PLANT-LEVEL PRODUCTIVITY AND IMPUTATION OF MISSING DATA IN THE CENSUS OF MANUFACTURES

by

T. Kirk White *
Economic Research Service, U.S. Department of Agriculture

Jerome P. Reiter *
Duke University

and

Amil Petrin *
University of Minnesota, Twin Cities and NBER

CES 11-02 January, 2011

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Cheryl Grim, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K130B, 4600 Silver Hill Road, Washington, DC 20233, **CES.Papers.List@census.gov**.

Abstract

In the U.S. Census of Manufactures, the Census Bureau imputes missing values using a combination of mean imputation, ratio imputation, and conditional mean imputation. It is well-known that imputations based on these methods can result in underestimation of variability and potential bias in multivariate inferences. We show that this appears to be the case for the existing imputations in the Census of Manufactures. We then present an alternative strategy for handling the missing data based on multiple imputation. Specifically, we impute missing values via sequences of classification and regression trees, which offer a computationally straightforward and flexible approach for semi-automatic, large-scale multiple imputation. We also present an approach to evaluating these imputations based on posterior predictive checks. We use the multiple imputations, and the imputations currently employed by the Census Bureau, to estimate production function parameters and productivity dispersions. The results suggest that the two approaches provide quite different answers about productivity.

* The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center and the Minnesota Census Research Data Center. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau or of the U.S. Department of Agriculture's Economic Research Service. All results have been reviewed to ensure that no confidential information is disclosed. We thank Bert Grider and Arnie Reznek for disclosure avoidance reviews, Shawn Klimek for help in obtaining access to captured data from the Census of Manufactures, and Cheryl Grim for documenting and helping us understand item-level edit/impute ags in the Census data. Correspondence to: T. Kirk White, Economic Research Service, U.S. Department of Agriculture, 1800 M St., NW, Washington, D.C. 20036. email: kwhite@ers.usda.gov. Phone: (202) 694-5415. Fax: (202) 694-5756.

1 Introduction

Nearly all economic surveys suffer from item nonresponse, i.e., respondents answer some questions but not others. Statistical agencies that collect data frequently impute the missing values before making data available for secondary analyses. The manner of imputation can strongly impact secondary analyses of the completed data and, hence, affect public policy (Little and Rubin (2003)). For example, when the Census Bureau reported on its website that interstate migration declined sharply in 2006, the supposedly sharp decline in labor mobility prompted concern from Assistant Treasury Secretary Alan Krueger (Fletcher (2010)), and a report from the International Monetary Fund suggested that the observed steep decline in labor mobility was increasing unemployment (Batini, Celasun, Dowling, Estevao, Keim, Sommer, and Tsounta (2010)). However, Kaplan and Schulhofer-Wohl (2010) find that nearly all of the observed decline in annual interstate migration between 2005 and 2006 is attributable to a change in the way the Census Bureau imputes for missing data in the Current Population Survey.

In this article, we consider the impact of imputations for missing data on another topic of considerable academic and policy interest: what determines within-industry differences in total factor productivity? This is currently one of the most important questions in industrial organization, and its answer has implications for several other areas of economics, including macroeconomics, trade, and labor economics. A large literature has been devoted to investigating within-industry productivity differences, surveyed by Bartelsman and Doms (2000), and more recently by Syverson (forthcoming). As both reviews emphasize, measured within-industry productivity dispersion is large and persistent. Averaging across all U.S. manufacturing industries, Syverson (2004) finds that plants at the 90th percentile of the productivity distribution are nearly twice as productive as plants at the 10th percentile. Explanations for these observed within-industry productivity differences include management practices, the quality of labor and capital inputs, information technology, research & development, international trade, and regulation (Syverson (forthcoming)). We add another factor to

the list: imputed data. In fact, we show that there may actually be *more* within-industry productivity dispersion than the existing literature suggests.

We investigate the impacts of imputation using the U.S. Census Bureau's Census of Manufactures, which supports much of the empirical research on plant-level productivity. Although the Census of Manufactures represents the best available data for studying U.S. plant-level total factor productivity, imputations for nonresponse comprise a large percentage of the data; in fact, we show that this percentage is far more than what is reported in the existing literature. The Census Bureau imputes missing values using a combination of mean imputation, ratio imputation, and conditional mean imputation. Their primary goal is to facilitate point estimation of industry aggregates; however, it is not clear if these imputations are appropriate for multivariate analysis of microdata, such as estimating plant-level total factor productivity. Our investigations suggest that they may not be. Functions of key variables in the completed data show evidence of attenuation and under-estimation of variability. Additionally, estimates of production function parameters appear to be strongly impacted by the imputations, as do estimates of the within-industry dispersion of productivity.

