Plant-level Productivity and Imputation of Missing Data in U.S. Census Manufacturing Data *

by
T. Kirk White
Economic Research Service, U.S. Department of Agriculture
Jerome P. Reiter
Duke University
Amil Petrin
University of Minnesota,Twin Cities and NBER

January 30, 2012

*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, the Economic Research Service, or the U.S. Department of Agriculture. All results have been reviewed to ensure that no confidential information is disclosed. We thank Randy Becker, Bert Grider, Cheryl Grim, John Haltiwanger, Shawn Klimek, Arnie Reznek, Pat Sullivan and participants in the Census Bureau Center for Economic Studies seminar for their comments. Reiter gratefully acknowledges support from National Science Foundation grant SES 1131897. Correspondence to: T. Kirk White, Economic Research Service, U.S. Department of Agriculture, 355 E St., SW, Washington, D.C. 20024-3221. email: kwhite@ers.usda.gov. Phone: (202) 694-5415. Fax: (202) 694-5756.

Abstract

Within-industry differences in measured plant-level productivity are large. A large literature has been devoted to explaining these differences. In the U.S. Census Bureau's manufacturing data, the Bureau imputes for missing values using methods known to result in underestimation of variability and potential bias in multivariate inferences. We present an alternative strategy for handling the missing data based on multiple imputation via sequences of classification and regression trees. We use our imputations and the Bureau's imputations to estimate within-industry productivity dispersions. The results suggest that there may be more within-industry productivity dispersion than previous research has indicated.

JEL codes: L60, C80, L11 Key words: Plant-level productivity; multiple imputation; missing data; manufacturing.

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 for 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, as noted by Kaplan and Schulhofer-Wohl (2010), 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 then-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 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 paper, 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 (2011). 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 (2011)). 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 (CMF) and Annual Survey of Manufactures (ASM), which support much of the empirical research on plant-level productivity. Although the CMF and ASM represent 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 imputation, 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 under-estimation 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 and the ASM, 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 de-

sirable 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 CMF and ASM 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 manufacturing data, and fits interactions and non-linear relationships without parametric assumptions. The resulting multiple imputations lead to substantially different estimates of plant-level 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 and ASM 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 (2011) 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 estimate equation (1) using both cross-sectional data and panel data. For both types of data we measure labor in production-worker-equivalent hours: $L_i = SW_i * PH_i/WW_i$, where SW are total salaries and wages, PH are production worker hours, and WW are production worker wages. When we use cross-sectional data, our other variables are nominal values. We use the total value of shipments to measure output, so our measure of productivity also includes any within-industry differences in prices. 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 2002 and 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 param-

eter vector is biased (Marschak and Andrews (1944)). Olley and Pakes (1996) develop an estimator to address this endogeneity issue using investment as a proxy. Zero investment–a common feature of plant-level data–causes identification problems for the Olley-Pakes estimator, so Levinsohn and Petrin (2003, LP hereafter) develop a similar estimator using intermediate inputs as a proxy. Wooldridge (2009, WLP hereafter) develops a version of the LP estimator that is robust to the critique of Ackerberg, Caves, and Frazer (2006).

Both the WLP and LP estimators require panel data. To create a panel of plants using only the 2002 and 2007 Censuses, we would need 6-digit NAICS industry-level price indexes for inputs and outputs to deflate the nominal values in the Census data to real values. Unfortunately such indexes are not yet available for 2007. However, we do have industry-level deflators for 2002-2005, and for 2003-2005 we have a sample of plants from the Annual Survey of Manufactures (ASM). To construct real values, we deflate our nominal measure of output, energy, and materials by the corresponding industry deflators. We construct real capital stocks from deflated initial book values and deflated investment expenditures using the perpetual inventory method, as described in Petrin, White, and Reiter (2011). For our 2002-2005 panel we estimate production functions using OLS, LP and WLP (2009).

3 The Impact of Missing Data in the Census of Manufactures and the Annual Survey of Manufactures

The quinquennial Censuses of Manufactures and the Annual Survey of Manufactures are available to researchers via the Census Research Data Center network. The Censuses 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. Our final sample size from the Censuses is approximately 200,000 plants in each year. The Annual Survey of Manufactures is an annual sample of roughly 50,000 to 75,000 plants. Large plants–usually defined as having more than 250 employees–are included in the sample with certainty, and smaller plants a sampled with a probability that primarily depends on the plant's size.

Over the years, the CMF and the ASM have been plagued by item nonresponse, and the Census Bureau has created imputations for this missing data. However, until the 2002 census, it was difficult for researchers to iden-

tify which, if any, items for a given plant were imputed due to item nonresponse, because item-level edit/impute flags were not made available. 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 and the 2003-2006 ASMs 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 key variables for all 6-digit NAICS industries from the 2002 and 2007 Censuses and the 2003-2006 ASMs. The book values of assets are collected in the Census years, but not in the ASM in these years, so for 2003-2006 we report imputation rates for total capital expenditures instead of assets. In identifying these records, we distinguish between edits or analyst corrections versus imputations.[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

[1]In the appendix we describe in detail how we identify imputed items.

