Exploratory Data Analysis of Economic Census Products: Methods and Results

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

Abstract

Beginning in 2017, the U.S. Census Bureau will begin using the North American Product Classification System (NAPCS) in the Economic Census to produce economy-wide product tabulations. This marks a major departure from the current data collection method that explicitly links products to industry. Motivated by this collection change, the U.S. Census Bureau conducted a study to investigate methods of treating missing product data in the Economic Census, with the goal of recommending a single imputation method to produce product data in all trade areas that is statistically defensible and operationally practical. The validity of an imputation method is highly dependent on the nature of both the reported data and on the nature of missing data (e.g., factors that contribute to response). This paper presents an exploratory data analysis of empirical data from selected industries with common products under NAPCS at the national industry level that explores these factors, describing the methods and presenting the results. The collective results were used to recommend candidate methods, to develop imputation cells, and to inform the subsequent evaluation study by providing realistic response propensity models.

Key Words: Economic Census, Missing Data, Exploratory Data Analysis, Response Propensity Models

1. Introduction

The U.S. Census Bureau conducts the Economic Census every five years in years ending in two or seven. Although the Economic Census is a single program, we process the Economic Census sectors, comprising industries, in eight separate trade area databases: Construction (CON), Finance, Insurance, and Real Estate (FIR), Manufacturing (MAN), Mining (MIN), Services Industries (SER), Retail Trade (RET), Transportation, Communication, and Utilities (UTL), and Wholesale Trade (WHO). Each trade area collects a core set of data items from each establishment called general statistics items or basic data. Examples include total receipts or value of shipments, annual payroll, and the number of employees in the first quarter. In addition, we collect information on the revenue obtained from sales of industry-specific products. All establishments of multi-establishment companies are asked to report product data. In the MAN and MIN trade areas, single establishment companies (called "single-unit establishments") that are larger than a predetermined size-threshold are asked to provide product data. In WHO trade area,

[1] Contact yukiko.tomabechi.ellis@census.gov with questions about this report. This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

product information is requested from all single-unit establishments (i.e., a complete census). Otherwise, a probability sample of single-unit establishments is used. [Note: The CON trade area is a probability sample of all establishments]. Prior to the 2017 Economic Census, respondents reported their product data on one of more than 400 industry-specific versions of the questionnaires (paper and electronic), with over 8,000 different products reported.

The methods of treating missing product data in the 2012 Economic Census and prior censuses vary greatly by trade area. In the MAN and MIN trade areas, we collect product values directly, and compare each establishment's aggregated products to the final edited and imputed value of receipts. We publish the difference between the sum of the reported products and the total receipts as "Not Specified by Kind (NSK)," making no attempt to impute missing product values. We imputed the CON industry data using a hot deck nearest neighbor donor imputation procedure. In the other trade areas, we rake the individual establishment product distributions to equal the total receipts value if the imbalance is within a trade-area specific tolerance level; otherwise, we classify the establishment's product data as unusable. For estimation, all usable records are ratio-adjusted so that the aggregated product data equals the value of total receipts within an industry by geography estimation cell. We refer to this procedure in-house as "expansion."

The following table summarizes the methods currently used to impute product data:

Trade Area	Products are imputed at the micro level for:	Method used to impute products *(in order applied)*
Manufacturing	N/A	• Allocated to NSK
Mining	Single product industries	• Total receipts allocated to the single product
Construction	Entire sample	• Nearest neighbor donor imputation • Industry average ratio *(if no acceptable donor)*
Other	Selected sample cases	• Raking to 100% (1) • Other sources (annual surveys, SEC filings, etc.)

(1) If the reported product values for an establishment do not sum to total receipts but are "close" then we rake the reported product values to equal total receipts. If the sum of the reported product values is not close to the corresponding value of total receipts or product values are not reported at all, we do not use the product data from the establishment in calculating products estimates.

Beginning with the 2017 Economic Census[2], data collection will be electronic and respondents will have more flexibility in reporting products. Moreover, NAPCS allows the collection of the same product in different industries, which will be summarized in cross-sector tabulation of products. The plan for the change in the data collection led to the creation of an interdivisional team in March 2014, tasked to determine a single

[2] Starting with the 2017 collection, the Economic Census will be published as the Census of U.S. Businesses.

imputation method for all products by the fall of 2014. The team comprised classification experts, subject-matter specialists, and methodologists. The subject matter and classification experts developed the test data used for all analyses and provided expertise on the current 2012 Economic Census procedures. The methodologists' familiarity with the subject matter and expertise on the current procedures ranged from completely novice to extremely knowledgeable about a selected subset of trade area procedures.