What can be done about this missing/imputed data problem? One solution, popular among economists, is to drop plants with missing/imputed values, and only analyze the plants with complete data. Unfortunately, it is well-known that unless the missingness mechanism is missing completely at random (MCAR), complete case analysis can lead to biased parameter estimates (Little and Rubin (2003)). We find that the missing Census data are not MCAR, probably in part because the Census Bureau makes a greater effort to collect complete data from larger plants. Hence, complete-cases is not a trustworthy solution. Further, the impacts of imputations are not mitigated by focusing on certain industries or by using statistics that are robust to outliers. The imputations are pervasive, affecting many industries that have been studied previously.

As an alternative to these strategies, we create completed datasets via multiple imputation (Rubin (1987)). Multiple imputation has the potential to avoid problems that plague strategies like mean imputation, ratio impu-

tation, and conditional mean imputation. First, it draws imputed values from models that potentially condition on all variables in the data, which enables imputations to reflect multivariate relationships. In contrast, mean imputation fails to preserve any multivariate relationships, and ratio imputation at best preserves selected bivariate relationships used in the ratios. Conditional mean imputation can condition on all variables in the data and preserve multivariate relationships. However, like mean imputation and ratio imputation, conditional mean imputation ignores the stochastic nature of the data. This can result in under-estimation of variability. In contrast, multiple imputation can generate appropriately dispersed values. Finally, multiple imputation offers secondary analysts the potential for valid variance estimation in multivariate models, including regressions useful in productivity analysis. In contrast, single imputation procedures result in underestimation of uncertainty, because typically analysts treat the imputations as if they were genuine values. See Little and Rubin (2003) for further discussion of the benefits of multiple imputation over mean imputation, ratio imputation, and conditional mean imputation.

The key to the success of multiple imputation, particularly with large fractions of missing data, is the validity of the imputation model. Finding good fitting models is particularly challenging in the Census of Manufactures, as models that seem to work well in one industry may not in another; for example, conditioning on geographic region (e.g., because of differences in prices) may be important in some industries, but not others. Given the large number of industries and variables to be imputed, it is desirable to have imputation procedures that flexibly fit each variable in each industry with minimal tuning by the imputer.

Recognizing this, we impute missing items in the Census of Manufactures data using a sequence of classification and regression trees, as recently developed by Burgette and Reiter (2010). This method automatically handles mixed categorical and continuous data, works for skewed distributions like those in the Census of Manufactures data, and fits interactions and nonlinear relationships without parametric assumptions. The resulting multiple imputations lead to substantially different estimates of plant-productivity

than those based on the Census Bureau completed datasets, verifying that the method of imputation has a strong impact on conclusions about plant-level productivity. Further, given the documented deficiencies of imputation techniques like those used by the Census Bureau, the differential results suggest that improved imputation procedures like the one presented here would benefit users of the Census of Manufactures microdata.

2 Background on Plant-level Productivity

Conceptually, total-factor productivity (TFP) is how much output is produced from a given level of all measurable inputs. Plants with higher TFP produce more output from the same level of inputs, or the same output with lower levels of inputs. Syverson (forthcoming) reviews several ways of estimating plant-level TFP and the measurement issues inherent in each approach. Here we take a very common approach: we estimate a production function. Specifically, for each industry, we assume that the technology of every plant can be approximated by a 4-factor Cobb-Douglas production function:

$$lnQ_i = \beta_0 + \beta_k lnK_i + \beta_l lnL_i + \beta_e lnE_i + \beta_m lnM_i + u_i \tag{1}$$

where Q_i is the output of plant i, K_i is the capital stock, L_i is labor, E_i is energy, M_i is materials, and u_i is an error term. The u_i can include both productivity and measurement error in the dependent variable. We use the total value of shipments to measure output, so our measure of productivity also includes any within-industry differences in prices. For capital, we use the book value of assets. We measure labor in production-worker-equivalent hours: $L_i = SW_i \frac{PH_i}{WW_i}$, where SW are total salaries and wages, PH are production worker hours, and WW are production worker wages. Energy is the sum of the cost of fuels and the cost of purchased electricity. For materials, we use the total cost of materials less energy costs. We estimate equation (1) separately for each industry for 2007 by OLS and take the estimated residuals (plus the estimated intercept $\widehat{\beta}_0$) as our estimates of total factor productivity. It is well known that if plant managers know

the plant's productivity and take it into account when choosing inputs, the OLS estimator of the β parameter vector is biased (Marschak and Andrews (1944)). However, our goal here is to assess the effects of imputations on TFP estimates, not to compare production function estimators.