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.

With the exception of total value of shipments and production worker hours, imputation rates tend to be lower in the ASM years than in the Census years. A higher percentage of the ASM samples are large plants, and the Census Bureau puts more effort into collecting complete records from large plants. The means and standard deviations of the industry imputation rates in the ASM are fairly stable over time.

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). Tables 2 and 3 present 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 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 TAE/TVS ratio is only 0.4 percent of the IQR of the TAE/TVS ratio when both variables are observed.

In the ASM years 2003-2006 (table 3), the ratios of IQRs are much higher than in the 2002 Census. This probably reflects the fact that the ASM samples include fewer small plants. As a result, compared to the Census data, a higher percentage of the ASM plants with missing data are similar to ASM plants with complete data. Thus in the ASM data the Census Bureau's imputations are able to do a better job of reproducing the within-industry dispersion in input-to-TVS ratios that we see in the complete data.

In 2007 the variation in the TAE/TVS ratio when TAE is imputed is much more similar to the variation in the TAE/TVS ratio 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 compared to 2002. However, in both Census 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 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.[2] To do so, we select a few detailed industries: coffee & tea manufacturing (NAICS 311920), fertilizer manufacturing (32531), flour milling (311211), ice manufacturing (312113), fluid milk processing (311511), pesticides manufacturing (32532), soy bean processing (311222), and sugar manufacturing (31131). We select these industries for several reasons. Some of them (fluid milk, ice, flour, soy beans) are relatively homogenous products, which should minimize within-industry differences due to product differentiation. 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

[2]Below, we check the robustness of our results using alternative production function estimators.

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.[3] Tables 4 and 5 show the sample sizes and imputation rates for key variables for each of our selected industries in, respectively, the 2002 and 2007 Censuses of Manufactures. The imputation rate for production worker hours increased significantly in 2007 in most of our selected industries increased significantly in 2007 compared to 2002. Other than production worker hours there is no clear pattern–imputation rates increased for some variables in some industries and declined for others. Comparing tables 4 and 5 to the imputation rates for 2002 and 2007 in table 1, most of our selected industries have imputation rates below the mean for most of these key variables.

For most of our selected industries, we do not have enough observations in the ASM years 2003-2006 to pass Census Bureau disclosure avoidance rules or produce reliable estimates. However, for one of our industries–fluid milk

[3]Foster, Haltiwanger, and Syverson (2008) study productivity dispersion in coffee, ice, and sugar manufacturing (as well as other industries); 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).

processing–we do have enough observations. For this industry, we construct a panel. Table 6 presents the imputation rates for key variables for fluid milk processing plants in the 2003-2006 Annual Surveys of Manufactures. Interestingly, the imputation rates increased significantly in 2004, remained high in 2005 and 2006, and dropped back to the 2003 levels in the 2007 Census (see table 5 for the 2007 rates). This pattern may be related to the fact that the probability sample portion of the ASM panel rotates out in 2003, and a new probability sample begins in 2004.

The first 4 columns of table 7 present OLS estimates of the production function parameter estimates from equation (1) for selected industries for 2002, based on the fully-observed (non-imputed) data. Table 8 shows the estimates based on the Census Bureau-completed data, which includes the fully-observed data as well as the Bureau's imputations. 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.[4] The next-to-last columns of tables 7 and 8 show the ratio of productivity at the 75th percentile of an industry's productivity distribution to productivity at the 25th percentile, and the final columns shows the ratio of the 90th percentile to the 10th percentile of productivity within each industry.

Not surprisingly, the industries with the highest imputation rates–coffee

[4]Previous research has found that exporters tend to be more productive than non-exporters (see, e.g., Bernard and Jensen (1999)).

& tea and sugar–exhibit the largest changes in their coefficient estimates. The point estimate for the exporter dummy switches from positive in the fully-observed data to negative in the Bureau-completed data, although neither estimate is statistically significant. The Bureau's imputations might affect the sign of the exporter dummy because exporters tend to be larger (and more productive), and large plants are less likely to have missing data.

The Bureau's imputations also affect estimates of within-industry TFP dispersion, although not always in the same direction. For coffee & tea, ice, and pesticides, the estimated 75-25 ratio is economically significantly lower in the Bureau-completed data, but for fertilizer and soybeans, the ratio is lower in the fully-observed data. Comparing the 90-10 TFP ratios, the Bureau's imputations sometimes make a quite a difference. For example, the pesticides manufacturing plant at the 90th percentile of the TFP distribution in the fully-observed data is 3.2 times as productive as the plant at the 10th percentile, while in the Bureau-completed data, the 90th percentile plant is "only" 2.6 times as productive.

Tables 9 and 10 present the production function parameter estimates for the same industries for 2007. Again, for several industries–especiallly fertilizer and soybeans–the differences between the results based on the fully-observed data and the completed data are economically significant. Although in the 2007 data exporters are more productive than non-exporters in both samples, 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 Bureau's imputations also affect estimates of within-industry productivity dispersion in 2007 using both the 75-25 ratio and the 90-10 ratio.