The first order of business for the team was to learn about the collection and the processing of product data in each trade area. Once the team was briefed, the second order of business was to study the properties of the product data in various trade areas so that viable imputation options could be considered. For this, the team conducted a series of exploratory data analyses (EDA) on empirical data from the 2012 Economic Census (excluding the CON trade area). These analyses were designed to study the characteristics of product data that vary greatly by industry (including candidate predictors) and to examine the factors that contribute to product response. We used the collective results from the first phase to recommend candidate imputation methods, to develop imputation cells, and to aid the design of the subsequent simulation study in Phase 2 by providing realistic response propensity models.

Prior to the EDA, the team had the following four candidate imputation methods in mind: ratio (expansion) imputation (EXP), hot deck nearest neighbor (HDN) and hot deck random (HDR) imputation, and Sequential Regression Multivariate Imputation (SRMI). The EDA provided needed information to validate the usage of these imputation methods and to implement imputation *models* within each imputation method. The logistic regression models for product nonresponse provided fitted response propensities necessary for the simulation study in the second phase of the study. Section 2 describes the collection of product data. Section 3 describes the test data. Section 4.1 describes investigation of the characteristics of reported product data while Section 4.2 describes exploring the response mechanisms for reporting product data. Section 5 presents the summary of the first phase of the product imputation study.

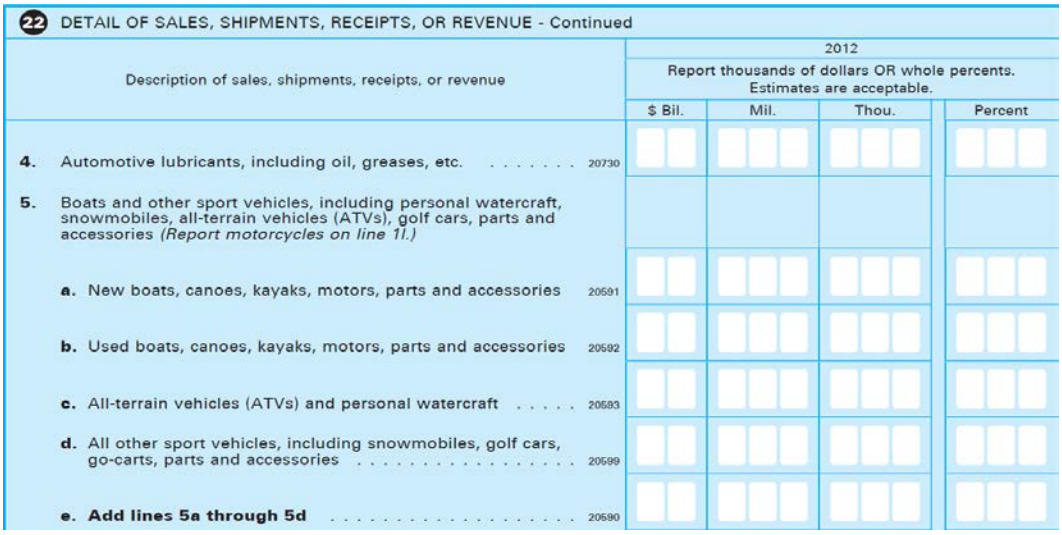
2. Product Data Collection

The Economic Census (EC) attempts to collect a total value for sales, shipments, receipts, or revenue from all sampled establishments. Figure 1 provides an example of this data item collection for a retail trade establishment.

Figure 1: Sample Economic Census Collection Instrument for Total Sales for a Retail Trade Establishment (Item 5)

Product data (labeled as "Details of sales, shipments, receipts, or revenue") are collected towards the end of the questionnaire in item 22. The types of products that an establishment is expected to produce or to sell are strongly related to the primary industry in which the establishment operates. In the 2012 EC, product information in item 22 contains a complete list of likely products for the industry. In the MAN and MIN trade areas, product data are reported as total values in dollars. The reported product dollar values in item 22 are expected to sum to the total receipts reported earlier in item 5 in the

questionnaire. In the other trade areas, each product is reported as a percentage of the establishment's total receipts, but many of the forms allow the establishment to provide either percentage distributions or total values in dollars. Figure 2 provides an example from the product collection for establishments located in the "Automobile Dealers" retail trade industry.



22 DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE - Continued

Description of sales, shipments, receipts, or revenue	2012 Report thousands of dollars OR whole percents. Estimates are acceptable. $ Bil.	Mil.	Thou.	Percent
4. Automotive lubricants, including oil, greases, etc. 20730				
5. Boats and other sport vehicles, including personal watercraft, snowmobiles, all-terrain vehicles (ATVs), golf cars, parts and accessories *(Report motorcycles on line 1l.)*				
a. New boats, canoes, kayaks, motors, parts and accessories 20591				
b. Used boats, canoes, kayaks, motors, parts and accessories 20592				
c. All-terrain vehicles (ATVs) and personal watercraft 20593				
d. All other sport vehicles, including snowmobiles, golf cars, go-carts, parts and accessories 20599				
e. **Add lines 5a through 5d** 20590				

Figure 2: Extract from 2012 Economic Census Data Collection Instrument for Retail Trade establishments in the Automobile Dealer Industry

Starting in 2017, we will only collect product value data in thousands of dollars in all trade areas and there will be no collection differences across trade areas.