3 The Impact of Missing Data in the Census of Manufactures

The 2002 and 2007 Censuses of Manufactures are available to researchers via the Census Research Data Center network. They include roughly 300,000 manufacturing plants in each year. Plants with fewer than five employees, which account for about a third of the plants in the census, are not sent a survey form. Hence, most data for these plants are imputed from administrative records (AR). Following most researchers who use the Census of Manufactures, we drop all these AR cases. We also only include tabulated establishments in our sample, since non-tabulated establishments are known to have data that is of poor quality in some way. The final sample size is approximately 200,000 plants in each year.

Over the years, the Census of Manufactures has been plagued by item non-response. However, until the 2002 census, it was difficult to identify which, if any, items for a given plant were imputed due to item nonresponse, because item-level edit/impute flags were not made available to researchers. Dunne (1998) developed several clever ways to identify some of the imputed values, although the item-level flags that became available in the 2002 Census show that a much higher percentage of observations are imputed than are identified by Dunne's methods (White and Reiter (2008)). The item-level flags available in the 2002 and 2007 censuses contain codes which provide some information about how each item was imputed. However, in most cases, the definition of the codes is rather vague. We have not been able to obtain the computer code used to generate these imputations.

Table 1 presents the means and standard deviations of the within-industry imputation rates for all 6-digit NAICS industries. In identifying these records, we distinguish between edits or analyst corrections versus imputa-

tions.[1] It is clear that high percentages of data are imputed. For example, in both 2002 and 2007, for the average industry about 27% of the data on Total Value of Shipments (TVS) are imputed. For some other key variables, the mean imputation rate is even higher: for the average industry 42% of the Total Cost of Materials (CM) data are imputed in both years, and 37% and 38% of the Cost of Purchased Electricity (EE) data are imputed in 2002 and 2007, respectively. There is also significant variation in the imputation rates across industries. For these key variables, the standard deviation of the 6-digit NAICS level industry imputation rates range from 7 percentage points to 14 percentage points in 2002, and from 9 percentage points to 13 percentage points in 2007. This means, for example, that an industry that is one standard deviation above the mean in terms of its cost of materials imputation rate would have roughly 52% of its cost of materials data imputed in 2007.

To get some sense of how the Census Bureau's imputations might affect the relationships between key variables, we compute the following ratio for several input variables X:

$$R_X = \frac{IQR(\frac{X_{imp}}{TVS_{impX}})}{IQR(\frac{X_{obs}}{TVS_{obs}})} \tag{2}$$

where $IQR(Z)$ is the interquartile range of Z, X_{imp} represents imputed cases for the variable X, TVS_{impX} are the corresponding observations for the total value of shipments (which may be either imputed or observed), X_{obs} are observed cases for the variable X, and TVS_{obs} are the corresponding TVS observations. A ratio less than one is evidence that there is less dispersion in the imputed data than there is in the observed data. We compute these ratios for several inputs: capital (TAE), production worker hours (PH), the cost of materials (CM), the cost of electricity (EE), and the cost of fuels (CF). Table 2 presents the ratio of IQRs for the industries at the 25th, 50th, and 75th percentiles of the industry distributions. The results suggest that the Census Bureau's imputations tend to reduce the amount of

[1] In an appendix–available from the authors–we describe in greater detail how we identify imputed items.

within-industry variation in the ratios of key variables, in some cases quite drastically. For example, in 2002 when the book value of assets (TAE) is imputed, for the median 6-digit NAICS industry the IQR of the $\frac{TAE}{TVS}$ ratio is only 0.4 percent of the IQR of $\frac{TAE}{TVS}$ when both variables are observed. In 2007 the variation in the $\frac{TAE}{TVS}$ ratio when TAE is imputed is much more similar to the variation in the $\frac{TAE}{TVS}$ in the fully observed data. The item-level edit/impute flags indicate that in the 2007 Census, the Bureau changed the way it imputed for capital (TAE) in the majority of cases. This may account for the increase in the R_{TAE} ratio in 2007. However, in both years, for most industries, and for all of these key input variables, when a variable X is imputed, there is much less variation in the $\frac{X}{TVS}$ ratio than there is when X is observed. Since total factor productivity essentially measures the relationship between output and these inputs, it seems likely that estimates of productivity dispersion will be affected by the Census Bureau's imputations.

