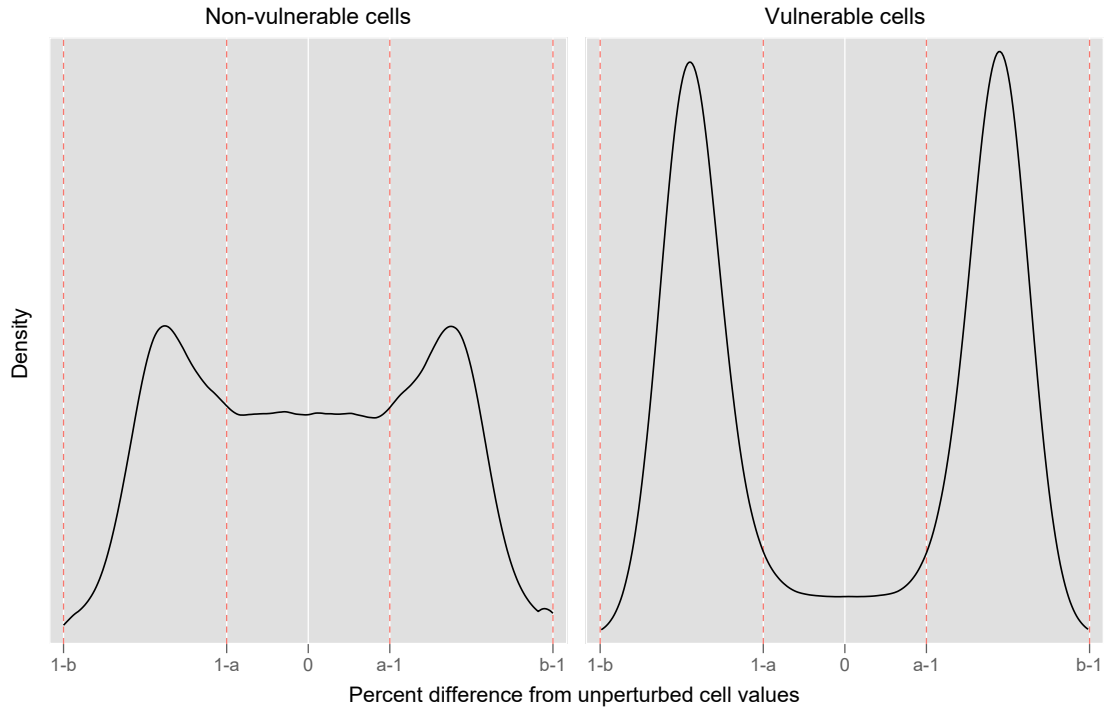
[16] Above we noted that using a bounded noise factor distribution results in bounded relative distortions for cells with only non-negative contributions. Boundedness in relative distortions is not guaranteed the total distortions that include the raking adjustments.

[17] Distortions in this section are typically calculated from a simulation exercise. In each replication, each record is assigned a noise factor under the selected EZS approach. Distorted cell values are computed and compared with the corresponding undistorted cell values to calculate a relative distortion for each cell. Distributional characteristics are calculated, either at the cell level or a more aggregated level, using relative distortions from each of the replications. In certain cases, cell-level standard deviations of relative distortions are directly calculated or approximated without reliance on simulations. Such direct calculation is not feasible for distortions in raked tables.

exports and imports (and, residually, for the balance of services trade).[18] Estimates for approximately 25 of the service types are based on source data other than BEA surveys and do not require protection. For this table, over 15,000 cells would be constructed in part or in full using noise-infused records.[19] Totals and other higher-level aggregates in International Services table 2.2 are consistent with the sum of quarterly estimates in the ITAs and the sum of monthly estimates in the trade in goods and services release.[20]

Figure 2 provides an example of the distributions of relative distortions that might arise with noise infusion.

Figure 2. Relative Distortions under Noise Infusion



[18] It also includes 12 supplemental columns related to insurance services. The measures in these columns partly underlie the insurance services measured in the main table columns.