Within an industry, product data are largely characterized by very low item response rates for all but the most frequently reported products. In some industries, one or more products are required – these are called "must-have" products. In fact, lack of a "must-have" product would call into question the industry classification of an establishment (an automobile dealer without revenue from the sales of automobiles, for example). Other industries do not have any "must have" products.

Missing product data can occur when an establishment does not respond to the census (unit nonresponse). Among unit respondents, it occurs when an establishment provides no product information or when an establishment provides product information that does not sum to its total receipts within the specified raking tolerance. Unit non-respondents were out-of-scope for this project. The remaining two types of establishments with missing product data will be called "product non-respondents" henceforth in this paper.

3. Test Data

For the exploratory data analyses and response propensity analyses, 2012 EC product data were used from seven trade areas: FIR, MAN, MIN, RET, SER, UTL, and WHO. Because there was no direct translation of its current product classification to NAPCS construction products at the time of the study, the CON trade area was excluded from these analyses. All data had undergone post-collection editing and imputation. In all trade areas, classification experts on the team selected ten to thirty industries per trade area with common products under the projected NAPCS structure.

The test data files contain all product level records from establishments in the EC product tabulation universe within the selected industries. However, we only used records from establishments that are full year reporters, had positive total receipt values (reported or imputed), and were used for product estimation. Products are edited and imputed after procedures for the general statistics items are completed so that all establishments have a valid value of total receipts along with a valid industry code.

Here, we define product respondents (donors) as establishments that provided "usable" products. These include establishments that provided products in perfect balance, i.e., the sum of reported product values equals the total receipts, and establishments with a product imbalance within an acceptable level (raking tolerance that varies by trade area) that was correctable. The detailed definition of product respondents can be found in the final team report (Thompson, 2014).

Table 1 shows the magnitude of the product non-response problem for five of the eight trade areas in the test data used for exploratory data analyses. The third column of the table shows the number of unit respondents that did not provide any product data, that is, a subset of the product non-respondents. The proportion of these units ranges from 15.2 percent for RET to 25.9 percent for FIR. The percentage distribution at the industry level is likely to be even more variable than the one at the trade area level shown here, possibly resulting in a dearth of donors in some of the imputation cells.

Table 1: Number and Percent of Unit Respondents that did not provide any product data, at Trade Area Level, based on Test Data

Trade Area	Unit Respondents that reported at least one product	Unit Respondents that did not report any product	Total Unit Respondents
FIR	44.212 (74.1%)	15,464 (25.9%)	59,676
RET	180,771 (84.8%)	32,361 (15.2%)	213,132
SER	108,360 (78.4%)	29,846 (21.6%)	138,206
UTL	9,718 (76.9%)	2,926 (23.1%)	12,644
WHO	38,896 (77.5%)	11,277 (22.5%)	50,173
Total	381,957 (80.6%)	91,874 (19.4%)	473,831

4. Exploratory Data Analysis

The following data analyses had two goals:

1. To understand the nature of reported product data to inform the four potential imputation methods: ratio (expansion) imputation, nearest neighbor hot deck, random hot deck, and sequential regression multivariate imputation; and
2. To understand the nature of missing product data to assess existing imputation cells, suggest refinements, and to provide response propensities necessary for later simulation.

The following sections describe these analyses and summarize the results. Hereafter, we use the term *imputation cell* to describe subdomains in which imputation is performed.

We define the imputation cells as follows (Note: NAICS is the North American Industry Classification System code, consisting of 6-digit Industry code and 2-digit US specific code):

- NAICS for FIR, MAN, MIN, RET, UTL
- NAICS and Type of Operation for WHO
- NAICS and Tax Exempt Status for SER

4.1 Exploring Reported Product Data

The set of EDA investigations described below provides information on the main characteristics of the usable products and provides some insights into strategies for imputation method implementation.

4.1.1 Distribution of Ratio of the Sum of Reported Products to Total Receipts (Establishment Level)

As described in Section 3, product respondents include establishments that provided products in perfect balance and establishments with product imbalance within an acceptable raking tolerance. This analysis provides the needed information to determine a reasonable raking tolerance for the MAN and MIN trade areas; this analysis was not necessary for the other trade areas, who had already established raking tolerances.

The analysis showed that nearly two thirds of all establishments are in balance with no discrepancy between the total receipts and sum of the product values in the studied trade areas. Given that over 80% of establishments have summed product data that are within 10-percent of the associated total receipts, we recommend setting raking tolerances at the slightly conservative value of 15%. With such a large base of in-balance or raked-into-balance cases, we should have a sufficiently large donor pool for imputation.

