Imputation in U.S. Manufacturing Data and Implications for Within-Industry Productivity Dispersion [*]

T. Kirk White[†], Jerome P. Reiter[**], and Amil Petrin[***]

[†] Center for Economic Studies, U.S. Census Bureau
[**] Duke University
[***] University of Minnesota,Twin Cities and NBER

Abstract

In the U.S. Census Bureau's 2002 and 2007 manufacturing data, respectively, 79% and 73% of observations have imputed data for at least one variable used to compute total factor productivity. 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 for every manufacturing industry. The results suggest that there is even more within-industry productivity dispersion in U.S. manufacturing than previous research has indicated. We also estimate relationships between plant exit, productivity, prices, and demand shocks. For these estimands, we find the results are substantively robust to using an alternative imputation strategy.

[*] Some of the research in this paper was conducted while the first author was a Census Bureau employee. Any opinions and conclusions expressed herein are those of the authors 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. Corresponding author: T. Kirk White. Email: thomas.kirk.white@census.gov.

1 Introduction

Nearly all economic surveys suffer from item nonresponse, i.e., respondents answer some questions but not others. Most statistical agencies 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 (Little and Rubin (2002)).[1] We investigate the impacts of imputation using data from the U.S. Census Bureau's Census of Manufactures (CM). Using edit/impute flags from the 2002 and 2007 CMs as well as edit/impute flags we recovered from the 1977-1997 CMs, we show that imputations for missing or faulty data comprise a far higher percentage of the CM observations than has been reported in the existing literature.[2] For example, in the 2007 CM, the imputation rates for total value of shipments, cost of electricity and cost of materials inputs in the *average* 6-digit NAICS industry are, respectively, 27%, 37%, and 42%.[3]

The missing data pattern in the CM is non-monotone — e.g., variable X is missing for some plants and Y is observed, while for other plants X is observed and Y is missing. Therefore the percentage of plants with missing data for *some* key variable is even higher than the imputation rate for any given variable. For example, in 2002 and 2007, respectively, 79% and 73% of the CM observations have imputed data for at least one variable used to calculate total factor productivity.[4]

The Census Bureau imputes for missing or faulty data using a variety of methods, some of which include use of administrative records data or lagged values from the same plant. However, we show that for many key variables, the vast majority of imputations in the CM are constructed using industry average ratios or univariate regression using only current-year reported data. The Bureau's primary goal in the CM is to facilitate point estimation of industry aggregates, not to produce plant-level data that is valid for multivariate regression. Imputations from industry average ratios or univariate regression are not appropriate for multivariate regression analysis of microdata — e.g., estimates of within-industry productivity dispersion — because these methods can distort covariances and correlations between variables and lead to underestimates of standard errors (Schafer and Graham (2002) and Little and Rubin (2002)). We find that functions of key variables in the completed CM data show evidence of attenuation and under-estimation of variability. Further, the impacts of imputations are not limited to a few industries and are not mitigated by using statistics that are robust to outliers. The imputations are pervasive, affecting industries throughout the manufacturing sector.

What can be done about this imputed data? One solution is to drop observations with imputed values, and only analyze the plants with complete (non-imputed) data.[5] Unfortunately, it is well-known that in general, complete case analysis

[1] See Kaplan and Schulhofer-Wohl (2010) for a recent example of how imputed data affected policymakers' assessment of the effect of labor mobility on unemployment in the U.S.

[2] One exception to this rule is a recent working paper by Foster, Grim, Haltiwanger, and Wolf (2015), which reports imputation rates in the 2002 and 2007 CMs and 2003-2010 Annual Surveys of Manufactures using the same edit/impute flags that we use (see tables A17-A18 in that paper).

[3] In calculating these imputation rates, following most researchers who use the CM data, we exclude so-called administrative records (AR) cases. These AR plants, which each have fewer than 5 employees, account for about a third of the total number of plants in the CM.

[4] After we reported the high imputation rates in the CM, Foster, Grim, Haltiwanger, and Wolf (2015) also calculated the percentage of CM observations with imputed data for at least one variable used to calculate total factor productivity (TFP). They report (table A17 in their paper) that in 2002 and 2007, respectively, 68.8% and 69.2% of observations have imputed data for at least one of their TFP variables. For these calculations, their set of TFP variables excludes inventories and the book value of assets; our calculations also exclude the book value of assets, but include inventories.

[5] See, for example, Foster, Haltiwanger, and Syverson (2008).

— i.e, using only the plants with no missing or imputed data — sacrifices efficiency and can lead to biased parameter estimates.[6] Suppose, for example, that productivity is correlated with plant size, and smaller plants are more likely to have missing data. Then estimates of within-industry productivity dispersion based on complete cases would be biased, because smaller, less-productive plants would be less likely to be in the sample. Even if complete case analysis does not create a selection bias for some sample, given the large percentage of missing data in the CM, the efficiency loss from using complete cases would also be large.