Table 11 presents the production function parameter estimates from our fluid milk panel. Column 1 shows OLS estimates of the production function coefficients, and TFP dispersion in the fully-observed data. For the coefficients, we pool observations across 2002-2005, but we estimate TFP dispersion within each year. For the ASM years 2003-2005, we compute the ASM-sample-weighted productivity distributions. Columns 2 and 3 present the estimates from the LP and WLP estimators, respectively. Columns 4-6 show the OLS, LP, and WLP estimates based on the Bureau-completed data.

Interestingly, the Bureau's imputations affect the OLS estimates much less than the LP and WLP estimates. For example, the WLP estimate of the energy coefficient falls from an incredible 1.93 in the fully-observed data to 0.20 in the Bureau-completed data, while the OLS estimates are quite similar in both samples. The LP and WLP estimators also tend to produce higher estimates of TFP dispersion than OLS whether we use only the fully-observed data or the Bureau-completed data. The LP and WLP estimates of productivity dispersion are also much greater in the fully-observed data than in the Bureau-completed data. To summarize, the Census Bureau's imputations have an important impact on productivity analyses whether we use OLS or estimators which try to account for the endogeneity of inputs.

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 and the ASM. 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 ten 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 CMF and ASM 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. For each of our industries, 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.

For two of our industries we include additional variables in the imputation model. A significant fraction of pesticides sales are exports, and exporters tend to be more productive (Bernard and Jensen (1999)). So in the pesticides industry imputation model we include a binary variable indicating whether or not the plant exports. For our fluid milk panel, in addition to the common set of predictors we also include a plant identifier, the year, and total capital expenditures. Since the book value of assets is not reported in ASM years, we initialize it for the imputation algorithm by added the plant's capital expenditures to its book value from 2002. Note that these are just initial values to be used by the imputation algorithm. When the imputations are completed we construct real capital stocks from the plant's initial book values

and investment expenditures using the perpetual inventory method.

Table 12 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 are imputed. For the measures of TFP dispersion we compute the standard deviation of the measures across the 20 estimates for each industry. For most of our selected industries the across-implicate standard deviation in the 90-10 ratio of TFP is substantial, indicating that missing data are responsible for a substantial amount of uncertainty about the degree of within-industry productivity dispersion in these industries.

Comparing table 12 to table 8, 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, in sugar manufacturing the 90-10 TFP ratio based on the CART-completed data is more than three times the estimate based on the Bureau-completed data, and the 75-25 TFP ratio from the CART-completed data is 60 percentage points higher.

The sugar industry has a relatively small sample size (83 plants), and the between-imputation standard deviation is quite high (8.1). However, even for industries with larger samples and lower between-imputation variances, such as ice manufacturing, the estimated within-industry TFP dispersion in the CART-completed data is economically signficantly higher than the estimate from the Bureau-completed data.

Table 13 presents the same statistics as 12, using the 2007 Census data. Comparing table 13 to table 10, the results are similar to those from the 2002 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.

Table 14 presents the means and between-imputation standard deviations of our estimates of production function parameters for the fluid milk industry in 2002-2005 based on 20 CART-completed panel datasets. Column 1 presents the OLS estimates. Columns 2 and 3 present the LP and WLP estimates. The first thing to notice is that, just as we saw in table 11, the LP and WLP estimators are more sensitive to the imputed data than the OLS estimator–the between-imputation standard deviations of the LP and WLP estimates tend to be much larger.

Table 15 presents the means and between-imputation standard deviations of the OLS, LP, and WLP estimates of within-industry productivity dispersion for fluid milk for each year, 2002-2005. Comparing table 15 to table 11, the mean OLS estimates of TFP dispersion in the CART-completed

are slightly larger than the OLS estimates from the Bureau-completed data. The mean LP estimates of TFP dispersion are sometimes slightly larger than those from the Bureau-completed data, and sometimes slightly smaller.[5] The mean WLP estimates of TFP dispersion from the CART-completed data are much higher than the estimates from the Bureau-completed data.

Why are the TFP dispersion estimates from the CART-completed data often higher than the estimates from the Bureau-completed data? There are at least two plausible 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 taking draws from the observed data (conditional on a set of predictors), we might expect to see more variability in the relationship between output and the input variables in the CART-completed datasets, and thus more measured TFP dispersion. Second, missingness in Census of Manufactures and the ASM are 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.

[5]For the LP estimator, we could not estimate the capital and energy output elasticities for 5 of our 20 datasets. The means of these (15) capital and energy coefficient estimates are substantially higher than the means of the (20) estimates from the OLS and WLP estimators.

