Variance Estimation for Product Sales in the 2017 Economic Census: Challenges in Implementing Multiple Imputation-Based Variance Estimation

Matthew Thompson[1], Katherine Jenny Thompson[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746

Abstract
The U.S. Census Bureau conducts an Economic Census every five years, producing key measures of American business and the economy. In addition to a core set of items collected from all establishments, the Economic Census requests information on the revenue obtained from products sold from all large businesses and a probability sample of smaller businesses. The 2017 Economic Census will – for the first time – publish variance estimates for product sales estimates.

A research team was established to recommend a variance estimation method that accounted for variance both due to sampling and imputation. The team's evaluative approach relied on simulation, using empirical data from a purposively selected, small subset of industries as the basis for the study. The research was complicated by the nature of product data, which are characterized by poor item response rates, few available predictors, additivity-within-establishment requirements, and many rarely reported products in an industry. The research team considered several alternative variance estimators on this limited number of industries and a subset of reported products, ultimately recommending a multiple imputation method that utilizes the Finite Population Bayesian Bootstrap (FPBB) to address the sampling variance and the Approximate Bayesian Bootstrap (ABB) to incorporate variance due to imputation.

The implementation of this proposed approach unveiled a number of modifications and enhancements needed to accommodate the complete set of variables. This paper describes the recommended variance estimation method and how this method is being implemented into the 2017 Economic Census production system. Examples are provided to illustrate implementation issues and the modifications and enhancements needed to fully implement the research-based recommendations.

1. Introduction

Improvements to official statistics programs may require complicated changes to existing methods or procedures to address new – emerging – requirements or to accommodate new

[1] Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

requests. Such changes are always constrained. Budget constraints could force an overall reduction in sample size; methodologists would need to revise the sampling design while maintaining predetermined reliability levels on key statistics. Project sponsors might commission the collection of additional data items, request preliminary tabulations (and publication) of survey estimates, or desire subdomain estimates not considered in the initial design. Again, these applications would be constrained by reliability and confidentiality mandates. In any case, the appropriate solution is rarely obvious; research is required. Of course, time and resource constraints can be prohibitive, so the research problem may be simplified to allow for a transparent and repeatable solution, thus delaying the unaddressed details or unforeseen nuances until the implementation stage, leaving little time or resources for additional research.

Efficiency frequently dictates the research and implementation processes. In the Economic Directorate of the U.S. Census Bureau, it is a common practice to establish a "dedicated" research team with a fixed duration comprising representatives from the relevant job series with differing experience levels, perhaps utilizing matrix management. Team responsibilities encompass defining and scoping the research problem, obtaining data, designing and conducting the research, writing and testing programs needed to carry out the research, documenting the findings, and presenting the "data-driven" recommendation. Assuming that the recommendation is endorsed, a subsequent implementation team is established. This team's composition can differ greatly from the research team, as expert staff are required – and production programmers must be included in the discussions – although some overlap in membership between the research and implementation teams is desirable. As with the research team, the implementation team usually operates under a fixed deadline. Team responsibilities include writing specifications that implement the research recommendation while addressing the issues that were "ignored" in the research. Logistical issues such as coding, testing, and validation are likewise included. An education component is not unusual, as the implementation team members may be unfamiliar with the methods under consideration.

The 2017 Economic Census leadership team endorsed a number of innovative updates, each introducing a new set of methodological and production implementation challenges. Historically, the Economic Census was a paper (mail out/mail back) collection; in 2017, collection will be primarily via the Web. In prior censuses, the collection unit for detailed breakdowns of dollar-valued totals varied by sector (percentage of total, $1000s, or both). In 2017, all detailed breakdowns of dollar-valued totals are collected in $1000s, as are the associated totals. Standard unit response rates will be released for the first time with the 2017 Economic Census, as will imputation rates for key statistics. Variability estimates for selected sample-based statistics will be published for the first time as well – of specific note for this paper, variance estimates for product sales.

Beginning in 2017, the Economic Census will use the North American Product Classification System (NAPCS) to produce economy-wide product tabulations. The Economic Census collects a core set of data items from each establishment called general statistics items: examples include total receipts or value of shipments, annual payroll, and number of employees in the first quarter. In addition, the Economic Census collects information on the revenue obtained from product sales. The introduction of NAPCS marks

a major departure from the prior collections which explicitly linked product codes to industry, allowing for different missing-data treatments for products by sector. Implementing a NAPCS-based collection necessitated the development of a single imputation approach for all Economic Census products to allow production of cross-sector tabulations.