4.1.2 Number of Products by Trade Area

An establishment can potentially report several different products. If, however, the majority of establishments within an industry report the same products and if these few products account for a high percentage of the industry receipts, then the product distributions within industry may be easy to preserve via a simple imputation method or model.

Our next analysis examines the distribution of the number of products reported per establishment within each trade area. We first obtain the number of products reported by each establishment that reported at least one product. We then obtain the distribution (mean, median, mode, and range) of the product counts within each industry. For example, the industry-level statistics for MIN are as follows:

Industry	Number of Establishments	Mean	Median	Mode	Range
21311300	127	1.65	1	1	4
21111100	2501	2.01	2	2	12
21239100	15	1.27	1	1	2
21111200	233	4.64	5	6	8
21211300	27	1.19	1	1	1
21311200	3504	2.06	1	1	13

Finally, we obtain the mean and median of the industry-level summary statistics at the trade area level. Table 2 presents the selected industry-level summary statistics of the product counts at the trade area level.

Table 2: Mean and Median of the Selected Industry-Level Summary Statistics (Median and Range) on Establishment Product Count by Trade Area

Trade Area	Number of Industries	Mean of the Medians	Median of the Medians	Mean of the Ranges	Median of the Ranges
FIR	8	2.25	1.50	6.25	6.50
MAN	15	2.40	2.00	12.53	11.00
MIN	6	1.83	1.00	6.67	6.00
RET	24	6.04	2.00	15.79	13.50
SER	23	1.43	1.00	9.00	7.00
UTL	11	1.00	1.00	4.82	5.00
WHO	58	1.25	1.00	7.69	6.50

The mean of the medians was below three for all studied trade areas except RET. Hence, half of the establishments in the study typically reported less than three products. However, there are establishments in each trade area that reported a much larger number of products than the trade-area averages. Because such cases appear in all trade areas, they cannot be easily discounted and should be considered in the imputation models.

4.1.3 "Importance" of Products by Trade Area

Given that the median of the medians was between one and two for all trade areas, we wanted to determine if establishments in the same industry generally reported the same products. To study this, we examined four unweighted proportions within each industry:

- Sum of all reported product values/unit respondent total receipts
- Sum of the top one product values/unit respondent total receipts
- Sum of the top one and two product values/unit respondent total receipts
- Sum of the top one, two, and three product values/unit respondent total receipts

(Note that the denominator includes product respondents as well as product non-respondents.)

After producing industry-level proportions, we obtained the trade area average for each of these proportions as shown in Table 3. Note that these analyses do not use the sample weights. As mentioned in Section 3, the selected test industries are not representative of the EC. Following Phipps and Toth (2012), our "population of inference" for this exploratory analysis is restricted entirely to establishments in the studied industries and in particular, to unit respondents. The analyses presented in Section 4.2 are weighted, as the models used should represent their EC populations.

In all trade areas, the top three products contributed more than 50 percent of the total reported product data. In the FIR, SER, and WHO trade areas, the summed product data across all products were about 20 percent below the total receipts for the industry, demonstrating that there is a non-negligible amount of product non-response. Thus, the recommended imputation procedure needs to capture a variety of products to create realistic distributions.

Table 3: Proportion Contributed to Industry Totals by Top Reported Products (Averaged Over Multiple Industries)

Trade Area	Number of Industries	All Products	Top 1 Product	Top 2 Products	Top 3 Products
FIR	8	0.76	0.37	0.56	0.62
MAN	15	0.97	0.48	0.57	0.64
MIN	6	0.98	0.58	0.80	0.84
RET	24	0.93	0.54	0.64	0.69
SER	23	0.77	0.35	0.49	0.57
UTL	11	0.91	0.74	0.82	0.85
WHO	58	0.81	0.58	0.67	0.72

4.1.4. Predictors of Product Data

First, we explore whether size of unit is a good predictor of number of reported products, specifically checking to see if the larger establishments are likely to report more products. To examine this, we grouped the establishments within each imputation cell into three equal sized groups (in terms of number of establishments) based on their annual payroll value and computed the mean number of products within each of these size categories. Table 4 summarizes these results by trade area. We see that the mean number of products increase *slightly* as the establishment size increases in all but the UTL trade area.

Table 4: Summary of the Mean Number of Products by Trade Area and Establishment Size

	Establishment Size Based on Annual Payroll		
Trade Area	Small	Medium	Large
FIR	1.56	2.44	3.13
MAN	1.67	2.64	2.92
MIN	1.83	1.83	2.00
RET	5.52	6.34	6.48
SER	1.26	1.57	1.83
UTL	1.00	1.00	1.00
WHO	1.11	1.34	1.62

The second analysis examines the correlation between total receipts (RCPTOT) and each product value (PRODUCT) by industry. The key assumption for the (ratio) expansion estimator is that RCPTOT is a strong positive linear predictor of each product value. If this model is true, then the correlation between each PRODUCT and RCPTOT should be positive and near 1. Obviously, it would be unrealistic to assume that this assumption has to hold for every single product in an industry, but it should be true for the majority of frequently reported products.

