# Evaluating Imputation and Estimation Procedures in a Survey of Wholesale Businesses

Martin Klein, Joanna Fane Lineback, Joseph L. Schafer*

**Abstract**

The Monthly Wholesale Trade Survey (MWTS) provides estimates of change in inventories and sales for wholesale businesses in the United States. In previous work, we developed a new procedure for multiply imputing values of sales and inventories for nonrespondents based on a multivariate linear mixed model. We conducted a simulation study, generating an artificial population with frame data and two years of monthly sales and inventory figures, and we observed how the current MWTS Horvitz-Thompson and random group estimators performed over repeated samples with and without missing data. We discovered that the complete-data point estimates, variance estimates and interval estimates did not behave as large-sample normal theory suggests they should. Over repeated samples, point estimates were highly skewed and interval estimates had poor confidence coverage. In this paper, we review our findings to date and describe ongoing research into new estimation procedures for Bayesian finite-population inference with stratified samples from highly skewed populations.

**Key Words:** Business survey, confidence interval coverage, estimation, longitudinal survey, modeling, simulation

## 1. Background

The Monthly Wholesale Trade Survey (MWTS), conducted by the U.S. Census Bureau, produces estimates of economic activity for merchant wholesalers in the United States (excluding manufacturing sales branches and offices). Estimates are released in the *Monthly Wholesale Trade Report*, published at `www.census.gov/wholesale/` six weeks after the close of each month. Results are tabulated by kind of business, corresponding to major categories of durable and nondurable goods defined by the North American Industry Classification System (NAICS). For each kind-of-business classification, three key estimates are reported: the percent change in sales relative to the previous month, the percent change in inventories relative to the previous month, and the inventories-to-sales ratio.

Several years ago, we began a research project to examine statistical methods used in the MWTS. At first, we focused on the imputation procedure used to fill in approximately 30% of the sales and inventories figures that are missing in any given month due to nonresponse. We developed an alternative model-based multiple-imputation procedure and compared its performance to the current method in a large simulation study. We discovered that, even apart from missing data, the current Horvitz-Thompson based estimation methodology has room for substantial improvement.

In this article, we review our findings to date and describe ongoing research to develop new methods for the MWTS. At this time, we are not proposing any specific changes to the

*U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233.

*Disclaimer:* This paper is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Authors' names are listed in alphabetical order.

survey. Rather, we envision possibilities for a new integrated methodology that leverages frame information and the power of statistical models to estimate total sales and inventories from the respondent data alone, bypassing the need for imputation entirely. In Section 2, we review key aspects of the current MWTS design and estimation procedures. In Sections 3 and 4, we describe our findings from our imputation research and simulation study. In Sections 5 and 6, we describe some possible alternatives to the current estimation techniques using model-based Bayesian finite-population inference.

## 2. Sample Design and Estimation

The MWTS is part of an elaborate data collection system for U.S. businesses, and its design is typical of many establishment surveys throughout the world (Smith, 2013). The system is built on the Business Register, a relational database continuously maintained by the Census Bureau with information for business entities at various levels of aggregation. Every five years, the Business Register is supplemented by data from the Economic Census. From these sources, a sample frame is generated. The largest companies in each industry class are selected with certainty, and smaller business units are sampled within strata defined by industry class and size. Approximately 8,700 units are selected to participate in the Annual Wholesale Trade Survey (AWTS), which collects detailed information each year. All certainty units and half of the noncertainty units from the AWTS are included in the MWTS, which has a total sample size of about 4,200.

Keeping the MWTS sample representative of its target population is challenging, because the economy is continually evolving. New businesses are created; others cease to exist; others change through mergers, acquisitions, and other types of restructuring; and new kinds of businesses emerge. Each quarter year, dead businesses are phased out of the MWTS and replaced by a sample of new businesses. Every five years, when new data from the Economic Census become available, the frame is recreated and new samples for the AWTS and MWTS are drawn.

Estimated sales and inventories within NAICS categories are based on the classical Horvitz-Thompson (HT) method. Each sampled unit's contribution to the estimated population total is weighted by the unit's inverse probability of selection into the sample. Standard errors are computed by the method of random groups (Wolter, 2007). Figures published in the *Monthly Wholesale Trade Report* are not strictly HT, however, because the data are heavily processed before and after the HT procedure. Many units fail to provide usable sales and/or inventories in a given month. On average, item nonresponse is approximately 30%, with inventories seeing higher rates of nonresponse than sales. These missing values are filled in by a ratio-imputation method described below. Moreover, the monthly estimates from the MWTS are revised to sum to corresponding total estimates from AWTS, a procedure known as benchmarking. Monthly estimates are also adjusted using the Census Bureau's X-13ARIMA-SEATS software, using a time series model that is periodically reviewed and updated.