As an alternative to these strategies, we create completed datasets via multiple imputation (Rubin (1987), Reiter and Raghunathan (2007)).[7] We replace the Census Bureau's imputations in the CM data with multiple imputations using sequences of classification and regression trees (CART), as recently developed by Burgette and Reiter (2010). We describe the method in detail in section 4. Here we provide some intuition for how it improves on the Census Bureau's imputation methods. First, the Bureau's univariate regression and average ratio methods put the imputed values on regression lines, thus underestimating the true variability in the data. In contrast, the CART method is designed to approximate the conditional distributions of the variables being imputed. Unlike the imputation methods the Census Bureau uses most frequently in the CM, the CART method works well for skewed distributions like those in the manufacturing data. It also handles nonmonotone missingness patterns, which are a common feature of economic data. Finally, the CART method flexibly and automatically determines which of the available variables are useful predictors in the imputation model, and flexibly includes interactions and non-linear relationships.

We use the CART method to investigate the effects of imputation on within-industry productivity dispersion in the CM. Within-industry differences in plant-level productivity are large. Averaging across all U.S. manufacturing industries, Syverson (2004b) finds that plants at the 90th percentile of the productivity distribution are nearly twice as productive as plants at the 10th percentile. Explaining the causes and consequences of these productivity differences are currently among the most important research agendas in productivity and industrial organization. Existing explanations include management practices, the quality of labor and capital inputs, information technology, product substitutability, competition, research & development, international trade, and regulation (Syverson (2011)). Within-industry productivity differences also have implications for several other areas of economics, including trade, labor, and macroeconomics (Bartelsman and Doms (2000), Syverson (2011)). The effects of imputed data on (measured) productivity dispersion has largely been ignored in the economics literature.

Ex ante it is not obvious how imputed data will affect estimates of productivity dispersion. Total factor productivity is a ratio of output over inputs. The Census Bureau's most frequently used imputation methods in the CM tend to decrease dispersion in both the numerator (output) and the denominator (inputs). Therefore imputed data could explain some of the existing estimates of measured TFP dispersion, or they could leave more dispersion to be explained.

Using the Census Bureau-completed data, we estimate within-industry productivity dispersion for each of the roughly 470 manufacturing industries in 2002 and 2007. For comparison, we also estimate within-industry dispersion for each industry-year after dropping from the sample any plant with data imputed by the Census Bureau using methods that are likely to reduce within-industry dispersion. Finally, we replace the Census Bureau imputations with multiple imputa-

[6] Complete cases can lead to biased parameter estimates unless the missingness mechanism is Missing Completely At Random (MCAR) (Little and Rubin (2002)). Missingness is MCAR if the probability that the data is missing does not depend on the values of the missing data or the values of the observed data. We find that smaller plants in the CM are more likely to have missing data. Thus the missingness in the CM is not MCAR.

[7] See Little and Rubin (2002) for discussion of the benefits of multiple imputation over some of the methods the Census Bureau uses in the CM, such as average ratio imputation and univariate regression.

tions using the CART method, and we re-estimate within-industry productivity dispersion for each industry-year. Our results suggest that there is actually *more* within-industry productivity dispersion than the existing literature suggests.

For eleven industries for which we have physical quantity data for relatively homogenous products, we also estimate within-industry dispersion in physical quantity-based productivity (TFPQ) and prices, first using the Census Bureau-completed data and then using the CART-completed data. For most of these industries we find that there is significantly more within-industry dispersion in both TFPQ and prices in the CART-completed data versus the Bureau-completed data. We also find that for these industries the effect of imputed data on within-industry dispersion is larger for physical TFP and prices than it is for a revenue-based TFP measure.

These results have implications for counterfactual policy experiments such as those suggested in Hsieh and Klenow (2009) and for cross-country comparisons of allocative efficiency (e.g., Bartelsman, Haltiwanger, and Scarpetta (2008)). If other countries' data have country-specific sources of measurement error (such different methods of imputation for missing data), then cross-country comparisons may also need to account for these differences.

We also investigate how imputed data affects some key empirical relationships between productivity, demand shocks, and plant exit. First, using the estimation sample in table 6 of Foster, Haltiwanger, and Syverson (2008), we exactly replicate their findings.[8] In particular, we estimate the probability of plant exit conditional on plant-level productivity, demand shocks, and product-year fixed effects, both with an without controls for plant's capital stocks. We reproduce FHS's results that both plant-level productivity and demand shocks are associated with negative, statistically significant decreases in the probability of exit.

