Evaluating Imputation Techniques in the Monthly Wholesale Trade Survey

Martin Klein, Joanna Fane Lineback, Joseph L. Schafer[1]

Abstract

The Monthly Wholesale Trade Survey (MWTS) provides estimates of the dollar value of sales and inventories for wholesale businesses in the United States. In this longitudinal survey, missing values for sales and inventories in any month are imputed via a ratio adjustment applied to data from the prior month. In this article, we describe ongoing research to evaluate the performance of the current imputation method and to investigate possible alternatives. Using information from the MWTS and the sample frame, we generated an artificial population of wholesale businesses with two years of monthly sales and inventory data. We repeatedly drew samples from this artificial population, imposed patterns of nonresponse on the samples, and filled in the missing values by two methods: the current ratio-based procedure and a new model-based multiple-imputation procedure. Preliminary results from this simulation are challenging to interpret because, apart from missing data, the inferential procedures (complete-data point estimates, variance estimates and interval estimates) do not behave as large-sample normal theory suggests they should. Based on these results, we recommend further research on improving the quality of the complete-data inferences, using methodologies that are better suited for stratified sampling from populations that are highly skewed.

Key words: business survey, item nonresponse, multiple imputation

1. Background

The Monthly Wholesale Trade Survey (MWTS) is a longitudinal business survey conducted by the U.S. Census Bureau. The MWTS yields estimates of end-of-month inventories and sales for merchant wholesale businesses in the United States, both overall and within industry groups defined by the North American Industry Classification System (NAICS). Public interest and media reports focus on the overall relative change in end-of-month inventories,

$$\%\ \text{relative change} = 100 \times \left(\frac{\text{current month inventories} - \text{previous month inventories}}{\text{previous month inventories}} \right).$$

Complete reports are released at www.census.gov/wholesale/ approximately six weeks after the close of each month. All aspects of surveys operations – data collection and processing, review, editing, imputation, estimation, benchmarking and seasonal adjustment – are completed within this short time frame. Proposals for methodological innovations and changes to the survey must be evaluated in light of this compressed production schedule and the need to maintain comparability in estimates across time to avoid breaks in the series. No microdata files from the MWTS are released; the monthly wholesale trade reports and the Manufacturing and Trade Inventories and Sales report are the major data products.

Missing data arise if a reporting unit fails to provide sales and/or inventories in a given month or if reported data are unusable. On average, item nonresponse is approximately 30%, with inventories seeing higher rates of nonresponse than sales. Currently, missing data are imputed based on estimated ratios. Within imputation cells defined by size and industry class, the imputed value for a unit's inventories (sales) is equal to that unit's prior month's inventories (sales), multiplied by the ratio of the total current month's inventories (sales) to the total previous month's inventories (sales). In our previous work (Lineback and Schafer, 2013), we identified potential problems with this method. Notably, this method implicitly relies on a regression model with intercept assumed to be zero and a slope assumed to be one, which may be unrealistic. It ignores other information that may be relevant for predicting the missing observations, such as sales and inventories from earlier months (prior to the last month), correlates from the census of wholesale businesses conducted every five years, and administrative payroll, revenue or employment data. Additionally, in some cells, the estimated ratios are based on very few cases, giving them high variability. Sales and inventories are imputed separately, so the relationship between them may be distorted. Finally, there is no attempt to assess missing-value uncertainty or include it in variance estimates.

[1] United States Census Bureau, 4600 Silver Hill Road, Washington DC 20233-9100. This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on statistical, methodological or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. Authors' names are listed in alphabetical order.

In previous work, we investigated strategies for multiple imputation under models that make better use of the available information (Lineback and Schafer, 2013). We proposed a multivariate normal longitudinal regression model (Schafer and Yucel, 2002) that allows flexible trends for inventories and sales within industry groups. Because inventories and sales are not normally distributed, we transformed them to approximate normality using smoothed empirical distribution functions and the Gaussian-quantile function (inverse cdf). We applied the new imputation method to monthly data from the MWTS over the 25-month period from December 2008 to December 2010. The new imputation method produced estimates of relative change that were smaller in magnitude (closer to zero) than the current method and standard errors that were much larger than the current method.

In the present article, we describe ongoing simulation work to evaluate the performance of these imputation methods. Using nonparametric techniques, we constructed an artificial population to mimic the behavior of the actual population over the two-year span from December 2008 to December 2010. Because we have access to the entire population, the truth (i.e., the set of "true" values for all population target estimands) is known, providing a gold standard against which the procedures may be compared. From this artificial population, we drew 1,000 probability samples using a stratified-sample design that closely resembles the MWTS sample design. For each sample, we calculated point estimates and standard errors for key population quantities with no missing data. We then imposed nonresponse on each sample by a stochastic mechanism that produces realistic rates and patterns of missing values. We filled in the missing values by the current ratio-based single-imputation method and a new model-based multiple-imputation method. Finally, we computed estimates and standard errors from each imputed sample and assessed their performance for drawing inferences for the known target quantities from the pseudo-population.

In the remaining sections, we describe these steps in greater detail. In Section 2, we explain how we constructed the artificial population. In Section 3, we describe our methods for selecting samples and imposing patterns of nonresponse. In Section 4, we describe our methods for ratio-based imputation and multiple imputation. In Section 5, we present some preliminary results, and in Section 6, we discuss our findings and offer tentative recommendations

2. The Artificial Population

2.1 Simplifications

Like most national surveys, the MWTS has many complicated features that are difficult to reproduce in a simulation study. For this preliminary round of simulations, we made three major simplifications.

First, we pretend that there are no "births" or "deaths." In reality, the population of wholesale companies doing business in the United States is constantly in flux. New companies appear, and existing ones go out of business or disappear due to mergers or acquisitions. The MWTS sample frame is updated each quarter-year to account for these birth and deaths. Devising realistic processes for simulating births and deaths proved to be challenging and time consuming. In this present article, we do not attempt to model the births and deaths. For these simulations, the population roster of units remained fixed for the entire two years.