## 3. Imputation Study Using 25 Months of MWTS Data

In current practice, missing values of sales and inventories in the MWTS are imputed separately but in a similar fashion by a ratio-based procedure. The sample units are classified into imputation cells defined by company size and industry class. Within an imputation cell, the imputed value for a unit's inventories (sales) in the current month equals that unit's inventories (sales) in the prior month, multiplied by the ratio of the weighted total current month's inventories (sales) to the weighted total previous month's inventories (sales). The

totals used to construct the ratio are taken over those units within the imputation cell having reported inventories (sales) in both the current and previous month.

Lineback and Schafer (2013) identified several concerns with this ratio imputation method: it implicitly assumes a linear model with zero intercept; the only predictor is the value from the previous month; some imputation cells contain very few cases; relationships between sales and inventories may not be preserved; uncertainty due to imputation is not taken into account; and it does not make use of other information available from the Economic Census and sampling frame.

Because of these concerns, we devised an alternative procedure to handle missing data and applied the current and new methods to 25 months of MWTS data from December 2008 to December 2010. The total sample size was $n = 4,468$. Our alternative involved multiple imputation under a multivariate linear mixed-effects model (Schafer and Yucel, 2002). This model was designed to preserve the non-normal marginal distributions of sales and inventories, to reflect multivariate relationships between sales and inventories within months, to allow flexible trends for inventories and sales within industry groups over time, and to make efficient use of information from the Economic Census. To handle the non-normal marginal distributions of sales and inventories, we transformed them to approximate normality using smoothed empirical distribution functions and the Gaussian-quantile function (Lineback and Schafer, 2013). We used four predictors from the 2007 Economic Census: receipts, payroll, employment, and operating expenses. Some units have missing values of the Economic Census predictors which poses an additional modeling challenge. We multiply imputed the missing predictors using a method of Gelman and Hill (2007, p. 541); we refer to Lineback and Schafer (2013) for more details. To model flexible trends of inventories and sales within industry groups, and within individual companies, we used natural cubic splines.

Suppose that $Y_{ijk}$ is the (transformed) response for unit $i$, during month $j$, to survey item $k$ (sales and inventories), $i = 1, \ldots, 4468$, $j = 1, \ldots, 25$, $k = 1, 2$. Let

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i11} & Y_{i12} \\ Y_{i21} & Y_{i22} \\ \vdots & \vdots \\ Y_{i,25,1} & Y_{i,25,2} \end{pmatrix}$$

denote the $25 \times 2$ dimensional matrix of responses for unit $i$. The model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where $\text{vec}(\boldsymbol{b}_i) \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$, $\text{vec}(\boldsymbol{\epsilon}_i) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I})$, independently over $i$. $\boldsymbol{X}_i$ is a matrix having 25 rows because there are 25 months of data, and $4 + 6 \times 19 = 118$ columns because there are four Economic Census variables, 6 basis functions used in the natural cubic splines, and 19 non-overlapping industry classes. $\boldsymbol{\beta}$ is a $118 \times 2$ dimensional matrix representing the fixed effects in the model. $\boldsymbol{Z}_i$ is a matrix having 25 rows (again because there are 25 months of data) and 6 columns because there are 6 basis functions used in the natural cubic splines. $\boldsymbol{b}_i$ is a $6 \times 2$ dimensional matrix representing the random effects in the model. The splines in the fixed-effects part preserve industry level trends, whereas the splines in the random-effects part preserve unit-level trends. Missing values are assumed to be missing at random (MAR), and diffuse Bayesian prior distributions were applied to the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. Multiple imputations were generated using a Markov Chain Monte Carlo (MCMC) approach (Schafer and Yucel, 2002).

Under the current ratio method, we imputed missing values as discussed above, and then computed Horvitz-Thompson estimates of total inventories (sales) and percent relative

change in inventories (sales) between a current and previous month, each within 22 industry classifications for each of 25 months. The standard error of each estimate was computed using the method of random groups. Under the new multiple imputation method, we created $m = 15$ imputed datasets. For each imputed dataset we computed Horvitz-Thompson estimates of total inventories (sales) and percent relative change in inventories (sales) between a current and previous month, within each of 22 industry classifications for each of 25 months, along with corresponding standard errors using the random group method. Finally we combined the results using Rubin's (1987) rules to get the final estimates and standard errors.