To investigate the effect of Census Bureau imputations on the FHS results, we use recently-recovered Census Bureau item-level edit/impute flags for the 1977-1997 Censuses of Manufactures. To the best of our knowledge, these item-level edit/impute flags from the 1977-1997 Censuses have not been used previously in publicly available research.[9] Using the edit/impute flags, we find that roughly 50% of the plants in FHS's estimation sample have imputed data for physical output quantities. When we re-estimate the FHS exit specifications using only plants with non-imputed data, we find that the estimated marginal effects of productivity and prices are somewhat attenuated, but overall the estimates are quite similar to FHS's estimates.

Finally, we replace the imputed data in FHS's sample with multiply-imputed CART data and re-run the exit probits using exactly the same sample of plants as in FHS. Using the CART-completed data, the estimated marginal effects for productivity, demand shocks and capital are again similar to FHS's estimates. We also check the validity of the CART imputation models for these regressions and find no evidence that the CART imputation models create biases in the marginal effect estimates.

Our results should be of interest beyond the community of researchers who study plant-level productivity and its causes and consequences. Plant-level U.S. Census manufacturing data has been used to study a variety of other topics, including why firms export (Bernard and Jensen (2004)), the effects of environmental regulation on manufacturing plants (Becker and Henderson (2001) and Greenestone (2002)), product switching (Bernard, Redding, and Schott

[8] We thank Lucia Foster, John Haltiwanger, and Chad Syverson for making their estimation samples and their programs available to us.

[9] Some of the same flags were used in an internal Census Bureau technical paper (Dunne (1998)), but this paper is not publicly available because it contains confidential information. Indeed, it seems that most researchers using confidential Census microdata were not aware that these edit/impute flags exist.

(2010)), industry agglomeration (Ellison, Glaeser, and Kerr (2010)), and firm structure and plant exit (Bernard and Jensen (2007)), just to name a few examples. Given the documented deficiencies of imputation techniques like industry average ratios and univariate regression, the different results for productivity dispersion suggest that improved imputation procedures like the one presented here would benefit users of these data or many other economic datasets containing imputed data.

The next section describes how we estimate plant-level productivity. Section 3 describes the data and how the Census Bureau's imputations affect ratios of key variables in the data. Section 4 describes the sequential CART multiple imputation method. Section 5 how we choose the variables for the imputation models and our strategy for choosing industries for the analysis of TFPQ and price dispersion. Section 6 shows how the imputations affect estimates of productivity dispersion and some key empirical relationships between productivity and other variables. Section 7 concludes.

2 Plant-level Productivity Estimation

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. We use a popular method: we use industry cost shares to estimate a production function. Specifically, for each industry, we assume that the technology of every plant within an industry can be approximated by a 4-factor Cobb-Douglas production function. We calculate two measures of TFP for plant i in a given year. First, we calculate TFP based on the quantity of the plant's physical output:

$$TFPQ_i = lnQ_i - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \tag{1}$$

where Q_i is the quantity of physical output of plant i, K_i is the capital stock, L_i is labor, E_i is energy, M_i is materials, and the βs are the respective output elasticities for each input. We also calculate a TFP measure based on the plant's revenues:

$$TFPR_i = ln(P_iQ_i) - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \tag{2}$$

where P_iQ_i is the total value of the plant's output.[10] We describe the variable construction in more detail in the appendix.

This method of estimating productivity has well-known deficiences (see, e.g., Griliches and Mairesse (1995)). Previous research (e.g., Van Biesebroeck (2004)) has analyzed the strengths and weaknesses of various other methods of estimating productivity. For example, proxy methods (Olley and Pakes (1996), Levinsohn and Petrin (2003), Wooldridge (2009)) address the well-known endogeneity issue (Marschak and Andrews (1944)) and do not impose constant returns to scale. However, our main goal in this paper is to show how different methods of imputing for missing data affect widely-used estimates of plant-level productivity dispersion and key empirical relationships between productivity and other variables. Using cost-shares to estimate the parameters in equations 1 and 2 is the approach used in some of the most influential research in the productivity literature.[11] Thus this choice facilitates comparisons with the existing

[10] FHS call this measure "Traditional TFP." However, much of the recent literature calls this measure "Revenue TFP" and uses the acronym TFPR (see, e.g., Hsieh and Klenow (2009)), so we adopt the TFPR label. Note that if the productivity analysis uses panel data (as we do in section 6.1 below), plant-level revenues and expenditures are deflated by industry-level price indexes.