We next directly investigate the impact of imputation on estimates of plant-level productivity by estimating the model in equation (1) by OLS. To do so, we select a few detailed industries: fluid milk processing (NAICS 311511), coffee & tea manufacturing (NAICS 311920), flour milling (NAICS 311211), sugar manufacturing (NAICS 31131), ice manufacturing (NAICS 312113), soy bean processing (NAICS 311222), pesticides manufacturing (NAICS 32532), and fertilizer manufacturing (NAICS 32531). We select these industries for several reasons. Some of them (milk, ice, flour, soy beans) are relatively homogenous products, which should minimize within-industry differences due to product differentiation. In other words, for these industries we would think that the Census Bureau's relatively simple imputation methods would have a better chance of preserving the relationships in the data, since the products produced by each plant in an industry are relatively similar. A relatively high percentage of pesticides shipments are exports, allowing us to investigate the effect of imputation on the estimated relationship between productivity and international trade. Most of these industries are inputs into or use inputs from the agricultural sector, and thus are of interest to agricultural policymakers. Finally, nearly all of these

industries have been studied in previous research.[2] Table 3 shows the sample sizes and the imputation rates for key variables for each of our selected industries in the 2007 Census of Manufactures. Comparing Table 3 to Table 1, these industries are not atypical in terms of their imputation rates.

Table 4 presents the production function parameter estimates for selected industries for 2007. For several industries, the differences between the results based on the observed data and the completed data are economically significant. For example, for soy beans processing the output elasticity estimates for capital and energy are 0.15 and -0.08 in the fully observed data, but 0.04 and 0.06 in the Bureau-completed data. For pesticides manufacturing the estimated output elasticities for capital, labor, energy, and materials in the full observed data are, respectively, 0.23, 0.25, -0.03, and 0.49, respectively. In contrast, in the Bureau-completed data, the corresponding output elasticity estimates for the same industry are 0.09, 0.15, 0.15, and 0.54.

Since a relatively large fraction of pesticides shipments are exported, for the pesticides industry we also include a dummy for whether or not a plant exports some of its shipments. Previous research has found that exporters tend to be more productive than non-exporters (see, e.g., Bernard and Jensen (1999)). Although in this sample we do not overturn the Bernard and Jensen result, the coefficient estimate on the exporter dummy in the Bureau-completed data is only about two-thirds the estimate based on the fully observed data.

The next-to-last column of Table 4 shows the ratio of productivity at the 75th percentile of an industry's productivity distribution to productivity at the 25th percentile. For a few industries (coffee & tea, fertilizer, and ice), the Bureau's imputations seem to decrease the estimated 75-25 ratio substantially. However, this result is not universal. The final column shows

[2] Foster, Haltiwanger, and Syverson (2008) study productivity dispersion in coffee, ice, and sugar manufacturing; Davis, Grim, and Haltiwanger (2008) study the effect of electricity prices on measures of electricity productivity dispersion in the ice and coffee manufacturing industries, among other industries; Roberts, Klimek, and Dunne (2004) study entry and exit in the fluid milk industry; pesticides manufacturing has been studied by Ollinger and Fernandez-Cornejo (1995).

the ratio of the 90th percentile to the 10th percentile of productivity within each industry. For this measure, the Bureau-completed data produces substantially lower estimates of TFP dispersion for two industries (coffee & tea and pesticides), but substantially higher estimates of TFP dispersion for two other industries (flour and soy beans). Clearly, the Census Bureau imputations have an impact on productivity analyses.

4 Multiple Imputation using Classification and Regression Trees

Given the documented deficiencies with mean, ratio, and conditional mean imputation in the statistical literature, the results of the previous section suggest that one can improve on the imputation strategy being employed by the Census Bureau for the Census of Manufactures. We now describe our attempt to do so based on multiple imputation via sequential regression trees. We present only a broad overview of the approach here and refer the reader to Burgette and Reiter (2010) for details on the method.

Classification and regression trees (CART) seek to approximate the conditional distribution of a univariate outcome from multiple predictors (see Breiman, Friedman, Olshen, and Stone (1984), Hastie, Tibshirani, and Friedman (2009), and Ripley (2009)). The CART algorithm partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. The values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define the leaf.

The imputation process is done separately for each industry. We begin the process in any industry by filling in initial guesses at the missing data to create completed datasets for the industry; see Burgette and Reiter (2010) for an explanation of how to obtain initial guesses. Then, we order the variables in terms of increasing percentages of missing data. For the first

variable in this ordering with missing data, say Y_1, we fit the tree of Y_1 on all other variables, say Y_{-1}, so that each leaf contains at least k records; call this tree $\mathcal{Y}^{(1)}$. We use $k = 5$, which is a default specification in many applications of CART, to provide sufficient accuracy and reasonably fast running time. We grow $\mathcal{Y}^{(1)}$ by finding the splits that successively minimize the deviance of Y_1 in the leaves. We cease splitting any particular leaf when the deviance in that leaf is less than 0.00001 times the deviance in the marginal distribution of Y_1 or when we cannot ensure at least k records in each child leaf. For any record with missing data, we trace down the branches of $\mathcal{Y}^{(1)}$ until we find that record's terminal leaf. Let L_w be the wth terminal leaf in $\mathcal{Y}^{(1)}$, and let $Y_{L_w}^{(1)}$ be the n_{L_w} values of Y_1 in leaf L_w. For all records whose terminal leaf is L_w, we generate replacement values of Y_{ij} by drawing from $Y_{L_w}^{(1)}$ using the Bayesian bootstrap Rubin (1981). Repeating the Bayesian bootstrap for each leaf of $\mathcal{Y}^{(1)}$ results in an initial set of plausible values.