5 Validity Checks

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. 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 θ_j, we compute a two-sided posterior predictive p-value:

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

where $I(x)$ equals one if $x > 0$ and equals zero otherwise. Here, $\widehat{\theta}_{imp,ij}$ is the estimate of parameter θ_j–a regression coefficient or TFP dispersion measure–from the ith completed dataset, and $\widehat{\theta}_{pred,ij}$ is the estimate from the ith predicted dataset. If the predicted data come from the same distribution as the completed data, we would expect $\widehat{\theta}_{imp,ij}$ to be higher than $\widehat{\theta}_{pred,ij}$ for about half the dataset pairs and lower than $\widehat{\theta}_{pred,ij}$ in the other half. A small p-value indicates that the $\widehat{\theta}_{pred,i}$ consistently differs from $\widehat{\theta}_{imp,i}$ in one direction. This would suggest that the imputation model does not adequately capture the relationships in the production function or the TFP dispersion in the data, and thus estimates based on the imputed data may be biased.

Tables 16 17 present the p-values for each production function parameter and one productivity dispersion estimate for selected industries in 2002 and 2007, respectively. With one or two exceptions, we find no evidence that our CART imputations are distorting the relationships between the variables in a way that would lead to biased estimates of production function parameters or within-industry productivity dispsersion.[6] Table 18 presents posterior predictive p-values for our OLS, LP, and WLP estimates for the fluid milk panel. Again, we find little evidence of bias from the imputation model. 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 CART-based multiple imputation provides reasonable answers.

[6]The exceptions are the 75-25 TFP ratios for flour in 2002 and 2007, which have posterior predictive p-values of 0.096 and 0.12, respectively.

6 Conclusions and Suggestions for Further Research

Much of the literature on U.S. plant-level productivity uses the Census Bureau's Census of Manufactures or the Annual Survey of Manufactures (ASM). Even after dropping Administrative Records, a surprisingly large percentage of the Census and ASM 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 technology, 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 or the ASM 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, the Census Bureau, and other statistical agencies to understand how missing and imputed data affect estimates produced from these data.

A Identifying Imputed Data

In this appendix we describe how we identify an element in the data matrix as imputed or not. As part of its edit and imputation process, the Census Bureau sets item-level edit/impute flags for each item on the survey. For most observations, we use the item's edit/impute flag variables to determine whether or not an item was imputed. To try to asses the accuracy of the edit/impute flags are, we also obtained access to the data file for the 2002 Census of Manufactures after only minimal editing at the Census Bureau's survey processing center and before the Census Bureau's main editing and

imputation processing. We call this the "captured data." We refer to the dataset resulting from the Census Bureua's edits and imputations the "final data."[7] In the vast majority of cases, if an imputation flag is set, the value for the item in the final dataset differs from the captured data item. However, in some cases the captured data is the same as the final data even though an impute flag is set. Therefore, for 2002 we define an observation as imputed if it meets the criteria below based on the edit/impute flags *and* the value in the final data is different from the value for the same observation in the captured data. For the 2003-2006 Annual Surveys of Manufactures and the 2007 Census of Manufactures, we do not have access to the captured data, so we rely solely on the edit/impute flags.

In our CART imputation models, for all industries we include the following plant-level variables: total value of shipments (TVS), total cost of materials and parts (CM), total cost of fuels (CF), cost of purchased electricity (EE), quantity of purchased electricity (PE), total book value of assets at the beginning of the year (TAB), total salaries and wages (SW), total employment (TE), production worker hours (PH), production worker wages (WW), and the number of production workers (PW). As described in the main text, we include additional variables in the imputation models for pesticides manufacturing and fluid milk processing. However, these additional variables are never imputed, so we do not discuss them here.

[7]The final data is available to researchers with approved projects via the Census Research Data Centers.

Each edit/impute flag consists of two or three characters. The first character is either a blank, indicating that the item was not reported on the survey form, or an 'R', indicating that it was reported. The second and (if applicable) third characters take one of 22 values. Table A1 list the 22 codes (including blank) and the names of each code. Table A2 briefly describes when each code is set. Every variable on the survey forms for the CMF and ASM has a corresponding edit/impute flag with some combination of these codes. For example, if total value of shipments (TVS) for a particular plant is reported on the survey form and not edited or imputed, then the edit/impute flag for TVS for that plant will be 'R ', indicating that the TVS value in the final dataset was reported on the survey form and was not edited or imputed. The third column of table A1 shows the Census Bureau categorization of each of these codes as either imputed or non-imputed. For example, if a data item is corrected by a Census Bureau analyst (code C), that item is not considered to be imputed.

In general, we define an item as imputed if the second or third character in its edit/impute flag is in the "imputed" category. We make an exception to this rule for the capital stock variables. In many cases the edit/impute flags for capital variables–total book values of assets beginning of year (TAB) and end of year (TAE)–and capital expenditures (TCE) are set to ' K'. The blank first character means that the item was not reported on the survey form. The K supposedly means that the sum of a set of detail items do not balance to a total, so the detail items are changed proportionally to correct

the imbalance. In the case of capital stock variables, TAB plus TCE should sum to TAE minus depreciation. However, in 2002 we find that for many plants the flags indicate that *none* of these capital variables was reported on the survey form and all of them were "raked." Since it is impossible to adjust to a total that was not reported, we treat these items as imputed.