This paper describes the research process used to determine the variance estimation method for product sales estimates and the process for implementing the research recommendations into the 2017 Economic Census production systems. Research and implementation were accomplished by two different teams, with a small fraction of membership overlap. Together, both teams developed and implemented the methodology that will be used in the 2017 Economic Census production systems.

Section 2 provides general background on the Economic Census along with more detailed background on product collection and estimation. Section 3 summarizes the research approach and resultant recommendations. In Section 4, we discuss the implementation of the recommended methods into a production system, specifically focusing on the unaddressed or unforeseen – but important – details that were excluded from the research study. We conclude in Section 5 with a few general observations about implementing research-based results in a production system.

2. Background on Economic Census and Product Data

The Economic Census is the U.S. Government's official five-year measure of American business and the economy. The term "Economic Census" is a bit of a misnomer as a sample of small single-unit establishments is surveyed in addition to all multi-unit establishments[2]. As mentioned in Section 1, the Economic Census collects a core set of data items from each establishment called general statistics items: examples include total receipts or value of shipments ("receipts"), annual payroll, and number of employees in the first quarter. In addition, the Economic Census collects data on the revenue obtained from products. All sectors construct a *complete universe of general statistics values* by using administrative data (or imputation) in place of respondent data for unsampled units. In contrast, sample weights are used to account for the unsampled single-unit establishments when producing product sales estimates. Finally, for each industry, the weighted sample estimates of product sales are further calibrated to ensure the sum of the product sales equals the total receipts (based on the complete universe) for the industry.

The 2017 Economic Census requests data on over 8,000 different products based on the North American Product Classification System (NAPCS); see https://www.census.gov/eos/www/napcs/more.html. However, evidence from prior economic censuses indicates that many products are rarely reported. Legitimate zero values are expected for many products in an industry, at both the individual establishment and total industry levels. Respondents can report data from a long, pre-specified list of potential

[2] A single-unit (SU) establishment owns or operates a business at a single physical location; whereas, multi-unit (MU) companies comprise two or more establishments that are owned or operated by the same company.

products in a given industry – some lists contain more than 50 potential products – and can write in descriptions of other products that were not pre-specified. Product lists can differ by industry within a sector. Furthermore, some product descriptions are quite detailed, and some products are mutually exclusive. Consequently, some establishments choose not to report any product data (complete product nonresponse). Those that do report often provide the same products typically reported by other responding establishments within a given industry (Fink, Beck, and Willimack 2015).

Several industries collect both "broad" and "detailed" products. Figure 1 presents a fictional illustration of broad and detailed products.

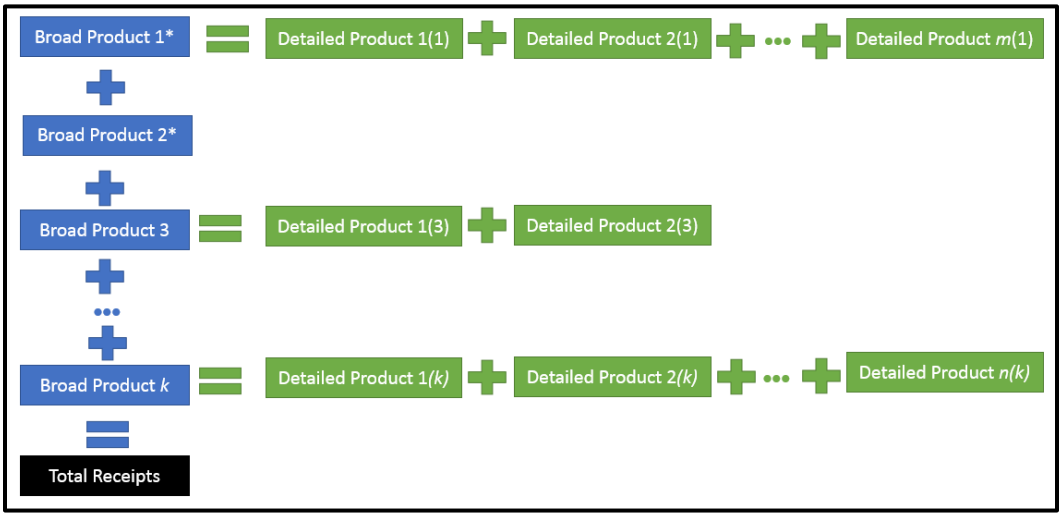

Figure 1: Illustration of nested balanced complexes defined by Broad and Detailed products in an industry. Broad products 1 and 2 are "must products" that have to be reported by an establishment to be classified into a particular industry. Broad products 1 through *k* sum to the establishment's value of total receipts. Not all broad products request detailed product breakdowns on the questionnaire, and the number of requested details differs by item (broad product). Finally, detailed product values are expected to sum to the associated broad product value. The establishment will still be considered to have reported product values as long as Broad products 1 and 2 are reported; other broad products are considered optional.