We next move to the second variable in the ordering with missing data, say Y_2. We fit the tree of Y_2 on all other variables, which we call $\mathcal{Y}^{(2)}$, using the newly completed values of Y_1. We run observations down $\mathcal{Y}^{(2)}$ to create plausible values for Y_2. The process continues for each Y_i in the ordering, each time using the newly imputed values of Y_{-i} to fit the tree and in locating leaves. We then cycle through this process five times to help move the trees away from the initial starting values. The end result is one completed dataset. We repeat this entire process m times to generate m completed datasets.

In the Census of Manufactures data, we delete (make missing) any Census-imputations identified by the item-level edit/impute flags, and run the sequential CART to create $m = 20$ completed datasets. The predictors for each tree include—whenever the variable is not the dependent variable—the total value of shipments, the total book value of assets, total salaries and wages, total employment, production worker wages, production worker hours, the number of production workers, the cost of purchased electricity, kilowatt hours of electricity, the cost of fuels, and the total cost of materials. We run the imputation procedure separately for each 5-digit or 6-digit NAICS industry.

Table 5 presents the production function parameter estimates and estimates of within-industry TFP dispersion based on datasets completed with the sequential CART method. The reported point estimates are the means of the parameter estimates and the means of the measures of TFP dispersion across the 20 implicates. For the production function parameter estimates, we compute the standard errors using Rubin (1987)'s combining formulas, which take into account the fact some of the data is imputed. For the measures of TFP dispersion we compute the standard deviation of the measures across the 20 estimates for each industry. For several industries (coffee & tea, fertilizer, sugar), the across-implicate standard deviation in the 90-10 ratio of TFP is substantial, indicating that missing data is responsible for a substantial amount of uncertainty about the amount of within-industry productivity dispersion in these industries.

Comparing Table 5 to Table 4, for most industries either the parameter estimates or the estimates of productivity dispersion–or both–differ substantially from the estimates based on the Census Bureau-completed data. For every industry, the sequential CART-completed data indicates that there is more within-industry TFP dispersion than the Bureau-completed data, and in some cases quite a lot more. For example, for sugar manufacturing, the average estimate of the 90-10 TFP ratio based on the CART-completed data is 131 percentage points higher than the estimate based on the Bureau-completed data; for coffee and tea manufacturing, the CART-completed average 90-10 TFP ratio is 98 percentage points higher than the 90-10 ratio based on the Bureau-completed data; for fertilizer the CART-completed average is 26 percentage points higher than estimate based on the Bureau-completed data; and for both fluid milk processing and ice manufacturing the CART-completed average 90-10 ratio is 26 percentage points higher.

Why are the TFP dispersion measures higher in the CART-completed data than they are in the Bureau-completed data? There are at least two possible reasons. First, as we saw in table 2, the relationship between total value of shipments and input variables has less variability in the Census Bureau's imputations than it does in the fully observed data. Since the CART imputations are essentially taking draws from the observed data (conditional

on a set of predictors), we might expect to see more variability in the relationship between TVS and the input variables in the CART-completed datasets, and thus more TFP dispersion. Second, missingness in Census of Manufactures is not completely at random (MCAR). In particular, smaller plants are more likely to have missing data. To the extent that plant size and productivity are correlated, imputation methods that fail to take this correlation into account will tend to reduce the amount of measured TFP dispersion.

To check the validity of our imputation models for these analyses, we use posterior predictive checks (He, Zaslavsky, Harrington, Catalano, and Landrum (2010)). Following Burgette and Reiter (2010), suppose that the n by k data matrix Y is arranged so that $Y = (Y_p|Y_c)$, where Y_p are the p partially observed columns of Y and Y_c are the remaining $k-p$ columns that are completely observed (payroll and total employment in our case). Let Y_{obs} denote the set of observed elements in Y, and let Y_{mis} denote the set of missing elements. For each industry, we use the CART method to create 500 pairs of datasets. The first dataset in each pair is a *completed* dataset, in which we create imputations for each element of Y_{mis}. To create the second dataset in each pair, we replace every element of Y_p, including elements that were not imputed in the original data. To do this, we take draws from the predictive distribution of Y_p conditional on Y_c using the tree fitted to create the first dataset in the pair. Let the second datasets in each pair be called the predicted datasets. We then use OLS to estimate the production function specified in (1) separately on each dataset. For each of the 500 pairs of datasets, we compute the differences between the parameter estimates from the completed dataset and those from the predicted dataset. Finally, for each parameter β_j, we compute a two-sided posterior predictive p-value:

$$P = \frac{2}{500} min\{\sum_{i=1}^{500} I(\widehat{\beta}_{imp,ij} - \widehat{\beta}_{pred,ij}), \sum_{i=1}^{500} I(\widehat{\beta}_{pred,ij} - \widehat{\beta}_{imp,ij})\} \qquad (3)$$

where $I(x)$ equals one if $x > 0$ and equals zero otherwise. Here, $\widehat{\beta}_{imp,ij}$ is the estimate of β_j for input j from the ith completed dataset, and $\widehat{\beta}_{pred,ij}$ is the

estimate from the ith predicted dataset. If the predicted data come from the same distribution as the completed data, we would expect $\widehat{\beta}_{imp,ij}$ to be higher than $\widehat{\beta}_{pred,ij}$ for about half the dataset pairs and lower than $\widehat{\beta}_{pred,ij}$ in the other half. A small p-value indicates that the $\widehat{\beta}_{pred,i}$ consistently differs from $\widehat{\beta}_{imp,i}$ in one direction. This would suggest that the imputation model does not adequately capture the relationships in the production function, and thus estimates based on the imputed data may be biased.

Table 6 presents the p-values for each production function parameter for selected industries in 2007. Basically, we find little evidence that our CART imputations are distorting the relationships between the variables in a way that would lead to biased production function parameter estimates. Although this does not confirm that the CART-based imputations result in the correct model, it does suggest that for at least this analysis the multiple imputation provides reasonable answers.

5 Conclusions and Suggestions for Further Research

Much of the literature on U.S. plant-level productivity uses the Census Bureau's Census of Manufactures. Even after dropping Administrative Records, a surprisingly large percentage of census data available to researchers is imputed. Our results suggest that these imputations have an economically significant effect on estimates of within-industry productivity dispersion. Using classification and regression trees, we provide a new set of imputations that seek to better preserve the joint distribution of key variables in the data and thus provide more accurate estimates of plant-level productivity dispersion and the relationship between productivity and other variables. The estimates of within-industry TFP dispersion using CART-completed data are often significantly higher than estimates based on the Census Bureau-completed data. These results suggest that there may be more within-industry productivity dispersion than the previous literature suggests. The existing literature provides a variety of explanations for within-industry productivity dispersion, including heterogeneity in management practices, the quality of labor and capital inputs, information tech-

nology, research and development, international trade, and regulation. To the extent that these factors are not part of the Census Bureau's imputation models (and they almost certainly are not), estimates of the effects of these factors on productivity dispersion in the Census data are probably biased. Researchers using the Census of Manufactures should consider how the Census Bureau's imputations may affect their estimates and consider alternative methods of imputation that try to preserve the key relationships in the data.

More broadly, as Kaplan and Schulhofer-Wohl (2010) illustrate, missing and imputed data can have a direct effect on policy discussions. As an increasing number of researchers conduct policy-relevant research using Census microdata made available via the expanding Census Research Data Center Network, this microdata will (hopefully) become increasingly important for policy debates. As a result it will be increasingly important for policymakers, researchers and the Census Bureau to understand how missing and imputed data affect estimates produced from these data.

References

Bartelsman, E. J., and M. Doms (2000): "Understanding Productivity: Lessons from Longitudinal Microdata," *Journal of Economic Literature*, 38(3), 569–594.

Batini, N., O. Celasun, T. Dowling, M. Estevao, G. Keim, M. Sommer, and E. Tsounta (2010): "United States: Selected Issues Paper," Working Paper 10/248, International Monetary Fund.

Bernard, A. B., and B. Jensen (1999): "Exceptional exporter performance: cause, effect, or both?," *Journal of International Economics*, 47(1), 1–25.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, FL.

Burgette, L., and J. Reiter (2010): "Multiple imputation for missing data via

sequential regression trees," *American Journal of Epidemiology*, 170(9), 1070–1076.

Davis, S., C. Grim, and J. Haltiwanger (2008): "Productivity Dispersion and Input Prices: The Case of Electricity," Working Papers 08-33, Center for Economic Studies, U.S. Census Bureau.

Dunne, T. (1998): "CES Data Issues Memorandum 98-1," CES data issues memorandum, Census Bureau Center for Economic Studies.

Fletcher, M. A. (2010): "Few in U.S. move for new jobs, fueling fear the economy might get stuck, too," *Washington Post*, July 30, p. A1.