[19] By value, BEA surveys provide the source data for approximately 55 percent of exports and 40 percent of imports. The shares are much higher (74 percent and 53 percent, respectively, in 2021) during the COVID pandemic because of the collapse of international travel and air passenger transport—service types whose statistics are not derived from BEA surveys. In calculating these shares, cells that combine service types are excluded from the calculation.

[20] The most detailed breakdown of quarterly services statistics in the ITAs is a cross classification including 13 service type columns (including a "total" row) for each of exports and imports and 35 geographic rows.

Figure 2 is based on a specific distributional approach to assigning noise factors and on specific parameters used within that approach. Different approaches and/or parameters would yield different distributions of distortions, as would the use of a different publication table. To avoid giving too much specific information about the choices we are considering, the distributions are shown with respect to the *a* and *b* parameters from the split distribution utilized for noise infusion (which is not necessarily the split *triangular* distribution shown in figure 1).

Figure 2 presents distributions for EZS noise infusion *without* constraining any aggregates to their undistorted values. The figure covers both distortions in exports and imports cells, but it, along with the other results shown in this section, excludes distortions from trade balance cells. As previously noted, the trade balance cells include negative components (because imports enter with a minus sign), and the interpretation of relative distortions is more complicated in such cases.

Two panels are included in figure 2, one for vulnerable cells, which we define as those cells that would be flagged for primary suppression under BEA's current cell suppression methodology, and one for the rest of the table cells ("non-vulnerable" cells). Approximately four to six percent of cells in International Services table 2.2 are vulnerable under this definition.[21] The two panels in figure 2 suggest that noise infusion achieves some degree of success in meeting its twin goals. For vulnerable cells, it provides relatively large distortions, which reflects protection of the most vulnerable (i.e., largest) contributions to those cells. For non-vulnerable cells, distortions are generally much smaller, suggesting that published values would be close to the underlying undistorted estimates.

Figure 3 shows corresponding distortion distributions for the case in which several aggregates are reset to their undistorted values after the application of noise infusion and the remaining non-vulnerable cells are raked to maintain additivity in the table. This process does not impact the distribution of noise infusions in the vulnerable cells, but it does impact the distribution of values for non-vulnerable cells. Specifically, raking increases the mass in the neighborhood of zero for non-vulnerable cells but also creates more diffusion around the edges of the distribution.

[21] The remainder of suppressed cells in International Services table 2.2—approximately six to eight percent of all cells in the table—represent complementary suppressions: cells that are themselves not vulnerable but that are suppressed to prevent attackers unwinding the value of primary suppressions by subtraction. Cells that would be complementary suppressions are included in the non-vulnerable cells category in figure 2 and figure 3.