Broad and detailed products comprise nested one-dimensional balance complexes. The broad product values within a given establishment are expected to sum to the total receipts value reported earlier in the questionnaire. Under NAPCS, the same broad products can be collected in different industries. Detailed product values are expected to sum to their associated broad product value. Under NAPCS, a particular detailed product is associated with only one broad product. Missingness tends to be higher with detailed products than broad products. Lastly, many industries have required "must products" for establishment classification. Thus, although the same product can potentially be produced in different industries, product reporting is intertwined with industry classification. Certainly, the product distributions will differ between industries, even when the same products are reported. And of course, the detailed products will differ by industry.

Recall from Section 1 that the implementation of a NAPCS-based collection necessitated the development of a single imputation approach for all Economic Census products. Davie et al (2017) describes the processes used to determine and implement this "single" missing

data treatment procedure. The imputation research preceded the variance estimation research, and the imputation implementation was performed in parallel with the variance estimation research (mostly separate teams). The imputation research team recommended using hot deck imputation for broad products in all industries, allowing different hot deck variations by industries (random or nearest neighbor). Detailed products are imputed using reported data when available and ratio imputation otherwise. This is further discussed in Section 4.

Many sample surveys only estimate sampling variance. Of course, failing to address the additional variance due to nonresponse and imputation can lead to severe underestimation, unless the unit and item response rates are high or the imputation method(s) is extremely effective.

The sampling variance is generally quite low in the Economic Census, as the majority of establishments are selected with certainty; only the smallest single unit establishments in an industry and state are sampled. However, many products have quite high imputation. Consequently, ignoring this component in the variance estimation is likely to lead to erroneous inferences for a majority of the products. To avoid this phenomenon, the Economic Census program managers agreed that the variance estimation procedure should address sampling and nonsampling errors to the extent possible.

3. Variance Estimation Research for Broad Products

At the beginning of the research project, methods of estimating sampling and imputation variance were studied separately. Thompson, Thompson, and Kurec (2016) give an overview of the initial research into various methods to calculate variance estimates under the condition of complete response (i.e. with no imputation needed), while Thompson and Thompson (2016) discuss the results of the initial research into methods for estimating variance due to imputation. The results of these two branches of study are combined and the final variance estimation method presented in Knutson, Thompson, and Thompson (2017).

The research was undertaken by a commissioned team comprised primarily of methodologists. Previous teams – specifically the team tasked with developing the imputation methods for product lines – used a broader cross-section of team members including subject matter experts and classification experts. However, these subject matter experts were committed to other teams in preparation for the upcoming Economic Census, as were the classification experts. Fortunately, the variance estimation research was able to leverage the work of the earlier team. Specifically, the imputation research team developed test data to be used for analysis and the classification experts provided a list of industries whose product distributions were expected to remain largely the same under NAPCS. In addition, the variance estimation team leader had served as leader of the preceding imputation research team and was extremely knowledgeable about the difficulties faced and addressed in the imputation research – many of which were also concerns for the variance estimation team.

The research team was given 12 months to complete its work to allow time for implementation into the production system. During this time, the team had to learn about product data, develop the research methodology, write and test programs, conduct method evaluations, analyze the results, and ultimately make a final recommendation.

The team divided the project into the three evaluation studies listed in Table 1. The project started with team members familiarizing themselves with both Economic Census processing – particularly the collection of product sales data – and the many variance estimation approaches available in the literature. Research teams at the Census Bureau are often seen as opportunities for staff members to gain valuable experience in new and interesting topics, making this phase of the research team especially crucial as many of the team members had little practical experience with implementing variance estimation methodologies. After a period of reviewing and discussing the available methodologies and the practicalities of implementing each in the context of the Economic Census, the team selected three candidate bootstrap methods (two design-based, one Bayesian) for estimating sampling variance and a Bayesian bootstrap method and two model-based approaches for estimating variance due to imputation.

Table 1: Research Components

Component	Purpose
Knowledge Sharing	Bring staff "up to speed" on data collections
	Familiarize staff with the various variance estimation techniques available
	Select which methods to include in the evaluation study
Evaluation Studies	Evaluate the performance of the considered sampling variance estimation methods over repeated samples under the assumption of complete unit and item response
	Evaluate the performance of the considered imputation variance estimation methods over repeated samples in the presence of item nonresponse
	Evaluate the performance of the final proposed method

The initial phase of the evaluation was split into two parts. For the first evaluation, a simulation study was performed in which repeated samples were drawn and variance estimates calculated with complete unit and item response. In this way the team could evaluate the variance methods ability to accurately estimate sampling variance. For the second evaluation, a simulation study was used to evaluate the performance of variance estimation techniques designed to incorporate variance due to imputation. In this second evaluation, the entire simulated population was used and nonresponse was induced in order to evaluate the ability of the methods to accurately estimate variance due to imputation in the absence of sampling.