We note that the majority of the Top 3 products are “must-have” products in our test data. The correlation between each of these Top 3 products and RCPTOT is very high, mostly over 0.90. However, as pointed out in Section 4.1.3, there is a non-negligible amount of product non-response not accounted for by the Top 3 products. Hence, we compute sample correlations for each product within each imputation cell. We excluded products that were reported by four or fewer establishments within an imputation cell from the

correlation analyses. However, it is useful to see how often this is the case. Table 5 presents the mean and median industry-level product correlations by trade area.

Table 5: Mean and Median Industry-Level Correlations between Product Value and RCPTOT, by Trade Area

Trade Area	Number of Estab	# of Product	Mean	Median
FIR	1-4	17		
	5-9	10	0.56	0.62
	10-14	5	0.69	0.84
	15+	95	0.66	0.68
	Total	127		
MAN	1-4	499		
	5-9	118	0.53	0.70
	10-14	54	0.47	0.53
	15+	137	0.60	0.63
	Total	808		
MIN	1-4	32		
	5-9	9	0.65	0.75
	10-14	5	0.72	0.99
	15+	51	0.80	0.93
	Total	97		
RET	1-4	179		
	5-9	55	0.45	0.58
	10-14	32	0.51	0.63
	15+	490	0.54	0.58
	Total	756		
SER	1-4	130		
	5-9	35	0.65	0.82
	10-14	24	0.66	0.73
	15+	170	0.64	0.66
	Total	359		
UTL	1-4	50		
	5-9	21	0.69	0.83
	10-14	11	0.60	0.67
	15+	61	0.76	0.84
	Total	143		
WHO	1-4	917		
	5-9	230	0.62	0.77
	10-14	96	0.64	0.72
	15+	527	0.67	0.72
	Total	1770		

The table indicates that the fewer the total number of products reported in a trade area the higher the correlations, as evident in the MIN (97) and UTL (143) trade areas. However, for the remaining trade areas, this table does not provide strong evidence that the product data satisfy the assumption for the ratio model used in the expansion method. It is also important to notice the high incidence of products reported by four or fewer establishments, especially in the WHO and MAN trade areas.

Note that the mean is lower than the median in all cases, indicating that there could be some very low correlations. Figure 3 graphs the distribution of product value correlations with RCPTOT by trade area.

On this boxplot, we hope to see small boxes with short tails, all centered above 0.70. We do not. Given the spread of the correlations and the existence of negative correlations, we have evidence against the use of the ratio (expansion) estimator for product data.

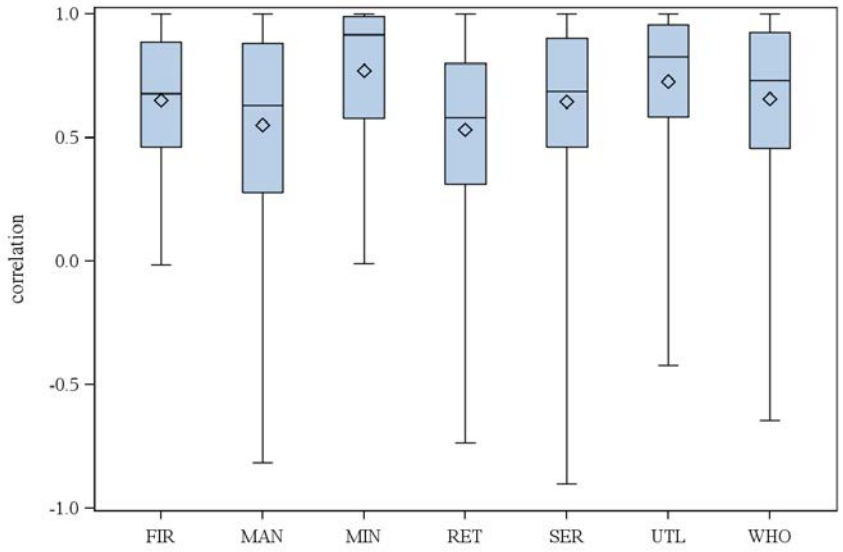


Figure 3: Distribution of Correlation of Product Values with RCPTOT by Trade Area. (Note: The upper and the lower ends of the whiskers represent the maximum and the minimum values. The mean is indicated by a diamond and the median by the horizontal bar in the rectangle.)

4.2 Exploring the Response Mechanisms for Reporting Product Data

The purpose of these analyses was to find covariates that are related to establishment reporting of usable product data. We developed two sets of response propensity models by trade area using logistic regression. We used the results of the study to assess and refine the existing imputation cells in each trade area. They are also helpful in implementing studied methods such as hot deck imputation, by providing sort variables within an imputation cell. We also used the models in our simulation study.