Figure 3. Relative Distortions under Noise Infusion and Raking

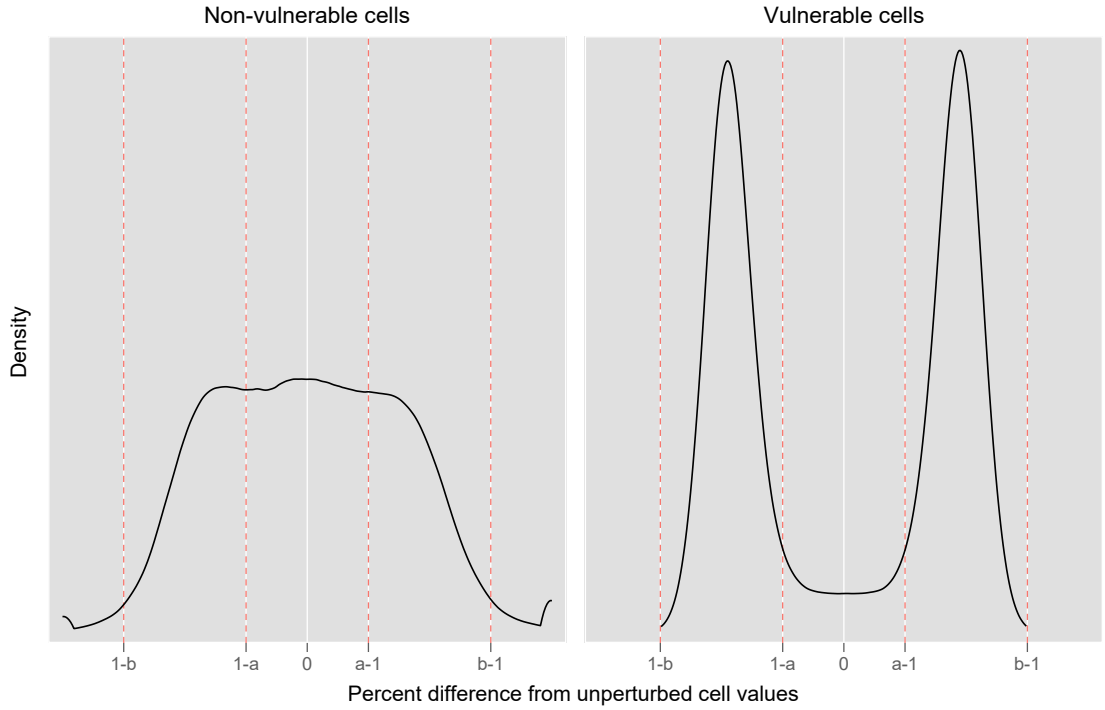


Table 3 uses a different criterion than vulnerability for classifying cells. Namely, it classifies cells according to their place in the hierarchy of International Services table 2.2. For instance, at the top level of both row and column hierarchies (level 1,1) are estimates of total exports of services and total imports of services. The next highest level in the geographic hierarchy includes estimates at the level of broad world regions (e.g., "Europe" or "Asia and Pacific"), while the next highest level in the service type hierarchy includes estimates for the 11 main service types (e.g., "telecommunications, computer, and information services" or "charges for the use of intellectual property not included elsewhere").

Average standard deviations of distortions are presented in table 3 instead of the full distribution of distortions, and results are presented for cells on both an unweighted and a weighted-by-dollar-value basis.[22] Results are shown separately for cases where key aggregates are and are not permitted to be distorted. Unlike with figures 2 and 3, numerical estimates are provided for distortions. However, these numerical estimates are only suggestive; they do not necessarily precisely reflect distortions that would

[22] To calculate average standard deviations, standard deviations are calculated for each cell in International Services table 2.2, and those standard deviations are then averaged over the cells in the relevant cell group.

be generated if BEA were to adopt EZS noise infusion, primarily because BEA may use different noise infusion parameters or distributions than underlie the estimates in table 3.

Table 3. Average Relative Standard Deviation by Cell Type and Noise Infusion Process