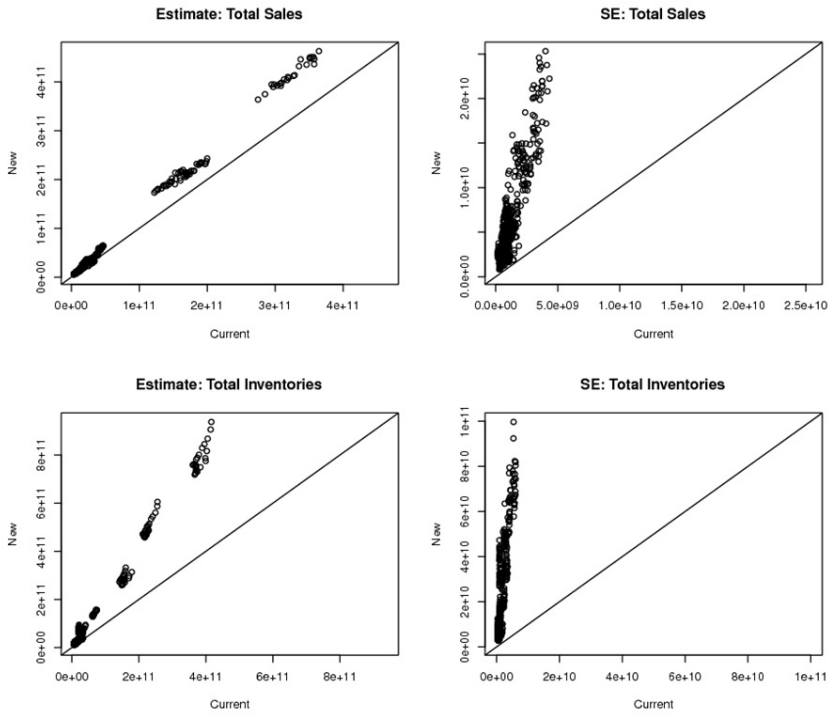


**Figure 1**: Total estimates and standard errors under current and proposed imputation

The scatter plots in Figure 1 compare the estimated total sales and inventories, and corresponding standard errors, obtained when imputing under the current ratio method ($x$-axis) versus the proposed multiple imputation based method ($y$-axis). The 45-degree line is also shown. Under the new method, the estimated totals are consistently larger, and the standard errors are much larger. Figure 2 shows results when the estimand is the percent relative change in inventories (sales) between a current month and previous month. We observe that under the new method the change estimates are smaller in magnitude and the standard errors are larger. The new multiple-imputation method yields larger standard errors because it accounts for uncertainty in predicting the missing values. The current ratio method imputes values along a straight line with no dispersion, ignoring the natural variability that would appear if the value had in fact been observed. The multiple-imputation method simulates draws from the posterior predictive distribution of the missing observations given the observed data, preserving the natural variability and reflecting uncertainty in model parameters.

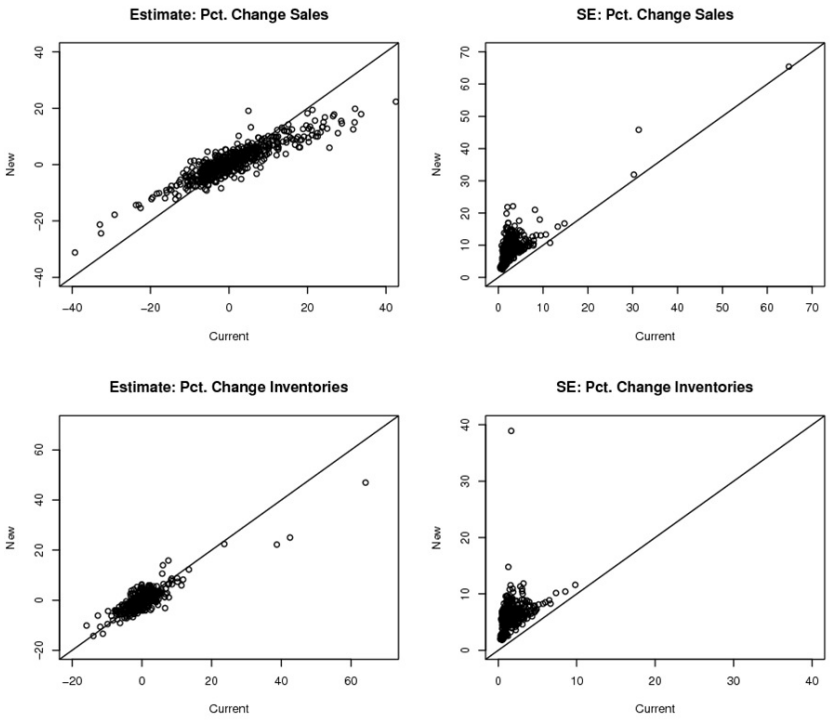


**Figure 2**: Change estimates and standard errors under current and proposed imputation

## 4. Simulation Study

Because it is impossible to draw firm conclusions based on one sample, we designed a simulation study to see how these imputation methods behave over repetitions of the MWTS sampling procedure (Klein, Lineback, and Schafer, 2014). The basic design of the simulation study was as follows.