Second, we pretend that each company operates in only one industry group. Each month, the U.S. Census Bureau publishes estimates of sales and inventories for wholesale companies at 22 levels determined by the North American Industry Classification System (NAICS): U.S. Total, Durable, Automotive, Furniture, Lumber, Professional Equipment, Computer Equipment, Metals, Electrical, Hardware, Machinery, Miscellaneous Durable, Nondurable, Paper, Drugs, Apparel, Groceries, Farm Products, Chemicals, Petroleum, Alcohol, and Miscellaneous Nondurable. Some of these publication levels are nested within others. For example, the U.S. Total includes Durables and Nondurables; the Durables includes Automotive, Furniture, Lumber, Professional Equipment, Computer Equipment, Metals, Electrical, Hardware, Machinery, and Miscellaneous Durables. If we rearrange these publication levels into categories that are mutually exclusive, we obtain 19 groups that cover the wholesale economy: Alcohol, Apparel, Automotive, Chemicals, Computer Equipment, Drugs, Electrical, Farm Products, Furniture, Groceries, Hardware, Lumber, Machinery, Metals, Miscellaneous Durable, Miscellaneous Nondurable, Other Professional Equipment, Paper, and Petroleum.

Some companies, especially large ones, operate in multiple industry groups. A company that sells machinery might also distribute hardware. During MWTS data collection, companies that operate in multiple groups are asked to report their sales and inventories for each industry group. Sometimes they are unable to do so, and instead they report company-wide totals; these totals are distributed across the industry groups using a predetermined set of multipliers (so-called kind-of-business factors) estimated beforehand from administrative data. Before a company is selected into the MWTS, we do not necessarily know all of the industry groups in which the company operates. When the MWTS sample frame is constructed, each company in the population is assigned to one group that represents the best guess for the area in which it does most of its business. That major industry group is used to assign the company to a sampling stratum.

For this round of simulations, we ignore the fact that many companies operate in multiple industry groups. That is, we pretend that all the sales and inventories for a population unit reside in the major industry group that was used to assign the company to a sampling stratum. This simplification is unrealistic, but it eliminates the daunting prospect of having to simulate kind-of-business factors for each unit in the population at each month. (This appears to be the reason why we had to adjust the values of sales and inventories in the simulated population by the "distributional raking" procedure to be described in Section 2.3.)

Third, we ignored benchmarking and seasonal adjustment. In the MWTS, total sales and inventories within industry classes are estimated by a standard Horvitz-Thompson (HT) procedure. The HT estimates are then benchmarked to figures from the larger Annual Wholesale Trade Survey, of which the MWTS is a subsample, and seasonally adjusted to eliminate the predictable cyclic variation that occurs within each calendar year (Brown, 2012). For this round of simulations, we use raw HT estimates with no benchmarking or seasonal adjustment.

Because of these simplifications, we do not claim that these simulations are a highly realistic representation of the MWTS. Our purpose was not to perfectly replicate the survey, but to create a laboratory for studying the missing-data procedures like those being used in the survey. These simplifications will affect our conclusions insofar as they *interact* with the missing-data procedures, skewing the comparisons by favoring one missing-date method over another. At present, we do not have any compelling reasons to conjecture that such interactions would be strong.

2.2 Establishments, companies, EINs and sample units

Thus far, we have been loosely referring to the MWTS sample units as companies. This is not strictly true, however, and before we proceed, we must clarify several terms.

Establishments and companies. An establishment is a single physical location where business is conducted or where services are performed, and a company (also called a firm; we use the two terms interchangeably) is a business organization that operates one or more establishments under common ownership or control. For example, in the retail sector, an establishment could be a single grocery store, and a company could be a regional chain of supermarkets.

Employer Identification Numbers or EINs. An EIN is an identifier issued by the Internal Revenue Service (IRS) to a legal entity that hires employees. Any business organization that hires workers must have an EIN. A firm with multiple establishments may centralize its payroll under a single EIN, or it may have a separate EIN for each establishment.

The MWTS sampling frame is constructed from an administrative list called the Business Register (BR). The IRS provides EINs and other basic information needed to maintain the BR, and the BR becomes the basis for economic censuses and business surveys conducted by the Census Bureau. The largest wholesale companies in the nation are identified and included in the MWTS with probability one. These so-called certainty units almost always consist of multiple EINs, and when the frame is constructed, basic information on these units (e.g. measures of size) are summed or "rolled up" as necessary to the company level. Noncertainty units, which are selected into the MWTS with probabilities less than one, are individual EINs representing a whole company or a portion of a multiple-EIN company. Thus, for present purposes, we will say that there are two types of sample units in the MWTS: certainty units, which are multi-EIN companies, and noncertainty units, which are single EINs.

2.3 Generating the population

From the actual MWTS sample frame, we obtained a roster of approximately 286,000 units (companies and EINs) to represent the population of merchant wholesale companies in the United States (excluding manufacturer sales branches and offices) during the last quarter of 2008. Approximately 3,200 of these units were present in the MWTS between December 2008 and December 2010. For most of these units, some variables were also available from the previous 2007 Economic Census. We gathered whatever non-imputed sales and inventories figures that were available from the actual MWTS data from December 2008 to December 2010, and four variables from the 2007 Economic Census (number of employees, annual payroll, first quarter payroll, and revenue), and brought them over to our population roster. After this merging, the roster had very high proportions of missing values for each variable from the MWTS. Our challenge was to fill in these missing values in a manner that preserves key aspects of the marginal distributions and inter-variable relationships, but without making strong modeling assumptions that would tend to favor either of the missing-data methods that we were going to use in the subsequent simulations.

We first filled in the missing values in the Economic Census variables using a sequential-regression random-forest method inspired by the recent work of Doove, Van Buuren and Dusseldorp (2014). Random forests average over large collections of classification or regression trees to generate predictions of an outcome variable from potentially large numbers of covariates, without making strong assumptions about how the outcome is related to the covariates. Doove, Van Buuren and Dusseldorp (2014) embed random forests into the sequential-regression multiple-imputation techniques that are currently implemented in the software packages IVEware (Raghunathan et al., 2001) and MICE (Van Buuren et al., 2006).