These analyses used the full-year reporter (active for 10 or more months in 2012) establishments that provided a non-zero value for total receipts in the same test industries as the EDA analyses described in Section 2.1, but they have different criteria for inclusion:

- We used the same criteria of 'reported data' given in the EC response rate report (Lineback, 2014).
- We retained selected not-specified-by-kind (NSK) cases in the MAN and MIN trade areas' data sets.
- We used the sampling weight in the logistic regression analyses. This effectively excluded establishments with zero sample weight from the analyses.

We considered the following covariates in the models:

Acronym	Variable Name	Levels	Description
IMPCELL	Imputation Cell		See introduction in Section 4
GEOREG	Census Region	4	Four groupings of states
LFO	Legal Form of Organization	2	Corporation or not
SURVUTYP	Unit Type	2	Multi-unit versus Single-unit
NAIC_CHANGE	Industry Change after Mailout	2	Yes/No
NON-NORM	Non-normal Unit Type	4	Births, deaths, seasonal, all others

4.2.1. Methods and Analysis Statistics

We define a *unit respondent* as an establishment whose value of total receipts is classified as "equivalent to reported data." A product respondent is a unit respondent that reported "usable" product data.

In each trade area, we partition the establishments in the test industries into four non-overlapping groups:

Group 1: Unit non-respondents
Group 2: Unit respondents that reported no usable Product data
Group 3: Unit respondents that reported exactly one usable Product item
Group 4: Unit respondents that reported more than one usable Product items

We do not compare Group 1 to other groups (Groups 2-4). This would be equivalent to examining unit response propensity, which is out-of-scope for this project. Instead, we focus on non-response with respect to usable product data among unit respondents. In Analysis 1, we compare unit respondents that reported no usable product data (Group 2) with unit respondents that reported at least one usable product data (Groups 3 and 4) to find a set of explanatory variables that are predictive of providing *any* respondent product data. In Analysis 2, we compare units in Group 3 against units in Group 4 to examine if there is any intrinsic difference between the two groups that might help us define imputation cells better. Separate logistic regression models were fit for each trade area.

4.2.2. Predicting Probability of Providing Any Usable Product Data (Analysis 1)

This analysis fits logistic regression models to find covariates that significantly contribute to the probability that a unit respondent will provide usable product data.

Let

$Y_{kj} = \begin{cases} 1 \text{ if the establishment j in imputation cell k provided any usable data} \\ 0 \text{ otherwise} \end{cases}$

$\boldsymbol{X}_{kj}^{w} = (x_{kj1}, x_{kj2}, \ldots, x_{kjw})$ denote the vector of w potential explanatory covariates from establishment j in imputation cell k.

We fit logistic regression models to estimate the conditional probability that an establishment reports usable product data, i.e., $\Pr\left(Y_{kj} = 1 \middle| \boldsymbol{X}_{kj}^{w}\right) = \pi(\boldsymbol{X}_{kj}^{w})$ for each candidate set of covariates w. Categorical covariates are represented as a collection of design (dummy) variables in $\boldsymbol{X}_{kj}^{w}$.

The logit of each multiple logistic regression model has the following form:

$$g\left(\boldsymbol{X}_{kj}^{w}\right) = ln\left(\frac{\pi(\boldsymbol{X}_{kj}^{w})}{1-\pi(\boldsymbol{X}_{kj}^{w})}\right) = \beta_0 + \beta_1 x_{kj1} + \cdots + \beta_w x_{kjw} = \beta^w \boldsymbol{X}_{kj}^{w},$$

where $\pi\left(\boldsymbol{X}_{\boldsymbol{kj}}^{\boldsymbol{w}}\right) = \frac{\exp\left(g\left(X_{\boldsymbol{kj}}^{\boldsymbol{w}}\right)\right)}{1+\exp\left(g\left(X_{\boldsymbol{kj}}^{\boldsymbol{w}}\right)\right)}$.

Since many trade areas in the EC employ probability samples, models are fit using PROC SURVEYLOGISTIC (SAS® Online Documentation) with strata defined by NAICS (industry). Unsampled cases are not included in this analysis, as their sampling weight is zero. The SURVEYLOGISTIC procedure uses the method of maximum likelihood to fit survey parameters, but incorporates complex survey sample designs features such as

stratification and unequal probability sampling in the test statistics using the methods outlined in Roberts, Rao, and Kumar (1987) and Lehtonen and Pahkinen (1995).

We performed response propensity modeling by trade area using a forward selection procedure derived by Wang and Shin (2011). Each additional covariate must be statistically significant given those already in the model in the forward selection procedure. We use the likelihood-ratio test to measure overall goodness-of-fit for each candidate model, whose test statistic is

$$D = -2 * \ln[\text{(likelihood of the fitted model) / (likelihood of the saturated model)}].$$

Under the null hypothesis ($\beta X=0$), D has an approximate chi-squared distribution. Each variable in the forward selected model must be statistically significant at the significance level of 0.05 using the Wald statistic.