Foster, L., J. Haltiwanger, and C. Syverson (2008): "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *American Economic Review*, 98(1), 394–425.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

He, Y., A. M. Zaslavsky, D. P. Harrington, P. Catalano, and M. B. Landrum (2010): "Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide," *Statistical Methods in Medical Research*, 19(6), 653–670.

Kaplan, G., and S. Schulhofer-Wohl (2010): "Interstate Migration Has Fallen Less Than You Think: Consequences of Hot Deck Imputation in the Current Population Survey," Working Paper 16536, National Bureau of Economic Research.

Little, R., and D. Rubin (2003): *Statistical Analysis with Missing Data, Second Edition*. John Wiley, New York.

Marschak, J., and W. Andrews (1944): "Random Simultaneous Equations and the Theory of Production," *Econometrica*, 12(3–4), 143–205.

Ollinger, M., and J. Fernandez-Cornejo (1995): "Innovation and Regulation in the Pesticide Industry," Working Papers 95-14, Center for Economic Studies, U.S. Census Bureau.

Ripley, B. (2009): "Tree: classification and regression trees," cran.r-project.org.

Roberts, M., S. Klimek, and T. Dunne (2004): "Entrant Experience and Plant Exit," Working Papers 04-12, Center for Economic Studies, U.S. Census Bureau.

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

Rubin, D. B. (1981): "The Bayesian bootstrap," *The Annals of Statistics*, 9, 130–134.

Syverson, C. (2004): "Product Substitutability and Productivity Dispersion," *The Review of Economics and Statistics*, 86(2), 534–550.

——— (forthcoming): "What Determines Productivity?," *Journal of Economic Literature*.

White, T. K., and J. P. Reiter (2008): "Multiple Imputation in the Annual Survey of Manufactures," in *2007 Research Conference Papers*, Washington, D.C. Federal Committee on Statistical Methodology, Office of Management and Budget.

Table 1: Imputation Rates for Key Variables At 6-digit NAICS Industry Level 2002 and 2007 Censuses of Manufactures

year	Statistic	Total Value of Shipments	Book Value of Assets	Production Worker Hours	Cost of Purchased Electricity	Cost of Fuels	Cost of Materials
2002	Mean	0.27	0.31	0.19	0.38	0.37	0.42
	s.d.	0.09	0.10	0.07	0.14	0.14	0.10
2007	Mean	0.27	0.32	0.31	0.37	0.35	0.42
	s.d.	0.09	0.10	0.13	0.13	0.12	0.10

The table shows the means and standard deviations of 6-digit NAICS industry-level imputation rates. The imputation rate is the percentage of tabulated non-Administrative Records cases that are imputed (not just edited) by the Census Bureau.

Table 2: Distribution of Ratios of Within-Industry Interquartile Ranges of Ratios of Key Variables in Imputed Data vs. Fully Observed Data (equation 2)

percentile	Book Value of Assets	Production Worker Hours	Cost of Purchased Electricity	Cost of Fuels	Cost of Materials
			2002		
25th	0.002	0.159	0.062	0.088	0.036
50th	0.004	0.293	0.112	0.174	0.208
75th	0.018	0.522	0.219	0.356	0.456
			2007		
25th	0.216	0.353	0.088	0.152	0.089
50th	0.369	0.486	0.179	0.370	0.262
75th	0.565	0.704	0.326	0.782	0.478

The table shows the 25th, 50th and 75th percentiles of the ratios of the within-6-digit-NAICS industry interquartile ranges (IQR) of inputs-to-TVS ratios in the imputed data relative to the IQR of the same ratio in the same industry in the fully observed data. A ratio less than one indicates that there is less dispersion in the input-to-TVS ratio in the imputed data for a given industry than in the fully observed data for the same industry.

Table 3: Imputation Rates for Key Variables, 2007 Censuses of Manufactures, Selected Industries

industry	Sample Size	Total Value of Shipments	Book Value of Assets	Production Worker Hours	Cost of Electricity Purchased	Cost of Fuels	Cost of Materials
coffee & tea	186	0.31	0.38	0.35	0.45	0.43	0.55
fertilizer	472	0.29	0.38	0.35	0.39	0.36	0.46
flour	210	0.22	0.20	0.18	0.21	0.20	0.31
fluid milk	362	0.13	0.18	0.17	0.20	0.20	0.32
ice	295	0.23	0.23	0.21	0.25	0.24	0.31
pesticides	196	0.24	0.30	0.27	0.40	0.39	0.48
soy beans	89	0.13	0.25	0.17	0.33	0.33	0.38
sugar	73	0.41	0.42	0.27	0.41	0.38	0.45

The table shows imputation rates for each 5- or 6-digit NAICS industry. The imputation rate is the percentage of tabulated non-Administrative Records cases that are imputed (not just edited) by the Census Bureau.