The test data was the same as used by the imputation research team and included only those industries provided by the classification experts. Note that the historical product data were

not expected to be perfect predictors of the 2017 product data because (1) some of the 2012 Economic Census product sales data were reported as percentages of total receipts; whereas, in 2017, all product data values will be collected in $1000s; and (2) in 2012 businesses could respond by paper and electronically, but in 2017 only electronic reporting will be used.

From the beginning, the team agreed to study only broad products (a decision consistent with the handling of product data by the imputation research team), as the rate of missing data for detailed products was quite high. And while industry-by-state level estimates were produced for each of the simulation studies, the data was ultimately so noisy at this level of detail that only national-level industry estimates were used in the evaluation.

The team decided on a simulation approach. In order to create industry "populations" from historical sample data in the 21 test industries, missing product values were filled in using nearest neighbor hot deck imputation for the sampled cases. Then, the SIMDAT algorithm was used to create completed records for the unsampled single-unit establishments in each industry (Thompson 2000). This nonparametric "nearest neighbors" simulation technique creates simulated data with the same correlation structure as the sample survey (training) data and similar quantile values. After establishing the complete population, 5,000 stratified SRS-WOR samples were independently selected in each industry, using the sampling parameters from the 2012 Economic Census for all sectors. Nonresponse was induced in each sample, and the identified variance estimation methods were applied to each. See Figure 2 for an illustration of the simulation study procedure.

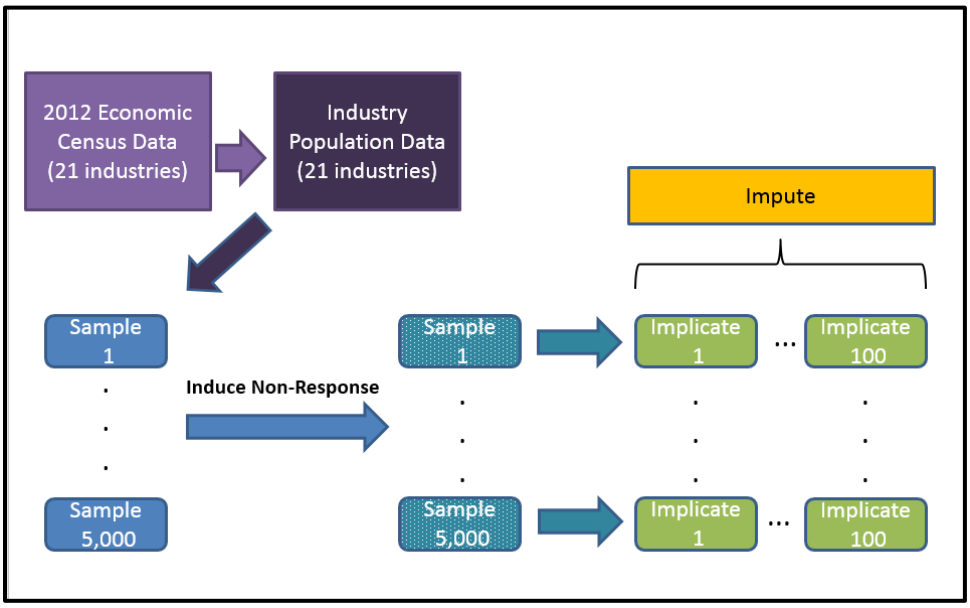

Figure 2: Illustration of the Simulation Procedure for the Variance Estimation Evaluation Research

Simulation was used to evaluate all of the considered variance estimation methods. By design and necessity, each simulation study made some simplifying assumptions beyond those previously mentioned. Small sample size effects were controlled by choice of

estimate level (national) and the selection of test industries. These choices sidestepped issues that would arise from small respondent sample sizes in imputation and estimation cells. The evaluation was restricted to the four best-reported broad products in each studied industry in terms of number of establishments that reported the product. Rescaling the size of the problem reduced computation time and increased available time for analysis, although it did impact the study's "representativeness." Lastly, the evaluation used rank-based tests within industry to compare the procedures, so that substantive improvements or deficiencies in specific situations were largely ignored. Instead, the evaluation procedures found common patterns among the methods on each evaluation criterion on the best-reported products.