Ideally, we want to minimize the number of covariates. Furthermore, any considered categorical variable must have a sufficient number of respondents per imputation cell for consideration. In addition to considering the goodness-of-fit test results described above, we examine the Rescaled R^2 from Tjur (2009). We calculate the mean predicted probability of an event for each of the two categories of the dependent variable and take the difference between those two means. Like the "traditional" R^2 used in linear regression, the upper bound is 1.0 and the interpretation is analogous.

4.2.3. Predicting the Probability of Reporting More than One Product (Analysis 2)

This analysis fits logistic regression models to assess whether there are intrinsic differences between unit respondents that report exactly one usable product and unit respondents that report two or more usable product items.

The analysis implemented the same forward stepwise procedure as Analysis 1. The definition of the dependent variable changed as follows:

Let

$$Y_{kj} = \begin{cases} 1 \text{ if unit respondent establishment } j \text{ in industry } k \text{ provided at least two products} \\ 0 \text{ if unit respondent establishment } j \text{ in industry } k \text{ provided exactly one product} \end{cases}$$

Initially, we planned to include the top two covariates from Analysis 1 into the model in Analysis 2 as the main effects and then examine the contribution of the remaining factors to the probability of reporting more than one product. However, we decided to conduct the forward selection procedure independently from Analysis 1 for two reasons. First, when we included the top two covariates from Analysis 1 in the initial model, the effect of at least one of these covariates was often not statistically significant in Analysis 2. This indicated that the mechanism that differentiated product respondents from product non-respondents might be different from the mechanism that differentiated respondents that reported two or more products from those that reported exactly one product. Second, the primary purpose of this analysis was to find covariates that further improved imputation cells. The more covariates included in the imputation cell definitions, the smaller the count of establishments in the defined cells. To respect the generally-accepted minimum cell size of ten establishments (Vittenghoff and McCullouch 2006), we decided to limit the number of covariates to two in Analysis 2.

4.2.4. Results and Recommendations

4.2.4.1. Analysis 1 Results

Table 6 presents the forward selection results for each trade area for the first model, showing covariates that are predictive of providing any usable product data. A dash (-) indicates that the covariate does not contribute significantly. An entry of 'n/a' indicates that the covariate is not applicable in a given trade area.

Table 6: Response Propensity Model Covariates (Analysis 1) Ordered by Descending Strength of Predictor

Predictor	MAN	MIN	FIR	RET	SER	UTL	WHO
RCPTOT	3	2	-	5	7	6	-
IMPCELL	1	1	1	3	1	1	2
GEOREG	-	3	6	7	5	3	6
LFO	-	-	2	1	4	4	1
SURVUTYP	2	4	4	2	6	-	3
NAICS_CHANG	n/a	n/a	5	4	2	2	4
NON_NORM	n/a	n/a	3	6	3	5	5
Rescaled R^2	.251	.066	.120	.267	.264	.362	.287

The results of this propensity analysis can be summarized as follows:

1. In all trade areas but RET and WHO, the currently defined imputation cell (IMPCELL) is the most significant contributor; in the WHO trade, the imputation cell is the second most significant contributor after LFO whereas in the RET trade, it is the third most significant contributor after LFO and SURVUTYP.

   The importance of IMPCELL in predicting the product response might be explained by the varying complexity of industry questionnaires. In some cases, item 22 is long but straightforward, requesting mutually exclusive products with short and clear descriptions. In others, the product definitions might be more detailed, products might not be seen as mutually exclusive, and in some cases, certain products may require additional sub-detail items with their own additively constraints. We speculate that the length, complexity, and burden of the item might affect nonresponse.

2. Except for MIN, total receipts is not a highly significant contributor, if at all. This provides further supporting evidence that size of unit does not seem to be related to providing product data.

3. Note that the secondary predictors are different by trade area. Hence, IF we have enough observations in imputation cells, we can tailor our response model to each trade area further by adding the secondary predictor.

4. The modest values of R^2 might indicate that a large portion of the explanation for product non-response may be due to idiosyncratic establishment effects.

4.2.4.2. Analysis 2 Results

Table 7 presents the forward selection results for each trade area for the second model, showing covariates that are predictive of providing more than one product, given that at least one product has been reported.

The top two most influential covariates are IMPCELL and SURVUTYP for all trade areas except RET. For RET, RCPTOT and IMPCELL are most influential. Recall that the RET imputation cells are defined entirely by NAICS. The significance of RCPTOT indicates that the probability of providing more than one product in a RET trade establishment is related to the size of the establishment. In other trade areas, the establishments tend to be more homogeneous in terms of size within imputation cell, especially after accounting for SURVUTYP.