Table 4: Production Function Parameter Estimates And Productivity Dispersion, Selected Industries, 2007 Census of Manufactures

	Production Function Coefficients					TFP ratios	
						75-25	90-10
	Capital	Labor	Energy	Materials	Exports	TFP	TFP
industry	β_k	β_l	β_e	β_m	Dummy	ratio	ratio
(a) Fully Observed Data							
coffee & tea	0.08	0.30	0.04	0.61		1.68	2.39
fertilizer	0.04	0.22	0.11	0.60		1.52	2.27
flour	0.02	0.13	0.11	0.72		1.20	1.42
fluid milk	0.05	0.15	0.05	0.70		1.28	1.82
ice	0.04	0.34	0.32	0.38		1.56	2.17
pesticides	0.23	0.25	-0.03	0.49	0.32	1.74	3.56
soy beans	0.15	0.11	-0.08	0.85		1.28	1.61
sugar	0.13	0.21	0.05	0.60		1.33	1.89
(b) Census Bureau-Completed Data							
coffee & tea	0.14	0.29	-0.02	0.62		1.36	2.08
fertilizer	0.05	0.21	0.13	0.58		1.40	2.15
flour	0.05	0.11	0.08	0.75		1.20	1.59
fluid milk	0.04	0.18	0.12	0.66		1.30	1.77
ice	0.05	0.31	0.30	0.39		1.40	2.13
pesticides	0.09	0.15	0.15	0.54	0.22	1.69	3.10
soy beans	0.04	0.11	0.06	0.84		1.25	1.84
sugar	0.15	0.19	0.08	0.52		1.37	1.83

Census Bureau-completed data include both fully observed cases and cases for which some variables are observed and other variables are imputed by the Census Bureau.

Table 5: Production Function Parameter Estimates And Productivity Dispersion, Selected Industries, 2007 Census of Manufactures, CART-completed Data

		Production Function Coefficients				TFP ratios	
industry	Sample Size	Capital β_k	Labor β_l	Energy β_e	Materials β_m	75-25, mean (s.d.)	90-10, mean (s.d.)
coffee & tea	186	0.083	0.292	0.088	0.559	1.78	3.06
		(0.043)	(0.105)	(0.050)	(0.078)	(0.13)	(0.90)
fertilizer	472	0.067	0.190	0.125	0.559	1.54	2.41
		(0.025)	(0.038)	(0.021)	(0.029)	(0.04)	(0.22)
flour	210	0.055	0.122	0.116	0.687	1.27	1.70
		(0.030)	(0.044)	(0.057)	(0.056)	(0.03)	(0.07)
fluid milk	362	0.069	0.182	0.109	0.613	1.36	1.99
		(0.038)	(0.045)	(0.039)	(0.073)	(0.05)	(0.11)
ice	295	0.070	0.340	0.343	0.287	1.59	2.35
		(0.031)	(0.060)	(0.034)	(0.039)	(0.03)	(0.08)
soy beans	89	0.065	0.091	0.033	0.826	1.48	1.95
		(0.093)	(0.098)	(0.052)	(0.047)	(0.04)	(0.12)
sugar	73	0.139	0.170	0.146	0.447	1.69	3.14
		(0.081)	(0.157)	(0.095)	(0.143)	(0.14)	(0.52)

The table shows the means across 20 CART-completed datasets of production function parameters and total factor productivity (TFP) dispersion. Standard errors are in parentheses. The standard errors of the estimates from each of the 20 implicates are combined using Rubin's (1987) combining formulas. For the TFP dispersion measures, the standard deviations of the TFP dispersion estimates across the 20 implicates are shown in parentheses.

industry	Capital β_k	Labor β_l	Energy β_e	Materials β_m
coffee & tea	0.696	0.204	0.500	0.976
fertilizer	0.736	0.748	0.500	0.716
flour	0.972	0.540	0.944	0.892
fluid milk	0.708	0.340	0.928	0.724
ice	0.316	0.460	0.852	0.380
pesticides	0.980	0.576	0.980	0.628
soy beans	0.840	0.492	0.920	0.980
sugar	0.928	0.988	0.820	0.828

Table 6: **Posterior Predictive P-Values for Production Function Parameter Estimates, Selected Industries, 2007 Census of Manufactures, CART-completed Data vs. CART-predicted Data.** The p-values indicate whether or not the parameter estimates from the CART-completed datasets consistently deviate from the estimates from the CART-predicted datasets, based on 500 pairs of completed datasets and predicted datasets for each industry.