After the census variables were completed, we filled in the sales and inventories each month by a sequential imputation procedure that took advantage of the fact that the missing-data pattern in these variables was nearly monotone. We arranged the MWTS variable in the following sequence,

$$(\text{sales in month } j, \text{ inventories in month } j), \text{ for } j=0, 1, \ldots, 24,$$

where j=0 represents December 2008 and j=24 represents December 2010. Under this ordering, the missing values for these 50 variables formed a near-monotone pattern; a unit that was missing the jth variable was almost sure to have missing values for every subsequent variable in the sequence. Monotone patterns are advantageous, because they allow us to impute missing values under a bonafide joint distribution without costly iteration, sweeping through the variable sequence only once. For each variable in this sequence, we fit a regression model to predict that variable from variables earlier in the sequence (up to ten of them) and the variables from the Economic Census. Because these variables are highly skewed, each one was transformed to its cube root. A subset of predictors was chosen by a forward-backward stepwise procedure. The missing values were then imputed using a regression-prediction-plus-hot-decked-residual procedure, with hot deck donor cells defined by the deciles of the fitted values. This entire procedure was carried out separately for each of the 19 non-overlapping industry groups mentioned in Section 2.1.

After the procedure was finished, we compared the marginal distributions of these 50 simulated variables to marginal distributions estimated from the MWTS. The MWTS-estimated marginals were obtained by expanding the set of observed values from the units included in the MWTS, replicating each value a number of times equal to its sampling weight (after rounding the weights to the nearest integers). We discovered that the quantiles of our simulated variables were substantially larger than the corresponding MWTS-estimated quantiles, especially in the upper tails. We believe that these discrepancies arose mainly from the second simplification described in Section 2.1: pretending that each company operates in only one industry group, the group where it had been thought to be conducting most of its business when the frame was created. To correct these discrepancies, we adjusted our simulated data by a procedure that we call "distributional raking." For each variable, we ordered the simulated population values from smallest to largest, and then replaced each value by its corresponding MWTS-estimated quantile.

These transformations forced the marginal distribution of each simulated variable in the artificial population to match its marginal distribution estimated from the actual MWTS. Key estimands from this population also resemble their corresponding estimates from the MWTS. Figure 1 compares simulated and estimated values for five important quantities for the 25-month period: total sales, total inventories, inventories-to-sales ratio, percent change in sales,

and percent change in inventories. In each plot, the solid line shows the "true" values from our simulated population roster, and the dashed line represents the raw Horvitz-Thompson estimates from the MWTS.

Figure 1. Simulated population vs. MWTS estimates (December 2008-December 2010)

3. Sample Selection and Nonresponse

After the artificial population was constructed, we drew 1,000 random samples using a stratified simple random sample design that mimics the actual design of the MWTS. Units from the MWTS that had been selected with probability one were included with probability one. All other units were assigned to 369 noncertainty strata representing the actual strata used for the MWTS, defined by industry class and size. Within each stratum, we drew a simple random sample (SRS) without replacement, with number of units selected equal to the number actually found in the MWTS. Sampling weights were computed as 1.0 for certainty units and N_j/n_j for each noncertainty unit in stratum j, where N_j and n_j are the population and sample sizes in stratum j.

Variance estimation in the MWTS is carried out by the method of random groups. The random-group method is described in Chapter 2 by Wolter (2007) for the Census Bureau's Retail Trade Survey, and the procedure used in the MWTS is nearly the same. For each of the 1,000 samples from our population, we ordered the noncertainty units by industry class and size. We chose a random integer s between 0 and 15, and then assigned the noncertainty units to 16 random groups in the order

$$s + 1, s + 2, \ldots, 16, 1, 2, \ldots, 16, \ldots$$

For each sampled unit in each sample, we imposed a pattern of missing values on the monthly sales and inventories (a vector length 50, containing 0's and 1's) by a hot deck within cells defined by industry class and size (essentially the sampling strata). Donors in each cell were the units from the actual MWTS. The simulated rates and patterns of missingness closely mimic the actual MWTS, and the missing values are missing at random (MAR), but not missing completely at random (MCAR), because probabilities of missingness depend on information from the frame.

4. Imputation Methods

For each sample from our simulated population, we imputed the missing values for sales and inventories by two different methods.

The first method ("Current Method") mimics the ratio-imputation procedure used in the MWTS. Each of the 19 non-overlapping industry groups was divided into one, two, or four imputation cells. The choice of one, two or four was made to ensure that each cell would have at least 15 donor cases with high probability. Cut-points for the cells were defined by the quartiles of sales and inventories in the population at the start of the simulation (December 2008). In each sample, sales (inventories) were imputed each month for 24 months (January 2009 to December 2010) using the MWTS ratio-imputation formula. After imputation, quantities of interest were estimated from the sample using the raw HT procedures with no benchmarking or seasonal adjustment, and a random-group variance estimate was computed for each point estimate.

The second method ("MI Method") is a Bayesian multiple-imputation procedure based on a multivariate normal regression model that preserves the correlation between sales and inventories each month and their relationships across months. Because these variables are not normally distributed, we normalized them using the empirical-cdf transformation described in Lineback and Schafer (2013). Then, in each of the 24 months (January 2009 to December 2010), we fit a multivariate normal regression using information available up to that month. The Y-variables were the sales and inventories for that month, the previous month, the month before that, and so on, up to a total for five occasions. The X-variables in the regression were dummy indicators for the 19 non-overlapping industry groups, plus the four variables from the 2007 Economic Census normalized by the empirical transformation procedure. Maximum likelihood estimates (MLEs) for the parameters were obtained by running an expectation-maximization (EM) algorithm (a slight generalization of the algorithm described by Schafer, 1997), and five MIs were generated by a Markov chain Monte Carol (MCMC) single-chain method. After imputation, we computed raw HT point estimates and their associated random-group variance estimates, and we combined the results across the multiple imputations using Rubin's (1987) rules.