1. Using data from the actual MWTS sampling frame, 2007 Economic Census, and 25 months of MWTS data (December 2008 to December 2010), construct an artificial population of values. The goal was to create an artificial population that could reasonably represent the target population of the MWTS during the December 2008 to December 2010 time frame.

2. Draw samples using a stratified sampling design that mimics the actual design of the MWTS.

3. For each unit in the sample, impose a pattern of missing values (a vector of length 50, containing 0's and 1's) of the MWTS sales and inventories variables over the 25 months using a hot deck within cells defined by industry class and size. The donors in each cell were the units actually in the MWTS.

4. Impute missing values.

5. Compute estimates of population totals and overall percent relative change (based on the Horvitz-Thompson estimator) and corresponding standard errors (based on random group variance estimator).

6. Repeat steps 2-5 a total of 1,000 times to study the properties of the estimators.

### 4.1 The Artificial Population

Details on the construction of the artificial population are given by Klein, Lineback, and Schafer (2014) To summarize, we used the actual MWTS sampling frame to obtain a list of approximately 286,000 units representing the population of merchant wholesale companies in the United States in the last quarter of 2008. The previous 2007 Economic Census contained variables for most of these units. Whenever available, we merged four variables from the 2007 Economic Census (number of employees, annual payroll, first quarter payroll, and revenue) into our population. After the merge, any of the four Economic Census variables that were missing in the population were imputed using a sequential-regression random-forest procedure, for details we refer to Klein, Lineback and Schafer (2014). Approximately 3,200 units in the population were present in the MWTS between December 2008 and December 2010. For each of these months, we merged any non-imputed sales and inventories figures that were available from the MWTS into the population. At this point, the population still contained a high proportion of missing values for each MWTS variable. We used a sequential-regression plus hot-decked-residual procedure to impute the MWTS variables, in conjunction with distributional raking; refer to Klein, Lineback, and Schafer (2014) for details.

We do not claim that the simulations are a highly realistic representation of the MWTS because, like most national surveys, the MWTS has many complicated features. In our simulation study we made three major simplifications. First, we did not attempt to mimic births and deaths, but kept the population units constant for 25 months. Second, we ignored the fact that some companies operate in multiple industry groups, and we assigned each company's economic activity to only one group. Third, we did not try to replicate the benchmarking and seasonal adjustment procedures used in the MWTS, but took the post-imputation HT estimate for each industry group as the final answer. Despite these simplifications, we believe the simulation was realistic enough to uncover major strengths and weaknesses of the MWTS imputation and estimation procedures.

### 4.2 Simulation Results

Our simulation results indicate that the most crucial methodological concern for the MWTS is not imputation but the performance of the estimation procedure apart from missing data. To see what would happen if there were no missing data, we generated 1,000 samples and computed HT estimates before imposing any missing values. First we consider U.S. total sales. For each of the 24 months (January 2009 to December 2010), we have 1,000 sample estimates of the total sales. Boxplots of those 1,000 estimates are shown in Figure 3, omitting some extreme outliers at the high end. The black center line of each boxplot represents the median of the 1,000 estimates, the edges of each box represents the 25th and 75th percentiles. Superimposed over each of the boxplots is a large blue dot indicating the average of the 1,000 sample estimates, and superimposed over the entire graph is a thick black line representing the known total U.S. sales in the artificial population. Figure 3 indicates that the Horvitz-Thompson estimates of U.S. total sales are unbiased (as expected), but the sampling distributions of the estimators are highly skewed.

Figure 4 shows what happens with point and variance estimators for total sales, not just for the entire U.S., but within the 22 industry classes for which the MWTS results are published. There are $22 \times 24 = 528$ population totals to estimate (22 industry classes and 24 months). The left plot in Figure 4 shows the true population value ($x$-axis) versus the average of the 1,000 sample estimates ($y$-axis), along with a 45-degree line through the origin. The sample estimates are unbiased because the points lie along the 45-degree line. The right plot in the figure shows the sample variance of the 1,000 point estimates ($x$-axis)

versus the average of the random group variance estimate ($y$-axis). Again we see that the points lie along the 45-degree line, indicating that the random group variance estimate is unbiased for the actual sampling variance of the estimate.