Ultimately, the team recommended a multiple imputation method that utilizes the Finite Population Bayesian Bootstrap (FPBB) to address the sampling variance and the Approximate Bayesian Bootstrap (ABB) to incorporate variance due to imputation (Knutson, Thompson, and Thompson 2017). The Finite Population Bayesian Bootstrap, takes a single sample and creates from it, B "pseudo"-populations (B=5 was recommended). Differences across FPBB populations capture variability due to sampling. Within each FPBB population, establishments identified as donors are resampled to create ABB replicates, thus introducing variability to the donor pool prior to imputation. This resampling is repeated C times (C=20 was recommended) and recipients are imputed using each of the C sets of donors. In total there will be B*x*C replicates.

While this multiple imputation (MI) method of variance estimation will produce estimates of variance at any level – both in terms of estimation cells and broad vs detailed products – it should be noted that the research team did not evaluate the quality of the variance estimates produced for detailed products. The recommended method worked well for the frequently reported products in terms of confidence interval coverage. However, it did not exhibit comparable performance for the less frequently reported broad products due to a lack of responses and the high instance of reported zeroes and imputed values. This is almost certainly the case with most detailed products as well.

4. Implementation Team

A new team was established for implementation. Leadership was provided by project management experts with extensive familiarity with the subject matter and with the planned Economic Census processes. The team consisted primarily of experts in post-survey processing and tables creation as well as production programmers. Production methodologists were recruited to develop the final specifications, and two methodologists from the research team were retained as consultants.

As with the research team, the project began slowly with an educational component. Team members each had their own set of implementation issues that needed to be addressed. The production programmers were concerned about the computing resource demands of fully implementing the recommended variance estimation methodology as it requires imputation be run 100 times over. There were also staffing concerns as team members had to juggle working on this project with other high priority projects.

One of the primary concerns about the FPBB-ABB variance estimation method was the time it would take to run this process to multiply-impute missing products for the entire Economic Census. To determine this, as well as test more scenarios than the research team was able to evaluate, the team made use of the test deck created by the earlier imputation implementation team. In order to test out the hot-deck imputation methods for the 2017 Economic Census, this (previous) team created a full size test deck with roughly 2.4 million donors (with over 20 million products) and 1.1 million full recipients covering all NAICS sectors in-scope to the Economic Census. Using a concordance mapping 2012 product codes to 2017 NAPCS codes, the 2012 Economic Census product data was converted to a 2017 NAPCS basis.

Based on this test deck, initial performance testing indicated that it would take about 110 minutes to fully process (create and impute) each replicate for the entire Economic Census (implying a run time of over a week for all 100 replicates). Server upgrades and program improvements have since significantly improved run times from early tests. Average replicate run time is now approximately 55 minutes. However, the production methodologists noted several potentially problematic results during their review of the test run output.

4.1 Estimate Calibration

The first unexpected issue that the implementation team ran into was in determining an appropriate way to incorporate the ratio adjustment made to product estimates into the variance estimates using FPBB-ABB. Recall from Section 2 that, for each industry, the weighted sample estimates of product sales are calibrated to ensure the sum of the product sales equals the total receipts for the industry.

The FPBB populations are created by drawing $(N_h - n_h)$ units from stratum h from the original sample, where N_h is the population size of stratum h and n_h is the sample size of stratum h. The probability for the k[th] selection is given by

$$p_{h,k} = \frac{w_i - 1 + l_{i,k-1}\frac{(N_h - n_h)}{n_h}}{N_h - n_h + (k_h - 1)\frac{(N_h - n_h)}{n_h}}$$

For the purposes of the research team study, this ratio adjustment was incorporated into the creation of the FPBB populations by using the calibrated weights in the above formula instead of the sample weights as recommended by Zhou, Raghunathan, and Elliot, M.R. (2012). This worked well in the context of the handful of industries included in the research team's simulation study. However, with the much broader set of industries in the implementation test deck, this approach proved problematic.

The production methodologists identified several industries in which the ratio adjustment resulted in weights less than one, which in turn produced $p_{h,k} < 0$. The Economic Census uses the Business Register as a frame, which is largely built from administrative records sources. When the Economic Census forms are processed, establishments are often re-classified into different industries. This can lead to sizeable ratio adjustments in both

directions. Initially, the implementation team attempted to develop an "unbiased" algorithm to adjust the resampling probabilities. That proved to be very difficult.

After much discussion with the implementation team and separate meetings of the methodologists on the team, a different approach was suggested: use the original sampling weights for the resampling procedure, obtain the unadjusted multiple imputation estimates and variance estimates, then adjust these estimates by the post-stratification (calibration) factor. Of course, the proposed approach conditions on the sample-based estimate of product sales in the industry and state and is not unbiased.