5. Results

When we apply missing-data methods, we typically make the assumption that, if there were no missing data, the inferences would conform to large-sample normal theory. That is, we assume that, if there had been no missing data,

- estimates of population quantities (point estimates) would be approximately unbiased for the population true values,
- variance estimates (squared standard errors of the point estimates) would accurately reflect the true variability in the point estimates, and
- confidence intervals would have actual coverage close to nominal levels (e.g., the point estimate plus or minus two standard errors would cover the population true values about 95% of the time).

If these assumption are correct – if the inferential procedures work well without missing data – then we may consider how the missing data and missing-data methods impact the quality of these inferences.

In the present study, however, we discovered that the inferential procedures without missing data were not working as well as we had hoped.

5.1 Performance of complete-data estimates and standard errors

To see what would happen if there were no missing data, we computed estimates and standard errors for quantities of interest from each sample before we imposed any missing values. First, we consider estimates of total U.S. sales. For each of the 24 months, we have 1,000 sample estimates of the totals sales for that month. Boxplots of those 1,000 estimates by month are shown in Figure 2. For clarity, we omitted some extreme outliers at the high end. These outliers, which fall outside of the plotting region, represent a small number of samples for which the HT estimates were extremely large.

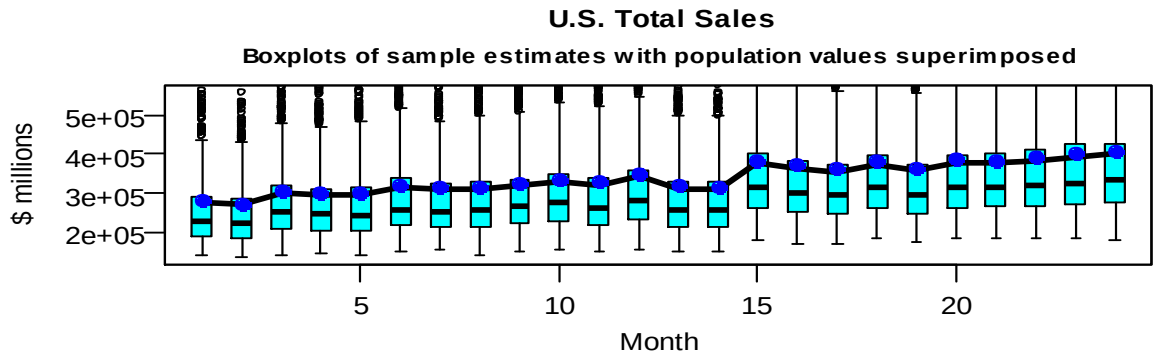

Figure 2. Boxplots of simulated complete-data sample estimates of total U.S. sales, with averages of the sample estimates (large blue dots) and known targets from the artificial population (thick black line) superimposed

The black center line of each boxplot represents the median of the 1,000 estimates, and the edges of each box represent 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.

One noteworthy aspect of Figure 2 is that the means of the sample estimates (blue dots) land squarely on top of the known population targets (thick black line). In other words, the sample estimates are unbiased. This is not surprising, because statistical theory tells us that HT estimates for totals are exactly unbiased. Another noteworthy aspect is that the distributions of the sample estimates are skewed. The skewness would appear to be even more dramatic if all of the outliers were shown. Because of this skewness, the means of the sample estimates lie well above their medians. The skewness arises because the distribution of sales across companies and EINs in the population are very highly skewed. These units vary greatly in size, and a large proportion of the total sales can be attributed to a small number of very large firms. The stratified sample design and the designation of the largest units as certainty units mitigates this to a degree. If this were a SRS, the skewness would be even more pronounced. The stratification does help, but the design does not get rid of all skewness in the sample estimates. (If the sample size were greatly increased, the Central Limit Theorem (CLT) – or a generalized version of the CLT for stratified samples – would eventually take hold, guaranteeing normality of the sampling distributions). Skewness is an issue because many inferential procedures assume that the sample estimates are approximately normally distributed. Notably, the usual method for obtaining a confidence interval (estimate plus or minus two SEs for a normal-theory 95% interval) might not have accurate coverage if the normal approximation is poor. Good behavior of ratios of totals and relative change estimates also depends on the asymptotic normality of the sample totals; if the sample totals are not normally distributed, then nonlinear functions of them might not be well-behaved.

Now we show what is happening with the point and variance estimates for total sales, not just for the entire U.S., but within the 22 industry classes for which the MWTS results are published. With 22 industry classes and 24 months, there are 22x24=528 population sales totals that we are estimating. In Figure 3, the plot on the left shows the true population value (the x-axis) and the average of the 1,000 sample estimates (the y-axis), along with a 45-degree line through the origin. The points lie along this 45-degree line, showing that the sample estimates are indeed unbiased. The plot on the right-hand side shows the behavior of the variance estimates. The x-axis is the sample variance of the 1,000 point estimates, i.e., the actual variability of the point estimate over repeated sampling, and the y-axis is the average value of the squared standard error, i.e., the average of the 1,000 variance estimates that come from the random-group method. Once again, these points sit on the 45-degree line, showing that the random-group variance estimator is an unbiased estimate of the sampling variance. The variance estimates are doing what they were designed to do, which is to give an unbiased estimate of the variability of the point estimate for a population total.