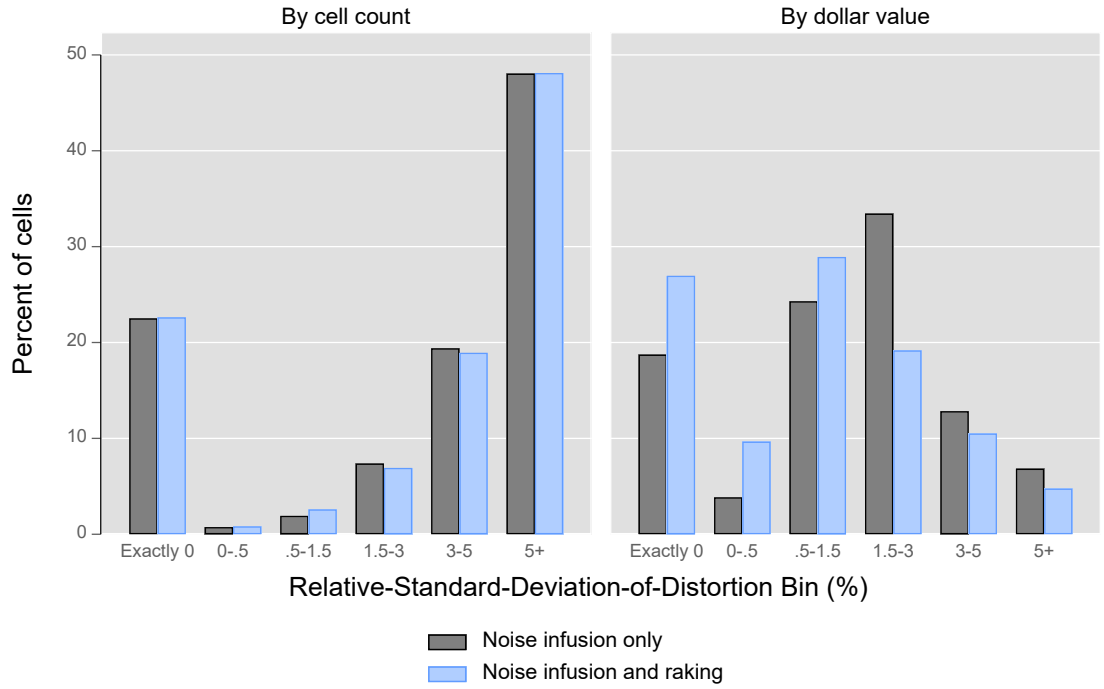
Cell level	Representative cell	Number of cells	Total value (billion USD)	Noise infusion		Noise infusion and raking	
				Unweighted average SD	Weighted average SD	Unweighted average SD	Weighted average SD
[1,1]	[Total services; all countries]	2	1,345	0.6%	0.7%	0.0%	0.0%
[1,2]	[Total services; Europe]	26	3,842	1.0%	1.0%	0.7%	0.5%
[2,1]	[Other business services; all countries]	22	1,345	1.6%	1.2%	0.0%	0.0%
[2,2]	[Personal, cultural, and recreational services; Latin America and Other Western Hemisphere]	262	3,840	2.6%	1.7%	2.4%	1.0%
[1,all other]	[Total services; South and Central America	160	1,529	1.7%	1.3%	1.5%	1.2%
[2,all other]	[Telecommunications, computer, and information services; Colombia]	1,718	1,529	3.5%	2.1%	3.8%	1.9%
[all other, 1]	[Computer services; all countries]	179	2,864	3.1%	1.9%	2.8%	1.3%
[all other, 2]	[Cloud computing and data storage services; Europe]	2,023	9,044	4.3%	2.4%	4.4%	1.9%
[all other, all other]	[Heritage and recreational services; Portugal]	11,772	3,141	5.5%	3.2%	5.7%	3.1%

(Unlike figures 2 and 3, table 3 includes results for table cells that are not derived from BEA survey data and, therefore, not subject to noise infusion; figure 4 and tables 4 and 5 below also includes results from cells not derived from survey data. As shown in figure 4, these cells account for slightly more than 20 percent of cells in International Services table 2.2. In terms of the cell levels shown in table 3, every level except [1,1], [1,2], and [1,all other] includes some cells that are not at least partly derived from BEA survey data.)

Table 3 makes clear that even if the higher-level aggregates are permitted to be distorted, their distortions are likely to be quite small in magnitude. This is good news regardless of whether higher aggregate distortions are permitted. If they are permitted, estimates in those cells will still be quite close to the original estimates. If they are not permitted, the raking performed to ensure additivity of the lower-level cells won't greatly change those smaller cells.

Figure 4 illustrates the same type of information in another way. It shows the share of table cells that, on an unweighted or a weighted basis, might fall into various standard deviation ranges.

Figure 4. Percent of Cells by Relative Standard-Deviation-of-Distortion Range



Tables 4 and 5 break out the distortion range results by cell types. These two tables only show distortions due solely to noise infusion; results from the combination of noise infusion and raking are omitted.

Table 4. Cell Count by Cell Type and Relative Standard Deviation Bin

Cell level	SD group 0	SD group 1	SD group 2	SD group 3	SD group 4	SD group 5
[1,1]	2	0	0	0	0	0
[1,2]	0	3	20	3	0	0
[2,1]	22	0	0	0	0	0
[2,2]	60	13	18	79	44	48
[1,all other]	0	12	75	59	12	2
[2,all other]	506	62	56	220	360	514
[all other,1]	40	0	6	49	46	38
[all other,2]	380	3	26	317	564	733
[all other,all other]	2,647	26	105	458	2,106	6,430