The Census Bureau uses different imputation methods for different variables. For example, the "industry average" method is used frequently for the energy input cost variables (cost of fuels, cost of purchased electricity), but only rarely for total value of shipments. On the other hand, the "Beta (Cold Deck Statistical)" method is used frequently to impute for total value of shipments and the total cost of materials. Note that although the edit/impute flags tell us what general method was used to impute each data element, we still do not know exactly how each element was imputed. For example, if the edit/impute flag for a plant's cost of purchased electricity is set to ' V', we know that the plant's electricity costs are set to the industry average by ratio imputation, but we do not know what the denominator of the ratio is. Similarly, a flag set to ' B' ("Cold Deck Statistical") means that the item was imputed using a regression model based on historical data. However, we do not know what sample was used for this regression or even what explanatory variables are in the regression model. One of the advantages of our imputations versus the Census Bureau's imputations is transparency–in the main text of this article we provide a detailed description of our imputation method, including all the variables in our imputation models.

References

Ackerberg, D., K. Caves, and G. Frazer (2006): "Structural Identification of Production Functions," Working Paper.

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.

Grim, C. (2011): "User Notes for 2002 Census of Manufactures," Unpublished Technical Note.

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.

Levinsohn, J., and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 70(2), 341–372.

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.

Olley, S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1298.

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.

Petrin, A. K., T. K. White, and J. P. Reiter (2011): "The impact of plant-level resource reallocations and technical progress on U.S. macroeconomic growth," *Review of Economic Dynamics*, 14(1), 3–26.

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.

——— (2011): "What Determines Productivity?," *Journal of Economic Literature*, 49(2), 326–365.

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.

Wooldridge, J. M. (2009): "On estimating firm-level production functions using proxy variables to control for unobservables," *Economics Letters*, 104(3), 112–114.

Table 1: Imputation Rates for Key Variables At 6-digit NAICS Industry Level, 2002 and 2007 Censuses of Manufactures and 2003-2006 Annual Surveys of Manufactures

year	Statistic	Total Value of Shipments	Book Value of Assets/ Capital Expenditures	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
2003	mean	0.27	0.31	0.28	0.30	0.26	0.35
	s.d.	0.12	0.13	0.13	0.13	0.11	0.13
2004	mean	0.25	0.27	0.26	0.30	0.26	0.33
	s.d.	0.11	0.11	0.11	0.13	0.11	0.12
2005	mean	0.23	0.26	0.25	0.29	0.24	0.32
	s.d.	0.10	0.12	0.10	0.12	0.10	0.11
2006	mean	0.25	0.29	0.34	0.32	0.27	0.34
	s.d.	0.11	0.11	0.14	0.12	0.11	0.11
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, 2002 and 2007 Censuses of Manufactures

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 within-industry interquartile range (IQR) of the ratio X_{imp}/TVS_{impX} divided by the IQR of X_{obs}/TVS_{obs}, where X_{imp} represents imputed cases for the variable X, TVS_{impX} are the total value of shipments for the same plants, and X_{obs}/TVS_{obs} is the ratio when both are observed.

Table 3: Distribution of Ratios of Within-Industry Interquartile Ranges of Ratios of Key Variables in Imputed Data vs. Fully Observed Data, 2003-2006 Annual Surveys of Manufactures

percentile	Total Capital Expenditures	Production Worker Hours	Cost of Purchased Electricity	Cost of Fuels	Cost of Materials
			2003		
25th	0.580	d	0.471	0.402	0.445
50th	1.042	0.844	0.687	0.637	0.651
75th	1.561	1.158	1.015	1.013	0.916
			2004		
25th	0.444	0.559	0.348	0.296	0.343
50th	0.715	0.807	0.584	0.518	0.522
75th	1.189	1.135	0.904	0.833	0.745
			2005		
25th	0.326	0.516	0.431	0.423	0.378
50th	d	0.759	0.654	0.875	0.556
75th	1.104	1.063	0.980	1.415	0.795
			2006		
25th	0.249	0.453	0.412	0.553	d
50th	0.512	0.655	0.660	0.960	0.528
75th	1.021	1.009	0.980	1.633	0.793

See notes for table 2

d=suppressed to avoid disclosure of confidential data.

Table 4: Imputation Rates for Key Variables, 2002 Census 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	154	0.44	0.43	0.23	0.44	0.39	0.56
fertilizer	504	0.26	0.32	0.21	0.29	0.28	0.40
flour	240	0.10	0.18	0.06	0.18	0.17	0.28
fluid milk	396	0.19	0.19	0.10	0.18	0.17	0.31
ice	256	0.28	0.48	0.17	0.26	0.25	0.31
pesticides	141	0.24	0.34	0.18	0.31	0.32	0.37
soy beans	94	0.13	0.15	d	0.14	0.14	0.35
sugar	83	0.41	0.39	d	0.35	0.33	0.49

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.
d=suppressed to avoid disclosure of confidential information.