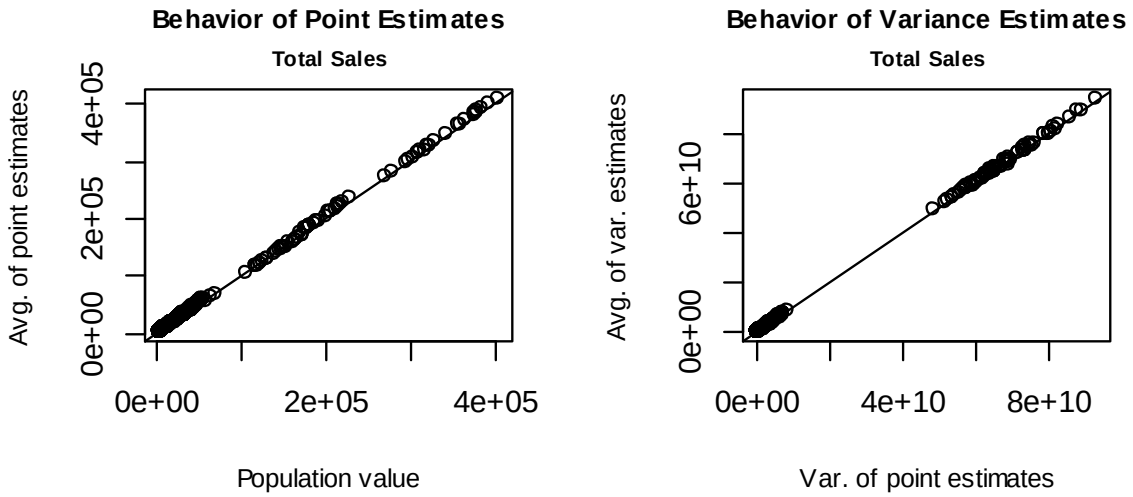

Figure 3. Complete-data average sample estimates (average variance estimates) for total sales, plotted against known population values (variances of point estimates) for 22 industry classes across 24 months

But when we combine the point and variance estimates to produce classic 95% confidence intervals, i.e., the point estimate plus or minus two standard errors, those confidence intervals do not cover the true population values 95% of the time. Refining this somewhat, there are 16 random groups, which give us 15 degrees of freedom for estimating the variance. Taking the estimate plus or minus 2.13 standard errors, because 2.13 is the 97.5th percentile of the t-distribution with 15 degrees of freedom, gives an interval that will have 95% coverage if the sampling distributions are normal. But the actual coverage rates are less than that, as shown by the histogram in Figure 4 below. If the procedures were working well, this histogram would spike at 95%, but the actual coverage rates are much smaller. If we restrict our attention just to the U.S. totals (there are 24 of these, one for each of the 24 months), the coverages of those intervals are approximately 67%. Over repeated samples, the distributions of the point estimates are far from normal, and the distributions of the variance estimates divided by the true variances are far from a chi-squared distribution with 15 degrees of freedom. The combination of these two factors makes the performance of the t-intervals suffer.

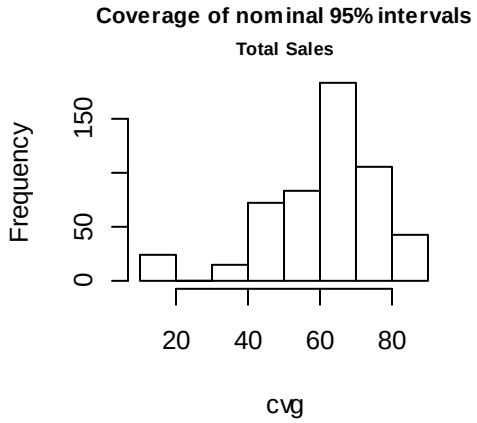

Figure 4. Histogram of simulated coverage rates of nominal 95% confidence intervals for total sales in 22 industry classes across 24 months

Now we turn our attention to estimates of percent change in sales from the current month relative to the previous month. The plot in Figure 5 is analogous to the one shown in Figure 2, except now it pertains to percent change.

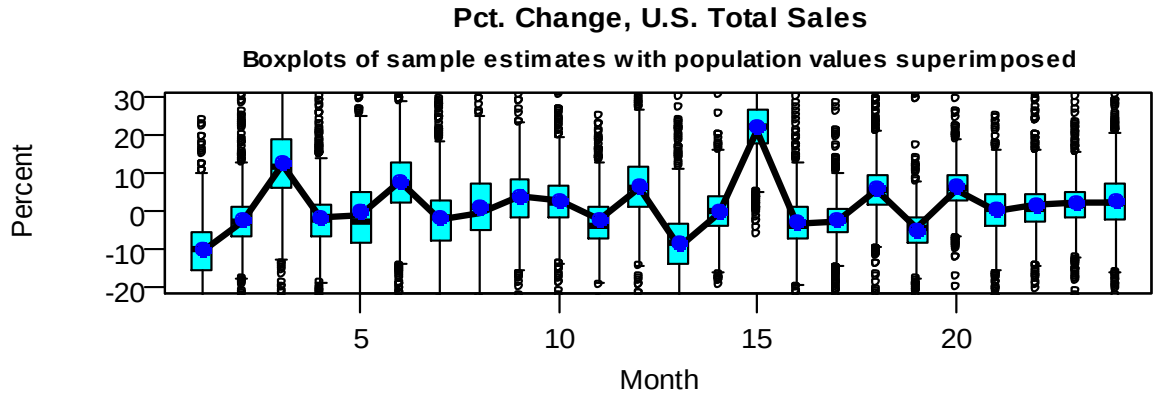

Figure 5. Boxplots of simulated complete-data sample estimates of percent change in total U.S. sales, with averages of the sample estimates (large blue dots) and known targets from the artificial population (thick black line) superimposed

As before, the sample estimates are essentially unbiased. Skewness is no longer a major issue; the relative change between two totals is less skewed than the totals themselves. However, these estimates are still far from normally distributed; the sampling distributions have very heavy tails at the upper and lower ends. (Some of the extreme outliers at the top and the bottom land far outside of the plotting region.) The leptokurtic sampling distributions of the point estimates creates problems, as does the behavior of the variance estimates. The performance of the point and variance estimates is summarized in Figure 6. In the plot on the left-hand side, the points cluster along the 45-degree line, which means that the estimated percent changes are approximately unbiased. In the plot on the right-hand side, however, the variance estimates behave erratically. If the variance estimates were unbiased, all of the points would lie close to the 45-degree line, but many of them are far from that line. Three of them were so far from the line that for clarity they had to be omitted from the plot. Those outliers are for the small domains (e.g., farm products) for which the sample sizes are small.