We combine the estimates and standard errors to form a nominal 95% confidence interval based on the $t$ distribution with 15 degrees of freedom. (The random group variance is based on 16 random groups, which gives 15 degrees of freedom for estimating the variance.) Figure 5 shows a histogram of the actual coverage rates of the nominal 95% intervals (as observed in the simulation) for each of the $22 \times 24 = 528$ estimands. The histogram shows that the coverage rates are well below the 95% nominal level. The low coverage occurs because the sampling distributions of the estimates are highly skewed, they are not approximately normal. The Central Limit Theorem (under stratified random sampling) does guarantee that these estimates would have approximate normal sampling distributions if the sample size is large enough. But in this case, the underlying population is highly skewed, and the sample sizes are not sufficiently large for the Central Limit Theorem to take hold.

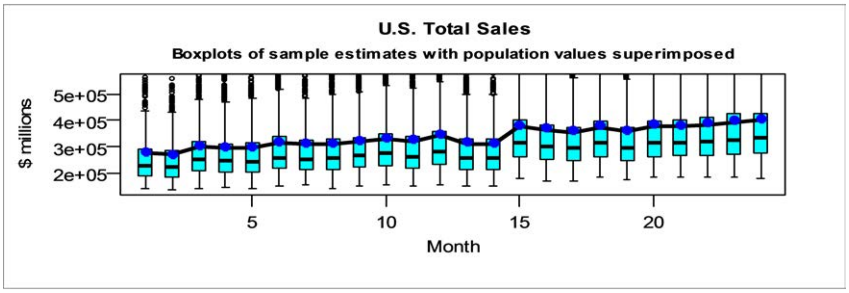


**Figure 3**: Complete data simulated distribution of U.S. total sales with no missing data

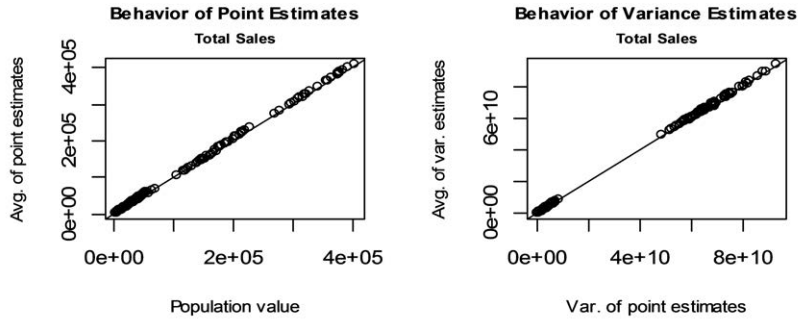


**Figure 4**: Complete data average sample estimates (average variance estimates) for total sales, plotted against known population values (variance of point estimates) for 22 industry classes across 24 months

If we look at estimates of percent change instead of totals, we again find that even when there are no missing data, the normal and $t$ based confidence intervals do not cover at the nominal rate. We refer to Klein, Lineback, and Schafer (2014) for more discussion and results. The key finding is that even when there are no missing data, the estimation procedures are not performing as expected, due to the highly skewed population distributions of sales and inventories. Taking this finding as our motivation, we will explore alternatives to the Horvitz-Thompson/large-sample normal approximation inferential procedure.

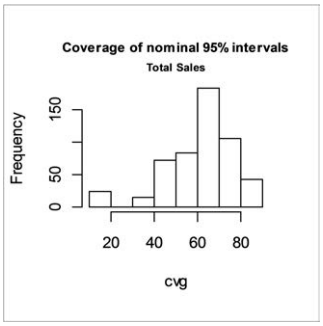


**Figure 5**: Histogram of simulated coverage rates of nominal 95% confidence intervals for total sales in 22 industry classes across 24 months

## 5. Bayesian Inference for Surveys

Using a highly simplified notation for a sample survey, let $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ denote a characteristic of interest (e.g., sales or inventories) for units in a population of size $N$. The goal is to use sample data to draw inferences about a population quantity $Q$ that is a function of $\boldsymbol{Y}$, for example, the population total $Q = \sum_{i=1}^{N} Y_i$. Define the sample inclusion indicators $\boldsymbol{I} = (I_1, \ldots, I_N)$, where $I_i = 1$ if unit $i$ is included in the sample, and $I_i = 0$ otherwise. Methods of traditional design-based inference (e.g., Cochran, 1977) treat the elements of $\boldsymbol{Y}$ as fixed constants, and they evaluate the distribution of an estimator for $Q$ over repeated realizations of $\boldsymbol{I}$ from hypothetical repetitions of the sampling procedure. The HT estimator for a population total, which is currently used in MWTS, is perhaps the best known example of a classical design-based method.