However, incorporating the same ratio adjustment factor in both the estimate and the variance results in the ratio adjustment canceling out in the coefficient of variance calculation, effectively sidestepping this concern. Equally important, this approach simultaneously addressed concerns about using a multiple imputation variance estimator for a singly imputed estimate.

4.2 Proper ABB Implementation

The next obstacle the team faced was the appearance of estimation cells with positive (sometimes quite large) variance estimates where the production methodologists were expecting to see variances of zero or close to zero. An extreme example of this would be cells comprised entirely of certainty units with complete response to the product data. In this instance, one would expect variances of zero, as there is no sampling and no imputation taking place, but this is not what the team was seeing. This specific example is a rather rare circumstance (as you might expect), but it did raise questions about the quality of the variance estimates being produced.

The implementation team speculated that the ABB was being incorrectly implemented by both the implementation team and the research team before it. Multiple imputation holds the non-missing values (donors) fixed in all implicates; only the missing values are changed in the multiply imputed replicates (hereafter referred to as implicates). To properly perform the ABB with hot deck imputation, donors are resampled with replacement in each implicate, so that each implicate has a different imputation base. Resampling the donors perpetuates uncertainty in the imputation process without allowing estimation of additional variability due to nonresponse.

Both the research and implementation teams made an obvious mistake, using the resampled set of donors in place of the original donors in each ABB implicate. In an extreme example as mentioned above, this leads to implausible variance estimates. Using the proper ABB imputation procedure eliminated this error. However, it introduced a different issue. Recall that the ABB was performed independently within each FPBB implicate. While the FPBB implicates contain different numbers of resampled respondents and nonrespondents (thus simulating some nonresponse variability), there were only five FPBB implicates used, in contrast with the 20 ABB implicates within each FPBB implicate. Knutsen, Thompson, and Thompson reported that the usage of five FPBB implicates tended to overestimate the sampling variance, while the usage of twenty ABB implicates tended to underestimate the sampling variance. Implementing the proper ABB would further reduce the imputation

variance component. Was it possible that the “proper” imputation method would lead to severe underestimation with the low response rates that are typical of product data?

At this point the implementation team decided to take a brief break from meeting so that the methodologists could confirm the appropriate implementation of the ABB method when combined with the FPBB, peripherally investigating why the incorrect implementation was not obvious in the results of the previous research study. The methodologists reproduced the original simulation study performed to evaluate the FPBB-ABB method (Knutson, Thompson, and Thompson 2017), but on a much smaller subset of industries and only evaluating nearest neighbor hot deck imputation. The decision was made not to include random hot deck imputation in this evaluation because it adds more variability to the estimates and could thus obscure differences between estimates when evaluating the proper implementation of the ABB method. Whereas the previous research induced nonresponse at the observed rates from the 2012 Economic Census for each industry, for this follow-up round of research nonresponse was induced at various levels – from 40% to 90% – so that the team could make inferences relating response rate levels to the statistical performance of the considered variance estimators.

Table 2 shows the coverage rates of 90% confidence intervals for an industry within the Manufacturing sector for the two most frequently reported products. The columns labeled “Resampled” use variance estimates calculated incorrectly using the resampled donors. The columns labeled “Original” use variance estimates calculated using the original set of donors.

Table 2: Coverage of 90% Confidence Intervals

Response	Product 1		Product 2	
Rate	Resampled	Original	Resampled	Original
40	89.30%	76.70%	89.90%	78.50%
50	94.10%	77.10%	93.60%	76.60%
60	95.00%	72.50%	96.50%	73.80%
70	98.20%	69.50%	98.10%	69.40%
80	99.20%	65.70%	99.90%	69.70%
90	100.0%	61.70%	100.00%	65.10%

Product response rates are generally closer to 40% or 50%. In these cases, the improper imputation method actually improved the coverage – to nominal levels. In fact, looking at the results of the three industries included in this second round of research, it was not immediately obvious that there was a significant improvement using the original set of donors. The improper method overestimates the variance when response rates are high, leading to extreme over-coverage. On the other hand, product response rates are generally quite low, and the proper imputation method consistently produced under-coverage, i.e. variance estimates that were too small.

Xie and Meng (2017) provides a principled response to the dilemma, arguing that hot deck imputation models are always uncongenial (imputer’s model ≠ analyst’s model ≠ true

model), so that the resultant multiply imputed variance estimates are always underestimates. Using a mathematical tautology, they show that correct coverage is obtained by doubling the variances before calculating confidence interval estimates or p-values. Indeed, doubling the variance estimates in this application resulted in nearly nominal coverage using the proper imputation method with realistic product response rates, as illustrated in Table 3.