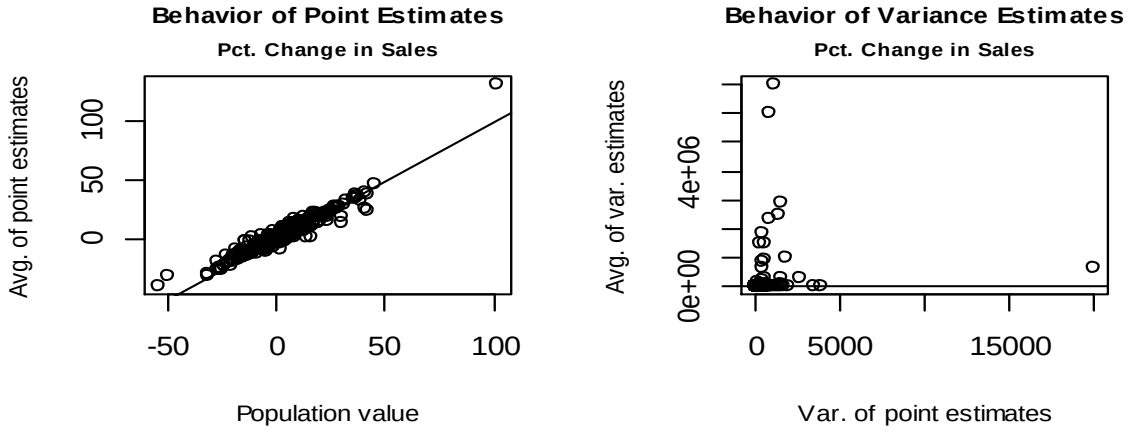

Figure 6. Complete-data average sample estimates (average variance estimates) for percent change in total sales, plotted against known population values (variances of point estimates) for 22 industry classes across 24 months

If we focus just on the national level – the percent change in sales for the entire wholesale economy – the situation is not much better. Figure 7 shows the corresponding plots just for the 24 national-level estimates. The point estimates are essentially unbiased, but in the plot on the right-hand side, all the points except one lie far below the 45-degree line. This indicates that the random group variance estimates for percent change are understating the actual variability. The standard errors for the change estimates are too small, and if you use them in confidence intervals, you get coverage that is not near 95%. At the national level, the confidence intervals for percent change in sales have an actual coverage of approximately 73%.

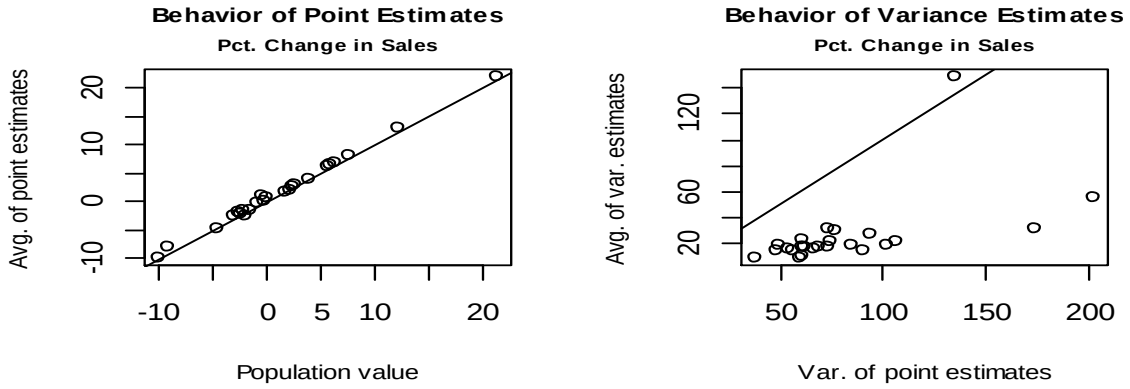

Figure 7. Complete-data average sample estimates (average variance estimates) for percent change in total sales, plotted against known population values (variances of point estimates) at the national level across 24 months

5.2 Performance of Current Imputation Method

We now show what happens when we impose patterns of missing values and impute by the current method. Boxplots of the estimated total sales at the national level are displayed in Figure 8. Once again, the black line represents the true population values, and the blue dots are the means of the 1,000 estimates. And, once again, some of the outliers at the top of the plot have been trimmed away. In this plot, the blue dots lie slightly above the black line, showing that the missing data and imputation mechanisms have introduced a slight upward bias.

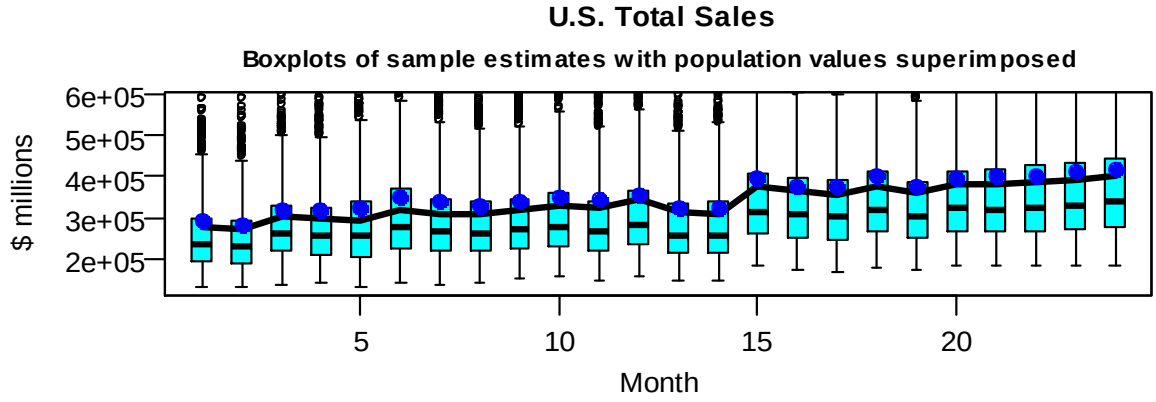

Figure 8. Boxplots of simulated sample estimates of total U.S. sales after imputation by the "current method," with averages of the sample estimates (large blue dots) and known targets from the artificial population (thick black line) superimposed