In recent decades, alternatives to design-based procedures have been emerging based on the idea of superpopulation modeling (Valliant, Dorfman and Royall, 2000). In this paradigm, a probability distribution $F$ is assumed for the values in the finite population, $Y_1, \ldots, Y_N \sim F(Y_i|\boldsymbol{X}_i, \boldsymbol{\theta})$, where $\boldsymbol{X}_i$ denotes variables for unit $i$ from the sample frame, and $\boldsymbol{\theta}$ is a vector of unknown parameters from the regression of $Y_i$ on $\boldsymbol{X}_i$. Superpopulation inference can be carried out using frequentist or Bayesian methods. Let $inc = \{i : I_i = 1\}$ denote the indices of population units that are included in the sample, and let $exc = \{i : I_i = 0\}$ denote the units that are excluded, so that the finite population total is

$$Q = \sum_{i \in inc} Y_i + \sum_{i \in exc} Y_i.$$

If the parameters $\boldsymbol{\theta}$ were known, we could predict an unsampled $Y_i$ by its expected value $\hat{Y}_i(\boldsymbol{\theta}) = E(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\theta})$. A frequentist estimate of $Q$ could be constructed as

$$\hat{Q} = \sum_{i \in inc} Y_i + \sum_{i \in exc} \hat{Y}_i(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is an estimate of the parameters computed from the sample data. The Bayesian approach places a prior probability distribution on $\boldsymbol{\theta}$ and conditions on the sampled values $\boldsymbol{Y}_{inc} = \{Y_i : i \in inc\}$ and frame variables $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ to generate a joint posterior distribution for $\boldsymbol{\theta}$ and the unsampled values $\boldsymbol{Y}_{exc} = \{Y_i : i \in exc\}$. The posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{Y}_{exc}$ induces a posterior distribution for the quantity of interest $Q$. For a formalization of the Bayesian argument and discussion of inclusion indicators and the role of nonresponse, see Chapter 2 of Rubin (1987).

In the 2007 Morris Hansen lecture, Joseph Sedransk called for the increased use of Bayesian methods in certain types of superpopulation inference (Sedransk, 2007). His first

example of an area of application where a switch to Bayesian methods would be beneficial was establishment surveys. In a typical establishment survey, there is a close and approximately linear relationship between the survey variable of interest $Y_i$ and one or more measure-of-size (MOS) variables in the sample frame. The skewed distribution for $Y_i$ suggests a stratified design where the largest units are selected with certainty.

Using the example of the Federal Reserve Board's program for estimating weekly monetary aggregates, Sedransk (2007) described an early Bayesian model that was implemented during the 1970's. For any given week $t$, let $Y_{hi}$ denote the value of monetary aggregate for bank $i$ in stratum $h$, and let $X_{hi}$ denote the measure of size for bank $i$ in stratum $h$ taken from the sample frame. The model was

$$Y_{hi} \;=\; \beta_h X_{hi} \;+\; \epsilon_{hi},$$

where the $\epsilon_{hi}$'s are independently distributed as $N(0, \sigma_h^2 X_{hi})$. The within-stratum variances $\sigma_h^2$ were replaced by point estimates and treated as known. A prior distribution was placed on the regression slopes, $\beta_1, \beta_2, \ldots \beta_H \mid \nu \sim N(\nu, \sigma_\beta^2)$, with a hyperprior $\nu \sim N(\beta_0, \sigma_\nu^2)$. The hyperparameters $(\sigma_\beta^2, \sigma_\beta^2, \sigma_\nu^2)$ were estimated by time-series analysis using five years of historical microdata. Under this model, posterior means for within-stratum finite-population totals were available in closed form using formulas given by Scott and Smith (1969) and Sedransk (1977).

With the advent of modern Bayesian methods and MCMC, compromises and simplifications in this model that were made for computational tractability in the 1970's would not be required today. For example, rather than substituting point estimates for the unknown $\sigma_h^2$'s, it is now possible to give a fully Bayesian treatment by applying a two-stage hierarchical prior to these variance components. The pooling of information across time that was done informally by plugging in time-series estimates of hyperparameters can now be accomplished by building an explicit longitudinal model for the unit-level responses $Y_{hi}$ across time. And it is now possible to handle arbitrary rates and patterns of missing values due to nonresponse, an issue that was not discussed at length by Sedransk (1977).