Table 5. Cell Value (million USD) by Cell Type and Standard Deviation Bin

Cell level	SD group 1	SD group 2	SD group 3	SD group 4	SD group 5
[1,1]	1,345,298	0	0	0	0
[1,2]	0	492,747	3,058,765	290,082	0
[2,1]	1,345,297	0	0	0	0
[2,2]	826,798	160,838	320,668	2,197,898	267,560
[1,all other]	0	152,280	884,414	359,103	130,655
[2,all other]	422,094	106,984	157,557	388,412	313,550
[all other,1]	698,929	0	220,375	1,359,851	414,106
[all other,2]	2,152,141	95,904	365,129	3,986,072	1,563,758
[all other,all other]	888,780	4,462	217,006	418,410	928,270

Communication with data users

One extremely important aspect of a potential adoption of EZS noise infusion at BEA is communication, particularly helping users understand how noise infusion works and how noise infusion will affect the usability of the data. Closely related to this is the question of how much information BEA can provide users about how much noise is in the statistics without compromising confidentiality. This background report, along with the FESAC presentation it accompanies, is meant to be a first step in communicating with data users in these areas. It is also meant to elicit suggestions about how best to communicate with other users—spanning the range from casual, occasional users to sophisticated power users—on these

topics, both at the time of a potential switch to noise infusion and on an ongoing basis if BEA implements noise infusion.

On the matter of communicating the benefits of the change to users, we want to get a sense of the relative importance to users of being provided "exact" estimates of published cells with no estimates for suppressed cells versus being provided reasonable estimates (with some variation in the degree of precision depending on cell type) for all cells. We would like to know which, if any, cells users view as untouchable in the sense that any distortion in such cells would dramatically degrade their usability.

On the matter of how much information can be provided to users about distortions, it is clear that providing a precise numerical value for each distortion would allow users to immediately unwind the undistorted value, thereby removing any protection meant to be inferred by noise infusion. To avoid partial unwinding of protection, we are also wary of providing an approximate magnitude of each distortion. Providing an exact description of the noise infusion process (distribution(s) and parameters) can also, in certain cases, assist attackers attempting to recover undistorted values of vulnerable cells, particularly if such a description accompanies other information about distortions or distributions of distortions.[23]

One relatively safe approach may be to provide users with information on the relative standard-deviation-of-distortions range into which a particular cell falls. For instance, the five standard deviation ranges from figure 4 and tables 4 and 5 could be used, either with cell labeling in the published table itself or a separate table of distortion standard deviations. For the cells shown in table 2 of this report, these options might yield information such as shown in tables 6 and 7.[24]

[23] The danger that arises from making public every aspect of disclosure avoidance is a feature EZS noise infusion shares with cell suppression. Under cell suppression, BEA informs data users that the p-percent rule is the basis used for suppression, but it does not reveal the threshold used with that rule.

[24] The standard deviations in these tables, like the cell values themselves, are purely hypothetical. Note that the standard deviations do not correspond to the distortions implied by the difference in cell values between tables 1 and 2. In this hypothetical set-up, each distortion implied by tables 1 and 2 represents only a single realization drawn from the universe of potential distortions.

Table 6. U.S. Exports of Services by Country and Service Type

	Total services	Computer services				Information services		
		Total	Computer software	Data storage services	Other computer services	Total	News agency services	Database services
All countries	19,519	11,660	4,236	2,674	4,750	7,859	3,405	4,454
Europe	11,003	6,577	2,860	896	2,821	4,426	2,009	2,417
France	3,364	1,457	589	202	666	1,906	888	1,018
Germany	2,257	2,002	925	176	901	255	124	131
Italy	3,076	2,055	952	208	895	1,021	187	834
United Kingdom	2,307	1,064	394	310	359	1,244	811	433
Africa	8,516	5,083	1,375	1,778	1,929	3,433	1,396	2,038
Morocco	2,468	1,414	366	725	323	1,054	251	803
Nigeria	3,276	1,443	50	573	819	1,833	868	965
South Africa	2,773	2,226	960	480	786	546	276	270

Key: relative standard deviations of distortions (percent)