For the sales totals at all publication levels, the behavior of point and variance estimates is summarized in Figure 9. In the plot on the left-hand side, most of the points lie above the 45-degree line, indicating that those point estimates have an upward bias. On average over these 22x24=528 estimates, the size of this bias is about 4% of the size of the estimands. One could argue that an upward bias of 4% in population totals is practically important, and perhaps it is. Interestingly, the missing-data procedure does not have much of an impact on the variance estimates. The plot on the right-hand side is clustered tightly around the 45-degree line, showing that the variance estimates are still unbiased. However, as noted before, the skewness of the sampling distributions makes the coverage of confidence intervals already well below 95%, and therefore introducing a bias of 4% in the estimates does not have much impact on the coverage.

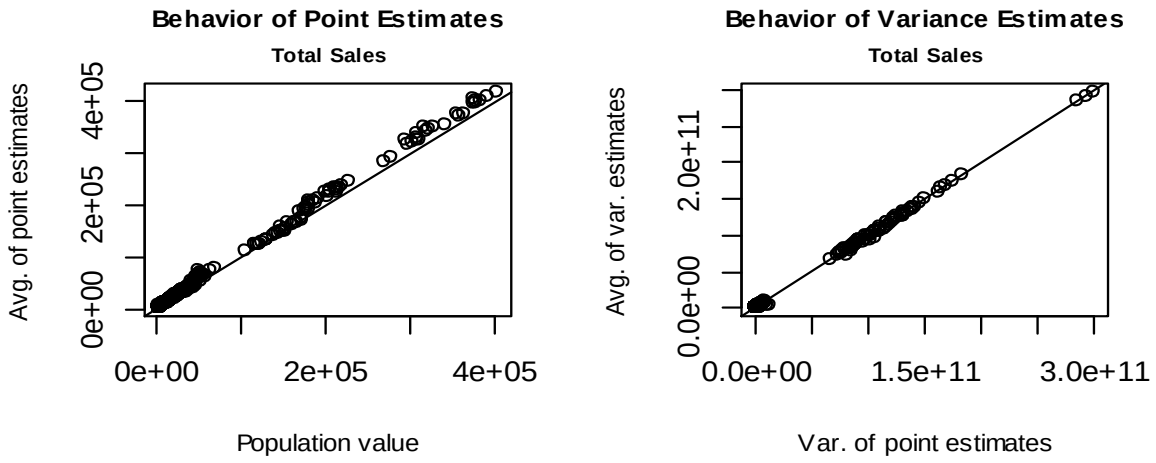

Figure 9. Average sample estimates (average variance estimates) for total sales after imputation by the "current method," plotted against known population values (variances of point estimates) for 22 industry classes across 24 months

Now let us examine the estimates of percent change, and let us focus on the national level. The plot on the left in Figure 10 shows that the points that before were tightly clustered along the 45-degree line have now strayed from the line. Most of them are slightly above, meaning that the estimates of change now have a slight upward bias. The real difficulty, however, is shown in the plot on the right. With complete data, the points were already below the 45-degree line, showing that the variance estimates were too small. Now the situation is even worse.

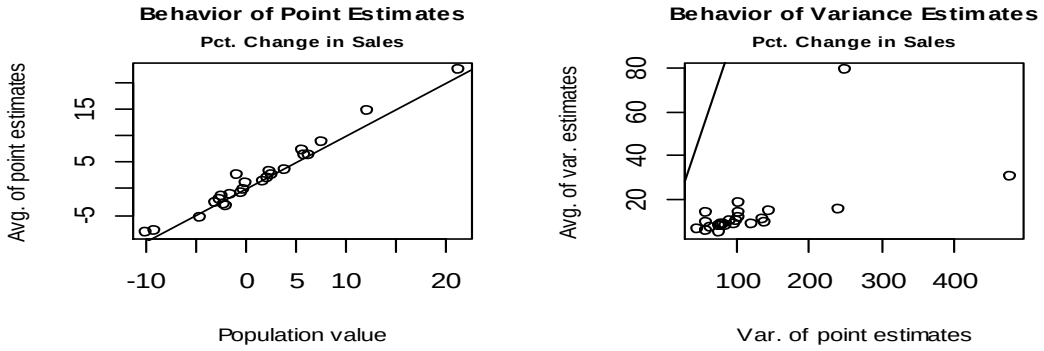

Figure 10. Average sample estimates (average variance estimates) for percent change in total sales after imputation by the "current method," plotted against known population values (variances of point estimates) at the national level across 24 months

In Figure 11, the x-axis is the average of the same variance estimates for the survey without missing data. The y-axis is the average of the variance estimates with missing data and imputation. The imputation procedure has shrunk the variance estimates by about 50%. The variance estimates were already too small, and the imputation procedure has made them much smaller. This distortion also has a negative impact on coverage. Without missing data, the average coverage of the confidence intervals for change at the national level (nominal 95% intervals) was only 73%. Now, with missing data and imputation, the average coverage has dropped to 52%. The imputation took a bad situation and made it worse.