## 6. Envisioning Bayesian Estimation Procedures for the MWTS

We believe that Bayesian methods are worth pursuing for the MWTS for several reasons. First, Bayesian estimates are likely to be more precise than the current HT-based procedures. It can be shown that the HT estimator for a finite-population total is in fact a Bayes estimator under a certain superpopulation model whose assumptions are restrictive and typically unrealistic (Ghosh and Meeden, 1997, Chap. 5). When the assumptions are violated, a Bayes procedure under a more plausible model will tend to be more efficient, because it is able to make better use of information contained in covariates from the sample frame. Second, with a Bayesian procedure, there is no need to invoke problematic large-sample normal approximations; estimates and intervals for any quantity of interest may be obtained directly from simulated draws from a posterior distribution. Third, there is no need to impute missing values for nonrespondents prior to estimation. In the Bayesian paradigm, inferences are based on data only from the respondents; data values that are missing due to nonresponse and data values for unsampled units are handled in the same conceptual manner, provided that the sampling and nonresponse mechanisms are both ignorable in the sense defined by Rubin (1987). Finally, a Bayesian approach can eventually be extended to a unified system of estimation that combines results from MWTS and AWTS (currently handled by benchmarking) and performs seasonal adjustment when desired.

A model for Bayesian inference in the MWTS ought to reflect important features of the distributions of sales and inventories in the population. The model should be attentive

to major features of the sample design, so that inferences will be design-consistent and well calibrated (Little, 2006). Finally, the model should describe or condition on important correlates of missingness, to make the assumption of ignorable nonresponse more plausible (Rubin, 1976, 1987).

The multivariate linear mixed model we described in Section 3, which we developed for multiple imputation of missing values, provides a useful starting point for a system of Bayesian estimation. We found that the normal model is an effective tool for preserving relationships between sales and inventories, and for pooling information and preserving trends within companies and within industry groups across time. Covariates from the sample frame and indicators for sampling strata are readily incorporated into the model as predictors.

A major difficulty with this model, however, is that it assumes the survey outcomes are normally distributed. In reality, sales and inventories are semicontinuous, a mixture of zeros (which are occasionally seen for small companies) and positive values that are highly skewed. Model-based inferences about finite-population totals in skewed populations may be sensitive to distributional shape, especially to assumptions about tail behavior. In our previous work, we transformed sales and inventories to approximate normality using their empirical cumulative distribution functions (cdf's) and the normal quantile function (inverse cdf). This ad hoc method performed well enough for imputing missing values for nonrespondents, but we are reluctant to apply it in a model for Bayesian finite-population inference, where we would effectively be imputing data for all unsampled units in the population. A general rule for missing-data problems is that, as rates of missing information increase, inferences become more sensitive to the specification of the model for the complete-data population. Our ad hoc procedure treated the transformations (which were highly dependent on the sample data) as a priori fixed and known, and thus failed to represent uncertainty about the shapes of the marginal distributions of sales and inventories in the population.

In the next phase of our research, we are seeking ways to combine the power of the normal model for preserving relationships and trends with flexible, nonparametric techniques for capturing the shapes of marginal distributions. At present, we are experimenting with the use of copulas (Sklar, 1959; Smith, 2011). A copula is a joint density function for a vector of random variables in which each random variable is marginally distributed as uniform on the unit interval (0,1). One of the simplest types is a Gaussian copula, which can be constructed from a multivariate normal by centering and scaling each coordinate to have mean zero and variance one, and then transforming each centered and scaled variate by the standard normal cdf. With copulas, we can effectively separate the modeling of marginal distributions from the modeling of intervariable relationships. MCMC algorithms for Bayesian inference have been developed for copula models combined with marginal distributions that are continuous, discrete, and mixed type (Pitt, Chan and Kohn, 2006; Smith and Khaled, 2012). Our near-term goal is to adapt copula methods to link the multivariate linear mixed model described in Section 3 with a nonparametric Bayesian methods for estimating the marginal densities for sales and inventories.

To illustrate how this density estimation might work in practice, Figure 6 shows a histogram of simulated inventories from the artificial population. A small but non-negligible percentage of the values are zero. The positive part of the distribution is displayed on a log scale, but tick labels for the x-axis correspond to the original scale. The dashed line represents the density from a lognormal fit. The lognormal model fails to capture the heaping of observations and minor mode seen at the left-hand side of distribution, and it assigns too much density to the tail on the right. Understating or overstating the right-hand tail can wreak havoc on an estimate for a finite population total (we are estimating total dollars, not

total log-dollars). None of the standard power transformations within the Box-Cox family can effectively normalize this distribution.