Table 7: Response Propensity Model Covariates (Analysis 2) Ordered by Descending Strength of Predictor

Predictor	**MAN**	**MIN**	**FIR**	**RET**	**SER**	**UTL**	**WHO**
RCPTOT	4	-	-	1	-	-	7
IMPCELL	1	1	1	2	1	1	1
GEOREG	3	-	4	3	4	-	3
LFO	-	-	6	-	5	-	5
SURVUTYP	2	2	2	-	2	2	2
NAICS_CHANGE	n/a	n/a	3	4	3	3	6
NON_NORM	n/a	n/a	5	5	6	-	4
Rescaled R^2	.156	.236	.496	.613	.468	.037	.187

At a minimum, these results indicate that SURVUTYP should be included in imputation model implementation, either to refine imputation cells, sort variables in hot deck methods, or as predictors in regression models. However, the optimal imputation cells would vary by trade area, as indicated by the first response propensity analysis.

5. Conclusion

This paper presents a series of analyses designed to gain an understanding on the reporting nature of establishments that provide valid product data to the EC. It is the first such study conducted at the U.S. Census Bureau in this area. Prior to this study, much of the "information" on product data was anecdotal. Collective wisdom differed by trade area. For example, many subject matter experts contended that establishments tend to report a single product, whereas our analyses showed that the typical number of products differs by trade area and indeed, the type of establishment (single unit or multi-unit) is predictive of reporting multiple products. Some trade areas believed that a simple ratio adjustment was the only appropriate missing data adjustment method, whereas others believed that no adjustment was appropriate. These analyses demonstrated the inappropriateness of the model assumption for the former and provided some indication that other methods such as hot deck imputation or sequential multiple regression imputation might be preferable.

Of course, not all of these analyses were as enlightening. Some are omitted as inconclusive. Others were more confirmatory of earlier findings, providing little new

information. Regardless, these analyses served as a very useful forum for gaining knowledge. They created an opportunity for hands-on experience and for knowledge sharing – both among team members (as results were shared) and with stakeholders (to confirm findings). They helped the team understand the nuances of the imputed product data that could be easily satisfied by the appropriate selection of imputation cells: for example, using industry and type of operation are both highly related to the establishment's reported products, and type of establishment (single or multi unit) appears to be related to the number of products reported. They also helped the team understand some of the implementation challenges ahead, such as the need to impute sparsely-reported products and the lack of available prediction variables.

These analyses provided invaluable background to a novice team. They helped lay groundwork for model development. Lastly, they informed the next phase of the project in a simulation study to assess the statistical properties of the four imputation methods over repeated samples in each of the trade areas.

Acknowledgements

The authors thank Scot Dahl, John Kern, and Michael Kornbau for their careful review and thoughtful comments on earlier versions of the manuscript. We also thank the collective members of the Product Line Imputation Team (Laura Bechtel, William Davie Jr., Kaili Diamond, Fay Dorsett, Maria Garcia, John Kern, Jeremy Knutson, Xijian Liu, Jared Martin, Darcy Steeg Morris, Jonathan Schuyler, Robert Struble, Kevin Tolliver, John Ward, and Yves Thibadeau (consultant)) for their dedicated work on the project.

References

Lehtonen, R. and Pahkinen, E. 1995. Practical Methods for Design and Analysis of Complex Surveys. Chechester, UK: John Wiley & Sons.

Lineback, J.F. 2014. "Economic Census Response Rate Programming Specifications." The Census Bureau internal document, EDMS #205254, dated April 25, 2014.

Phipps, P. and Toth, D. 2012. "Analyzing Establishment Nonresponse Using and Interpretable Regresssion Tree Model With Linked Administrative Data." *The Annals of Applied Statistics*: 6(2): 772-793.

Roberts, G., Rao, J.N.K., and Kumar, S. 1987. "Logistic Regression Nalaysis of Sample Survey Data." *Biometrika* 74: pp. 1-12.

SAS Online Documentation: SAS/STAT 92. User's Guide, Second Edition, Overview: SURVEYLOGISTIC Procedure. Institute for Digital Research and Education, University of California at Los Angeles. www.ats.ucla.edu/stat/sas/sasdoc.htm (Last access on November 3, 2014)

Thompson, Katherine J. 2014. "Recommendation for Product Line Imputation for 2017 Economic Census: Report from the Product Line Research Team." The Census Bureau internal document, dated 12/9/2014.

Tjur, T. 2009. "Coefficients of Determination In Logistic Regression Models – A New Proposal: The Coefficient Of Discrimination." *The American Statistician* 63: 366-372.

Vittenghoff, E. and McCulloch. 2006. "Relaxing the rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*: 165(6), pp. 710 – 718.

Wang, F. and Shin, H., 2011. "Model Selection Macros for Complex Survey Data Using PROC SURVEYLOGISTIC/SURVEYREG." *MWSUG Proceddings.*