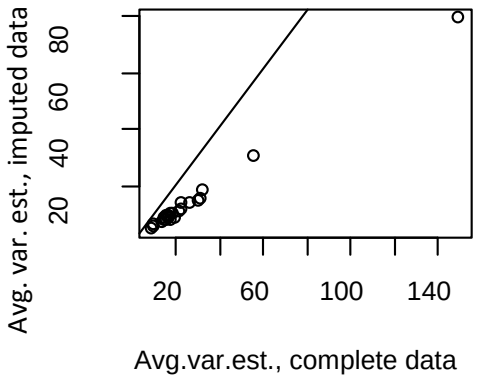

Figure 11. Average variance estimates for percent change in total sales under current imputation method versus average complete case variance estimates, national level for 24 months

5.3 Performance of MI Method

Finally, we describe what happens when we use the MI method. Boxplots for the estimated total sales are shown in Figure 12. In this case, the blue dots lie farther above the black line than before. The estimates of total sales from the MI method are biased upward, more so than with the current method. The average size of this bias is about 14% of the true population value. In terms of the bias, the MI method is not performing as well as the current method. Notice, however, that the center lines of the boxplots, which represent the medians of the sample estimates, are closer to the black line than they were before. The edges of the boxplots, which enclose the middle 50% of the estimates, are almost centered over the true values. Although the estimates now have greater bias, the "bulk" of the distribution for each estimator now lies closer to the truth than it did before, and the overall error (as would be measured, for example, by the mean squared error) has been reduced. This shift has a positive impact on the coverage of confidence intervals. Without imposing any missing values, the average coverage for these U.S. sales totals (nominal 95% intervals) was only 61%, because the sampling distributions were skewed. With the current imputation method, the average coverage dropped slightly to about 59%, but with MI the coverage goes up to about 81%. Although MI biases the point estimates, it biases them in a direction that actually helps. This may seem counterintuitive, but unusual happenings like these are possible with sampling distributions that are not normal.

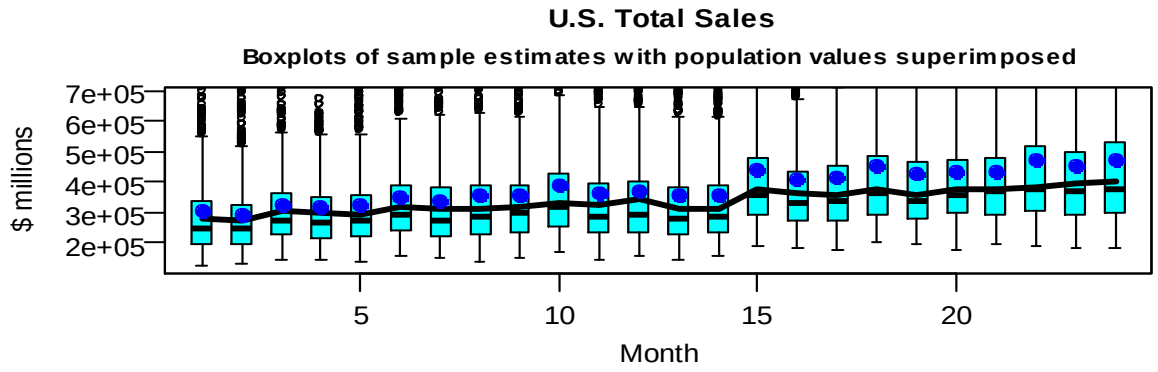

Figure 12. Boxplots of simulated sample estimates of total U.S. sales after imputation by the "MI method," with averages of the sample estimates (large blue dots) and known targets from the artificial population (thick black line) superimposed

6. Discussion

Building on our previous work (Lineback and Schafer, 2013), in which we identified potential problems with the current MWTS imputation procedure, we have now gathered more evidence that the current procedure is problematic. These simulations suggest that the current method does not substantially impair estimates for totals, but for percent changes, it appears to makes a bad situation worse. The current imputation method tends to deflate the standard errors for estimates of change. These standard errors would be too small even if there were no missing data. Under the rates and patterns of missing data that do occur in the MWTS, the current imputation method shrinks them further. It appears that our MI method might introduce bias into the change estimates, but it also appears to make them more precise, and the standard errors from the MI method are larger but more accurate.

At present, however, we do not suggest that the imputation procedures in the MWTS ought to be replaced with model-based MI. Now that we have constructed an artificial population and seen how the current MWTS estimators (HT and random group) perform over repeated samples with no missing data, we believe that attention should be turned to those complete-data procedures. What began as a missing-data problem has become an estimation problem. The sampling distributions of the point estimates are highly skewed and leptokurtic, and the variance estimates, although not necessarily biased, may behave in ways that data customers might not anticipate. In combination, the point and variance estimates produce confidence intervals that could be very misleading.

In light of these results, further comparison of imputation techniques seems less urgent. Instead, these results lead us to recommend re-directing this research towards investigating model-based and model-assisted methods, as well as Bayesian procedures, which may improve the quality of point and interval estimates. Some of these procedures operate on the observed values alone, with no imputation of missing values needed. The MWTS estimation procedures were developed nearly four decades ago (Wolter et al., 1976; Isaki et al., 1976), and since that time, a variety of new methodologies have become available. Thus, we are presented with an opportunity to modernize MWTS estimation procedures. As this line of research continues, we intend to take a more holistic approach, striving to make the best possible use of all the observed data by whatever methods we have in our statistical toolkit.

References

Brown, I. (2012). Modernization of benchmarking economic time series at the U.S. Census Bureau. Proceedings of the Federal Committee on Statistical Methodology.

Doove, L. L., Van Buuren, S. and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, 72, 92-104.

Isaki, C.T., Wolter, K.M., Sturdevant, T.R., Monsour, N.J. and Trager, M.L. (1976). Sample redesign of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Section of the American Statistical Association*, 90-98.

Lineback, J.F. and Schafer, J.L. (2013). Multivariate linear mixed-effects models for missing data applied to a business survey, presented at the 2013 Joint Statistical Meetings.

Raghunathan, T.R., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-96.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. and Yucel, R.M. (2002). Computational strategies for multivariate linear mixed models with missing values. *Journal of Computational and Graphical Statistics*, 11, 421-442.

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

Wolter, K.M., Isaki, C.T., Sturdevant, T.R., Monsour, N.J., and Mayes, F.M. (1976). Sample selection and estimation aspects of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Section of the American Statistical Association*, 99-109.

Van Buuren S., Brand, J.P.L., Groothuis-Oudshoorn, K. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 1049-1064