Table 5: Imputation Rates for Key Variables, 2007 Census 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 6: Imputation Rates for Key Variables, 2003-2006 Annual Surveys of Manufactures, Fluid Milk Processing

year	Sample Size	Total Value of Shipments	Total Capital Expenditures	Production Worker Hours	Cost of Electricity Purchased	Cost of Fuels	Cost of Materials
2003	240	0.16	0.15	0.15	0.17	0.13	0.22
2004	262	0.31	0.26	0.27	0.28	0.24	0.32
2005	264	0.25	0.22	0.18	0.26	0.23	0.33
2006	266	0.26	0.26	0.29	0.27	0.24	0.31

Note: book values of assets are not reported in the 2002-2006 Annual Surveys of Manufactures. Also see notes for table 4.

Table 7: OLS Estimates of Production Function Parameters and Productivity Dispersion, Selected Industries, 2002 Census of Manufactures, Fully-Observed Data Only

	Production Function Parameters					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
coffee & tea	0.04	0.05	0.10*	0.78***		1.34	1.77
fertilizer	0.01	0.26***	0.12***	0.60***		1.47	2.08
flour	0.06**	0.22***	0.04	0.68***		1.22	1.60
fluid milk	0.04*	0.16***	0.19***	0.61***		1.28	1.71
ice	0.08**	0.40***	0.19***	0.32***		1.50	2.39
pesticides	0.11	0.14	0.03	0.67***	0.07	1.96	3.21
soybeans	0.17**	0.05	0.07	0.71***		1.28	1.78
sugar	0.13*	0.48***	0.01	0.40***		1.31	1.74

** = significant at the 10% level; ** = significant at the 5% level;*

**** = significant at the 1% level.*

Table 8: OLS Estimates of Production Function Parameters and Productivity Dispersion, Selected Industries, 2002 Census of Manufactures, Census Bureau-Completed Data

	Production Function Parameters					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
coffee & tea	0.10***	0.12***	0.16***	0.64***		1.22	1.73
fertilizer	0.03***	0.29***	0.12***	0.55***		1.57	2.11
flour	0.07***	0.23***	0.07***	0.66***		1.22	1.68
fluid milk	0.05***	0.13***	0.23***	0.60***		1.25	1.74
ice	0.11***	0.34***	0.21***	0.32***		1.30	1.91
pesticides	0.19***	0.08	0.02	0.65***	-0.03	1.55	2.64
soybeans	0.17***	0.05	0.10***	0.73***		1.36	1.90
sugar	0.15***	0.26***	0.07*	0.48***		1.30	1.68

*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. * = significant at the 10% level; ** = significant at the 5% level; *** = significant at the 1% level.*

Table 9: OLS Estimates of Production Function Parameters and Productivity Dispersion, Selected Industries, 2007 Census of Manufactures, Fully Observed Data Only

	Production Function Parameters					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
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

** = significant at the 10% level; ** = significant at the 5% level;*

**** = significant at the 1% level.*

Table 10: OLS Estimates of Production Function Parameters and Productivity Dispersion, Selected Industries, 2007 Census of Manufactures, Census Bureau-Completed Data

	Production Function Parameters					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
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. * = significant at the 10% level; ** = significant at the 5% level; *** = significant at the 1% level.*

Table 11: Production Function Parameters and Productivity Dispersion, Fluid Milk Processing, 2002-2005

	(a) Fully Observed Data			(b) Bureau-Completed Data		
	OLS	LP	WLP	OLS	LP	WLP
	(1)	(2)	(3)	(4)	(5)	(6)
	Production Function Parameters					
Capital	0.01	0.00	-0.06	0.07	0.21	0.21
Labor	0.16	0.15	0.06	0.16	0.16	0.15
Energy	0.21	1.00	1.93	0.20	0.54	0.20
Materials	0.61	0.61	0.69	0.57	0.57	0.59
	75-25 log TFP differences (weighted distributions)					
2002	0.24	0.73	1.59	0.21	0.60	0.32
2003	0.20	0.66	1.35	0.22	0.48	0.30
2004	0.27	0.62	1.53	0.24	0.62	0.35
2005	0.24	0.75	1.54	0.23	0.64	0.37
	90-10 log TFP differences (weighted distributions)					
2002	0.52	1.93	3.98	0.51	1.36	0.67
2003	0.52	1.65	3.45	0.53	1.06	0.64
2004	0.56	1.85	3.57	0.53	1.45	0.71
2005	0.54	1.49	3.04	0.55	1.59	0.69

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 12: OLS Estimates of Production Function Parameters and Productivity, Selected Industries, 2002 Census, CART-completed Data

		Production Function Parameters					TFP ratios	
industry	Sample Size	Capital β_k	Labor β_l	Energy β_e	Materials β_m	Exporter Dummy	75-25, mean (s.d.)	90-10, mean (s.d.)
coffee & tea	154	0.06	0.09	0.12*	0.73***		1.51	2.24
							(0.07)	(0.17)
fertilizer	504	0.05	0.29***	0.10***	0.54***		1.52	2.28
							(0.05)	(0.10)
flour	240	0.06*	0.19***	0.08*	0.65***		1.27	1.76
							(0.03)	(0.07)
fluid milk	396	0.06**	0.15***	0.19***	0.61***		1.32	1.79
							(0.02)	(0.05)
ice	256	0.08**	0.39***	0.24***	0.25***		1.53	2.33
							(0.03)	(0.12)
pesticides	141	0.09	0.15**	0.07	0.63***	0.04	2.00	3.65
							(0.15)	(0.29)
soybeans	94	0.14	0.07	0.12	0.67***		1.58	2.34
							(0.11)	(0.51)
sugar	83	0.14	0.22	0.14	0.52***		1.90	5.36
							(0.54)	(8.10)

Means (standard deviations) across 20 CART-completed datasets of production function parameters and total factor productivity (TFP) dispersion. Standard errors of the estimates from each of the 20 implicates are combined using Rubin's (1987) combining formulas.