0 to 0.5	0.5 to 1.5	1.5 to 3	3 to 5	5 or more

Table 7. Relative Percent Standard Deviations of Distortions in Export Estimates

	Total services	Computer services				Information services		
		Total	Computer software	Data storage services	Other computer services	Total	News agency services	Database services
All countries	0 to 0.5	0 to 0.5	0 to 0.5	0 to 0.5	0 to 0.5	0.5 to 1.5	0.5 to 1.5	0.5 to 1.5
Europe	0 to 0.5	0.5 to 1.5	0.5 to 1.5	0.5 to 1.5	0.5 to 1.5	0.5 to 1.5	1.5 to 3	0.5 to 1.5
France	0 to 0.5	0.5 to 1.5	5 or more	5 or more	1.5 to 3	0 to 0.5	1.5 to 3	1.5 to 3
Germany	0 to 0.5	0.5 to 1.5	0 to 0.5	1.5 to 3	1.5 to 3	0.5 to 1.5	1.5 to 3	3 to 5
Italy	0 to 0.5	0.5 to 1.5	5 or more	5 or more	3 to 5	0.5 to 1.5	3 to 5	1.5 to 3
United Kingdom	0 to 0.5	0 to 0.5	1.5 to 3	1.5 to 3	1.5 to 3	0.5 to 1.5	1.5 to 3	1.5 to 3
Africa	0.5 to 1.5	0.5 to 1.5	3 to 5	3 to 5	3 to 5	3 to 5	3 to 5	3 to 5
Morocco	1.5 to 3	3 to 5	5 or more	5 or more	5 or more	3 to 5	5 or more	5 or more
Nigeria	3 to 5	5 or more	5 or more	5 or more	5 or more	3 to 5	5 or more	5 or more
South Africa	1.5 to 3	3 to 5	5 or more	5 or more	5 or more	5 or more	5 or more	5 or more

We seek feedback on whether information such as this is too detailed, not detailed enough, or just about right in describing the variability in our trade in service statistics due to noise infusion, or if information of a different sort would be more useful to users of the statistics.

Questions for the Committee

This background report has outlined the reasons BEA is investigating a switch from cell suppression to EZS noise infusion in protecting survey data on trade in services, the methods under consideration for implementing such a change, the likely impacts of a change, and some potential means for communicating the resulting table-cell distortion to users of the statistics. Along with welcoming any general feedback on the topic of noise infusion or any specific feedback the on the content of the report, we raise the following questions for FESAC members:

- Is EZS noise infusion an improvement over cell suppression for this product?
    - Is there any additional information you'd need to confidently answer this question?
- What level of distortion is too much?
    - How should we weigh the tradeoff between exactness in major aggregates and additional distortion in lower-level cells?
- How much insight should be provided to users into the noise infusion process?
    - Is it important that they know the distribution(s) used to draw noise factors?
    - Is it important that they be informed about distributional bounds?
- How much information should be provided to users on the variability of cell distortions?
    - Would flags on individual cells be useful?
    - Would more aggregate information suffice (e.g., overall distributions)?
    - Something else?
- Are there any other questions users might like to see answered?
    - Are there any key considerations we haven't addressed?

References

Evans, Timothy, Laura Zayatz, John Slanta (1996), "Using Noise for Disclosure Limitation of Establishment Tabular Data. *In Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 65-86. (noise-data.pdf (census.gov))

Evans, Timothy, Laura Zayatz, John Slanta (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data," *Journal of Official Statistics*, 14(4), 537-551. (Using Noise for Disclosure Limitation of Establishment Tabular Data (scb.se))

Federal Committee on Statistical Methodology, "EZS-Noise Addition" In *Data Protection Toolkit: Protecting Data While Increasing Access*. Accessed April 28, 2023. (DPT | EZS-Noise Addition (ed.gov))

Martin, Jared (2013), "Evaluating Noise Infusion for Disclosure Protection for Two Time Periods," In *Proceedings of the 2013 Federal Committee on Statistical Methodology Research Conference*. (Evaluating Noise Infusion for Disclosure Protection for Two Time Periods (ed.gov))

Massell, Paul. B. and Jeremy M. Funk (2007), "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," in *Proceedings of the 2007 Federal Committee on Statistical Methodology Research Conference*. (Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection)

McKenna, Laura and Matthew Haubach (2019), *Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau*, Research and Methodology Directorate, U.S. Census Bureau, Washington, DC. (Research and Methodology Directorate (census.gov))