Table 3: Coverage of 90% Confidence Interval

Response	Product 1			Product 2		
Rate	Resampled	Original	2xOriginal	Resampled	Original	2xOriginal
40	89.30%	76.70%	89.80%	89.90%	78.50%	90.80%
50	94.10%	77.10%	90.50%	93.60%	76.60%	90.00%
60	95.00%	72.50%	87.40%	96.50%	73.80%	88.90%
70	98.20%	69.50%	84.70%	98.10%	69.40%	83.80%
80	99.20%	65.70%	82.60%	99.90%	69.70%	84.90%
90	100.0%	61.70%	79.70%	100.00%	65.10%	81.80%

This adjustment resulted in coverage rates much closer to 90% for all three reviewed industries – mitigating the under-coverage of using the original donors considerably.

4.3 Detailed Products

The team then shifted its attention to the estimation of variances for detailed product sales data. The research team had excluded detailed products from its evaluation in order to limit the scope of the research and for practical reasons, due to the extreme sparsity of data for many detailed products. However, the production methodologists on the implementation team pointed out that there were examples of detailed product data that were quite robust in terms of the amount of data collected (and the quantity of sales) and variances estimates would be needed for these products.

The same FPBB-ABB methodology used for estimating variances for broad products can easily be used for detailed products. However, the methodology used for imputing detailed product data created complications. When implementing the imputation methodology, the decision was made to utilize as much respondent data as possible in the imputation process. To this end, the imputation team decided to create “complete” donor records (i.e. fill in the missing detailed products) prior to hot deck imputation. Category averages were computed for each detailed product within a broad product for each potential imputation cell. Designated missing detailed products were imputed from their associated broad product total after matching the appropriate category averages. Once this process was finished and all donors were made “Complete,” the hot deck process was performed to impute products for all “Full” recipients.

Table 4 presents the establishment categorization used for product imputation. These categories maximize the use of reported data in the hot deck imputation procedures, but complicate the variance estimation of detailed products due to the partial donor/recipient establishments.

Table 4: Establishment Classification for Imputation

Donors	Broad products usable
Complete	All broad and detailed products usable
Partial	All broad products usable and some detailed products usable
Minimal	All broad products usable; detailed products missing and required
Recipients	Missing products
Full	Need broad and detailed products
Partial	Need some (designated) detailed products
Minimal	Need all detailed products
Ineligible	All products usable, but not “typical”; excluded from donor pool

What this means for variance estimation is that there is a subset of establishments that are both recipients (in category-average imputation for detailed products) and donors (for hot-deck imputation of broad-products). So the question arises, how should these cases be handled when identifying donors and recipients when resampling donors for creating the ABB replicates?

Treating these partial and minimal donor/recipient establishments as donors would handle the variance estimation properly for broad products, but would erroneously identify the cases as donors for detailed products. This would have the effect of resampling some detail product recipients in creating the ABB replicates. Treating the partial and minimal donor/recipient establishments as recipients, however, would negate – in part – what the resampling of donors is attempting to accomplish in the ABB since many true donors would not be eligible for resampling in this scenario. Alternatively, the broad and detailed products could be processed separately with donors and recipients being correctly identified for each. But program running times are already a huge concern, as the variance estimation process takes approximately 1 – 2 weeks to run. Doubling that run time is simply infeasible from a production perspective.

In the end, the current imputation methodology is a composite imputation that does not lend itself to a smooth implementation of FPBB-ABB. After much discussion the team decided on the following methodology. All donors – complete, partial, and minimal – will be identified as such for ABB resampling. In addition, each of the five FPBB populations will be run through imputation in order to apply the category averages to “complete” the partial and minimal donors. When creating the variance estimates, the original set of donors will be used (per Section 4.2), but these “original” donors will be pulled from the completed FPBB populations mentioned above. For broad products, this will have no impact as broad product data was fully reported for complete, partial, and minimal donors. For detailed products, this will use some category-average imputed data – specifically, for partial and minimal donors – as if it was reported. A note will be added to the published data product to make clear that the variance estimation method used was not optimized for detailed products. The team also decided to include a calculation of the percentage of the detailed product estimate derived from category-average imputation. This information can then be used in developing cell suppression rules and implemented during the disclosure avoidance processing. If the percentage is high, it is likely that the variance estimates will be an underestimate since a large portion of recipient data was being treated as donors.