Table 13: OLS Estimates of Production Function Parameters and Productivity Dispersion, Selected Industries, 2007 Census, CART-completed Data

		Production Function Parameters					TFP ratios	
industry	Sample Size	Capital β_k	Labor β_l	Energy β_e	Materials β_m	Exporter Dummy	75-25, mean (s.d.)	90-10, mean (s.d.)
coffee & tea	186	0.08*	0.29***	0.09*	0.56***		1.78	3.06
							(0.13)	(0.90)
fertilizer	472	0.07***	0.19***	0.13***	0.56***		1.54	2.41
							(0.04)	(0.22)
flour	210	0.06*	0.12***	0.12**	0.69***		1.27	1.70
							(0.03)	(0.07)
fluid milk	362	0.07*	0.18***	0.11***	0.61**		1.36	1.99
							(0.05)	(0.11)
ice	295	0.07**	0.34***	0.34***	0.29***		1.59	2.35
							(0.03)	(0.08)
pesticides	196	0.10**	0.17**	0.15**	0.50***	0.13	2.07	4.36
							(0.04)	(0.43)
soy beans	89	0.07	0.09	0.03	0.83***		1.48	1.95
							(0.04)	(0.12)
sugar	73	0.14*	0.17	0.15	0.45***		1.69	3.14
							(0.14)	(0.52)

See notes for table 12

Table 14: Production Function Parameters, Fluid Milk Processing, 2002-2005, CART-completed Data

	OLS	LP	WLP
	(1)	(2)	(3)
	Production Function Parameters		
Capital	0.05	0.09	0.02
	(0.01)	(0.27)	(0.58)
Labor	0.14	0.13	0.08
	(0.01)	(0.01)	(0.15)
Energy	0.20	0.46	0.19
	(0.02)	(0.07)	(1.55)
Materials	0.60	0.60	0.78
	(0.03)	(0.03)	(0.25)

Between-imputation standard deviations are in parentheses.

See notes for table 12.

Table 15: Productivity Dispersion, Fluid Milk Processing, 2002-2005, CART-completed Data

	OLS	LP	WLP
	(1)	(2)	(3)
	75-25 log TFP differences (weighted distributions)		
2002	0.25	0.6	0.73
	(0.02)	(0.30	(0.93)
2003	0.25	0.55	0.73
	(0.03)	(0.21)	(1.00)
2004	0.28	0.62	0.74
	(0.02)	(0.31)	(0.96)
2005	0.28	0.59	0.76
	(0.02)	(0.29)	(0.97)
	90-10 log TFP differences (weighted distributions)		
2002	0.57	1.22	1.52
	(0.03)	(0.60)	(1.89)
2003	0.55	1.12	1.47
	(0.05)	(0.50)	(1.87)
2004	0.56	1.25	1.49
	(0.03)	(0.63)	(1.81)
2005	0.63	1.30	1.75
	(0.07)	(0.62)	(2.45)

Between-imputation standard deviations are in parentheses.

See notes for table 12.

Table 16: Posterior Predictive P-Values for Estimates of Output Elasticities and Productivity Dispersion, Selected Industries, 2002 Census of Manufactures, CART-completed Data vs. CART-predicted Data.

industry	Capital β_k	Labor β_l	Energy β_e	Materials β_m	75-25 TFP ratio
coffee & tea	0.932	0.736	0.936	0.972	0.424
fertilizer	0.392	0.208	0.276	0.368	0.564
flour	0.804	0.408	0.768	0.852	0.096
fluid milk	0.848	0.288	0.384	0.460	0.244
ice	0.836	0.668	0.780	0.576	0.388
pesticides	0.744	0.668	0.808	0.860	0.868
soy beans	0.892	0.620	0.944	0.940	0.768
sugar	0.944	0.548	0.740	0.664	0.752

Note: The p-values indicate whether or not the 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.