In contrast, a nonparametric density estimate, shown by the solid line, captures the shape of this distribution extremely well. We estimated this function by fitting a loglinear penalized B-spline to binned histogram frequencies (Eilers and Marx, 1996). This method transforms the problem of density estimation into a Poisson regression with fixed and random effects. Many other methods for nonparametric Bayesian density estimation are available: Dirichlet process modeling, Gaussian mixtures, and Gaussian process models, to name a few. Selecting an appropriate and feasible method for density estimation, and linking it to a copula version of our multivariate linear mixed model, is a topic of ongoing research.

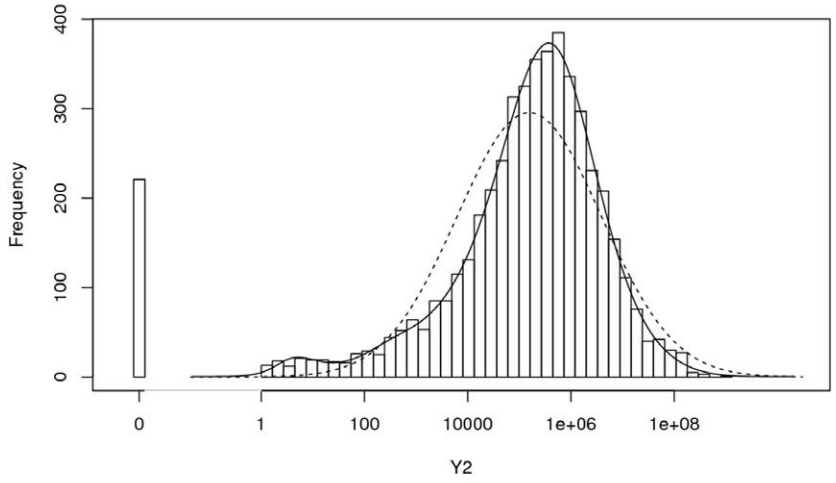


**Figure 6**: Histogram of simulated inventories from artificial population, positive values logarithmically transformed, with densities estimated by a lognormal fit (dashed line) and penalized B-spline (solid line)

## REFERENCES

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley & Sons.

Eilers, P. H. C., and and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89-121.

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*, London: Chapman & Hall.

Klein, M., Lineback, J.F., and Schafer, J. L. (2014), "Evaluating Imputation Techniques in the Monthly Wholesale Trade Survey," In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 1814-1827.

Lineback, J. F., and Schafer, J. L. (2013), "Multivariate Linear Mixed-Effects Models for Missing Data Applied to a Business Survey," Presented at the *Joint Statistical Meetings*.

Little, R. J. (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, 99, 546-556.

Little, R. J. (2006), "Calibrated Bayes: A Bayes/Frequentist Roadmap," *The American Statistician*, 60, 213-223.

Pitt, M., Chan, D., and Kohn, R. (2006), "Efficient Bayesian Inference for Gaussian Copula Regression Models," *Biometrika*, 93, 537-554.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-592.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New Jersey: John Wiley & Sons.

Schafer, J. L., and Yucel, R. M. (2002), "Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values," *Journal of Computational and Graphical Statistics*, 11, 437-457.

Scott, A. and Smith, T.M.F. (1969). "Estimation in Multistage Surveys," *Journal of the American Statistical Association*, 64, 830–840.

Sedransk, J. (1977), "Sampling Problems in the Estimation of the Money Supply," *Journal of the American Statistical Association*, 72, 516–522.

Sedransk, J. (2007), "Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities," *Journal of Official Statistics*, 24, 495–506.

Sklar, A. (1959), "Fonctions de Répartition à n Dimensions et Leurs Marges," *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229-231.

Smith, M. S. (2015), "Bayesian Approaches to Copula Modelling," in *Bayesian Theory and Applications*, eds. P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, New York: Oxford University Press, pp. 336-358

Smith, M. S., and Khaled, M. A. (2012), "Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation," *Journal of the American Statistical Association*, 107, 290-303.

Smith, P. (2013), "Sampling and Estimation for Business Surveys," in *Designing and Conducting Business Surveys*, New York: John Wiley & Sons, pp. 165–218.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley & Sons.

Wolter, K.M. (2007), *Introduction to Variance Estimation*, Second edition, New York: Springer.