5. Conclusion

When developing a research plan that applies to an ongoing survey, finding balance is hard. On one hand, making the scenario as simple as possible reduces the probability of treatment effects (solutions) being confounded by factors such as sample size or random noise. On the other hand, oversimplification can lead to very impractical solutions. Of course, it is crucial to limit the scope so that the research can be timely enough to be relevant when completed. However, it should be acknowledged that compressing the scope can lead to hasty decisions later in the implementation process, when there is no time left for careful further investigation.

In the case of the variance estimation research and implementation discussed in this paper, a couple of key factors played into the success of the teams. A key contributor was the awareness of team members of other research projects that had preceded us. In particular, the team was able to leverage the work of the previous imputation research team in a big way. Utilizing the comprehensive test deck created by that earlier team saved a considerable amount of time and effort. Time that was then able to be used to not only address issues as they arose but to actively research questions raised by the team in a more large-scale format than would have otherwise been possible.

The simulation performed to confirm the correct implementation of the ABB methodology was also made possible in large part because of the overlap between the research and implementation teams and a willingness (and the available time) from team leadership to halt the implementation process to properly investigate the questions being raised by the review of the early test runs. Due to the presence of the research team representatives throughout the implementation process, the team was able to utilize the research team's simulation apparatus quickly and efficiently and run simulations with a clear understanding of the problems being faced in implementation.

The hardest tasks tend to come from implementation, where every "cut corner" in the research needs filling in, and not every situation is ideal (e.g. small sample sizes, fewer donors than recipients, limited predictors). There are real advantages in establishing (almost) separate research and implementation teams as discussed in this paper. Having two teams approach the same problem from different perspectives leads to innovative applications. Often, these teams provide practical opportunities for methodologists to learn about data and data collection and for subject matter experts to learn about alternative methodologies.

Of course, there are equally real disadvantages. The limited scope in research can lead to missed requirements, which can be revealed as unexpected results in implementation testing or in production. The situations described in Section 4 are all examples of this. Clearly the decision to not include detailed products in the research phase left the implementation team with a lot of work in determining the best way to handle these estimates. Limiting the scope to a handful of industries also excluded many smaller industries - industries that are more likely to exhibit characteristics like the all certainty, 100% response example discussed in Section 4.2. Delaying decisions until implementation

can preclude having sufficient time for careful investigation, and quick decisions are often made for convenience based on anecdotal justification, with no alternatives tested.

When the end-result is a theoretically solid and operationally viable system, this approach is a success. It certainly was in the case study presented in this paper. The two-phase team approach has since been used with equally successful results for other 2017 Economic Census applications such as determining and implementing an imputation method for product estimates (Thompson and Liu 2015) and for developing standard response rates. Certainly in these examples, the advantages outweighed the disadvantages, with workable solutions and buy-in as well as shared understanding of implemented methods. And of course, the imperfect solutions provide plenty of exciting research ideas and opportunities for the upcoming 2022 Economic Census.

References

Davie Jr, W., Dahl, S., and Thompson, K.J. (To appear). From Research to Implementation of Product Estimation in the 2017 Economic Census: Hard, Harder, Hardest. *Proceedings of the Government Statistical Section,* American Statistical Association.

Fink, E.B., Beck, J.L. and Willimack, D.K. (2015). Data-Driven Decision Making and the Design of Economic Census Data Collection Instruments. *Proceedings of the FCSM Research Conference.*

Knutson, J. and Martin, J. (2015). Evaluation of Alternative Imputation Methods for Economic Census Products: The Cook-Off. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Knutson, J., Thompson, M., and Thompson, K.J. (2017). Developing Variance Estimates for Products in the Economic Census. *Proceedings of the Government Statistics Section*, American Statistical Association.

Thompson, J. R. (2000). Simulation: A modeler's approach. New York: Wiley.

Thompson, K.J. and Liu, X. (2015). On Recommending a Single Imputation Method for Economic Census Products. *Proceedings of the Government Statistics Section,* American Statistical Association.

Thompson, K.J. and Thompson, M. (2016). Estimating the Variance Due to Hot Deck imputation for Product Value Estimates in the 2017 Economic Census. *Proceedings of the Government Statistics Section,* American Statistical Association.

Thompson, M., Thompson, K.J., and Kurec, R. (2016). Variance Estimation for Product Value Estimates in the 2017 Economic Census Under the Assumption of Complete Response. *Proceedings of the Government Statistics Section*, American Statistical Association.

Xie, X. and Meng, X.L. (2017). Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's, and Analyst's Models are Uncongenial? Statistica Sinica 27(4): 1485-1544. doi:10.5705/ss.2014.067

Zhou, H., Raghunathan, T.E., and Elliot, M.R. (2012). A Semi-Parametric Approach to Account for Complex Designs in Multiple Imputation. Proceedings of the FCSM Research Conference.