Table 17: Posterior Predictive P-Values for Estimates of Output Elasticities and Productivity Dispersion, Selected Industries, 2007 Census of Manufactures, CART-completed Data vs. CART-predicted Data.

industry	Capital β_k	Labor β_l	Energy β_e	Materials β_m	75-25 TFP ratio
coffee & tea	0.696	0.204	0.500	0.976	0.384
fertilizer	0.736	0.748	0.500	0.716	0.408
flour	0.972	0.540	0.944	0.892	0.128
fluid milk	0.708	0.340	0.928	0.724	0.612
ice	0.316	0.460	0.852	0.380	0.236
pesticides	0.980	0.576	0.980	0.628	0.676
soy beans	0.840	0.492	0.920	0.980	0.512
sugar	0.928	0.988	0.820	0.828	0.364

See notes for table 16

Table 18: Posterior Predictive P-Values for Estimates of Output Elasticities and Productivity Dispersion, Fluid Milk Processing, 2002-2005, CART-completed Data vs. CART-predicted Data

	OLS	LP	WLP
	(1)	(2)	(3)
	Output Elasticities		
Capital	0.680	0.955	0.925
Labor	0.535	0.490	0.905
Energy	0.225	0.195	0.870
Materials	0.325	0.310	0.895
	75-25 TFP Ratios		
2002	0.605	0.475	0.525
2003	0.455	0.405	0.555
2004	0.610	0.380	0.530
2005	0.905	0.345	0.500
	90-10 TFP Ratios		
2002	0.290	0.475	0.510
2003	0.340	0.475	0.520
2004	0.440	0.370	0.515
2005	0.725	0.240	0.495

See notes for table 16

Table A1: Edit/Impute Flags in the 2002 and 2007 Census of Manufactures and the 2003-2006 Annual Survey of Manufactures

Code	Name	Category
(blank)	Flag Not Set	Non-imputed
A	Administrative Records Data	Imputed
B	Beta (Cold Deck Statistical)	Imputed
C	Analyst Corrected	Non-imputed
D	Donor Model Record	Imputed
E	Endpoints of Limits (Upper/Lower)	Imputed
G	Goldplated	Non-imputed
H	Historic Values	Imputed
J	Subject Matter Rule	Imputed
K	Raked	Non-imputed
L	Logical	Imputed
M	Midpoints of Limits	Imputed
N	Rounded	Non-imputed
O	Override Edit with Reported Data	Non-imputed
P	Prior Year Administrative Records Data	Imputed
S	Direct Substitution	Imputed
T	Trim and Adjust Algorithm	Imputed
U	Unable to Impute	Non-imputed
V	Industry Average	Imputed
W	Warm Deck Statistical	Imputed
X	Unusable	Non-imputed
Z	Acceptable Zero	Non-imputed

Source: Grim (2011).

Table A2: Definitions of Edit/Impute Flags

Edit/Impute Action	Occurs when...
Administrative (A)	the item is imputed by direct substitution of corresponding administrative data (for the same establishment/record).
Cold Deck Statistical (B)	the item is imputed from a statistical (regression/beta) model based on historic data.
Analyst Corrected (C)	the reported value fails an edit, and an analyst directly corrects the (reported or imputed) value.
Model (Donor) Record (D)	the item is imputed using hot deck methods.
High/Low (E)	the item is imputed by direct substitution of value near (high or low) endpoints of imputation range.
Goldplated (G)	the reported value for the item is "protected" from any changes by the edit. The value of a goldplated item is not changed by the editing system, even if the item fails one or more edits. In general, the goldplate flag is set by an analyst.
Historic (H)	the item is imputed by ratio imputation using historic data for the same establishment (for example, prior year data imputation in Manufacturing)
Subject Matter Rule (J)	the item is imputed using a subject matter defined rule (e.g. y=1/2x).

Table A2: Definitions of Edit/Impute Flags (continued)

Edit/Impute Action	Occurs when...
Raked (K)	the sum of a set of detail items do not balance to the total. The details are then changed proportionally to correct the imbalance. This preserves the basic distribution of the details.
Logical (L)	the item's imputation value is defined by an additive mathematical relationship (e.g., obtaining a missing detail item by subtraction).
Midpoint (M)	the item is imputed by direct substitution of midpoint of imputation range.
Rounded (N)	the reported value is replaced by its original value divided by 1000.
Restore Reported Data (O)	the reported value fails an edit. Either an analyst interactively restores the originally reported value of an edit (set by the interactive update system) or the ratio module later "imputes" originally reported data for an item which was imputed in the previous edit pass.
Prior Year Administrative (P)	the item is imputed by ratio imputation using corresponding administrative data from prior year (for same establishment).
Direct Substitution (S)	the item is imputed by direct substitution of another item's value (from within the same questionnaire.)

Table A2: Definitions of Edit/Impute Flags (continued)

Edit/Impute Action	Occurs when...
Trim-and-Adjusted (T)	the item was imputed using the Trim-and Adjust balancing algorithm (balance module default).
Unable to Impute (U)	the reported item is blank or fails an edit, and the system cannot successfully substitute a statistically reasonable value for the original data.
Industry Average (V)	the item is imputed by ratio imputation using an industry average.
Warm Deck Statistical (W)	the item is imputed from a statistical (regression/beta) model based on current data.
Unusable (X)	the sum of a set of detail items cannot be balanced to the total because none of the scripted solutions achieved a balance.
Acceptable Zero (Z)	the reported value for an item is zero, and the item has passed a presence (zero/blank) test. This often occurs with part time reporters (e.g., births, deaths, idles). The zero value will not be changed, even if it fails one or more edits.

Source: Grim (2011).