[11] For example, Bailey, Hulten, and Campbell (1992), Syverson (2004b), Syverson (2004a), and Foster, Haltiwanger,

literature on plant-level productivity. To the extent that other methods of estimating production functions are more sensitive to imputed data, the resulting productivity estimates may also be more sensitive to imputed data.[12]

3 The Impact of Imputed Data in the Census of Manufactures

The quinquennial Census of Manufactures (CM) includes data on roughly 300,000 manufacturing plants. The data for the smallest plants — about a third of the sample — are almost entirely imputed. Following most researchers who use the CM, we exclude these so-called administrative records plants from all of our analysis.

Over the years, the CM has been plagued by item non-response, and the Census Bureau has created imputations for this missing data. However, until the 2002 census, it was difficult for researchers to identify which, if any, items for a given plant were imputed due to item nonresponse, because item-level flags were not made available to researchers. Previous researchers developed several clever ways to identify some of the imputed values.[13] However, the item-level flags that became available to researchers starting with the 2002 Census show that a much higher percentage of observations are imputed than are identified by these methods.

Table 1 presents the means and standard deviations of the within-industry imputation rates for key variables for all 6-digit NAICS industries (the most detailed level of industry classification) from the 2002 and 2007 Censuses. 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 the total value of shipments are imputed. For some other key variables, the mean imputation rate is even higher. There is also significant variation in the imputation rates across industries. For example, an industry that is one standard deviation above the mean cost of materials imputation rate would have roughly 52% of its materials data imputed in 2007. Note that these are imputation rates for a given variable, and the missingness pattern in the CM data is nonmonotone. In a multivariate analysis — such as estimating total factor productivity (TFP) — the percentage of plants that have imputed data for *some* variable is usually larger than the percentage of plants that have missing data for any given variable. In 2002 and 2007, respectively, 73% and 79% of the observations in the mail sample of the CM have imputed data for at least one variable used to calculate TFP.

The Census Bureau uses a variety of methods to impute for missing data, and the frequency of use of each method varies across variables.[14] In the Census of Manufactures, for most variables used to compute TFP, the Census Bureau most frequently uses univariate regression imputation[15] or industry average ratio imputation. For example, in 2007, for total value of shipments, cost of materials, cost of fuels, cost of electricity, production worker hours, production worker wages, beginning of year inventories, and end of year inventories, respectively 58%, 67%, 87%, 87%, 80%, 78%, 62%, and 78% of imputations were constructed using either univariate regression with only current-year data or using industry average ratios.[16] The industry average ratio method imputes for missing values of variable Y by

and Syverson (2008), all use this approach. Hsieh and Klenow (2009) also use cost shares, although they estimate two-factor value-added production functions.

[12] An earlier version of this paper showed greater sensitivity of proxy methods for estimating production functions to imputed data. Further investigation of this sensitivity should be an important area for future research (see, e.g., Foster, Grim, Haltiwanger, and Wolf (2015)).

[13] See for example, Foster, Haltiwanger, and Syverson (2008), Roberts and Supina (1996), and Roberts and Supina (2000).

[14] Tables A1 and A2 in the appendix describe the various imputation methods that the Census Bureau uses in the Census of Manufactures.

[15] Univariate regression imputation replaces missing data with the predicted value from a univariate regression.

[16] Appendix tables A4 and A5 present the percentages of observations by imputation method for key variables in the 2002 and 2007 Censuses of Manufactures.

multiplying the observed value of variable X by an industry average ratio:

$$Y_i^{imp} = X_i \overline{(\frac{Y}{X})} \tag{3}$$

where Y_i^{imp} is the imputed value of Y for plant i, X_i is the observed value of X for the same plant, and $\overline{(\frac{Y}{X})}$, is an average ratio of $\frac{Y}{X}$ for plant i's industry. Thus all the imputed values for Y lie on a regression line running through the origin, where the slope is the industry average ratio $\frac{Y}{X}$. Estimates of the variance of Y conditional on X using data imputed this way understate the true conditional variance. The same is true for the univariate regression imputation method.

Furthermore, these methods can introduce bias into estimated relationships between variables. To see this, suppose that we use the industry average ratio method to impute for Y conditional on X, but in the true (unobserved) data Y is linear in X and Z. Then suppose we regress Y (including the imputed values of Y) on X and Z. The coefficient on Z will be attenuated, because the imputed values of Y incorrectly generate conditional independencies in a subset of observations.[17] The same logic applies to regression imputation if an important predictor is omitted from the imputation model.

To get some sense of how the Census Bureau's imputations are affecting the relationships between key variables in the CM data, we compute the following ratio for several input variables X:

$$R_X = \frac{IQR(\frac{X_{imp}}{TVS_{impX}})}{IQR(\frac{X_{obs}}{TVS_{obs}})} \tag{4}$$

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 ratio X/TVS in the imputed data than there is in the observed data. We compute these ratios for several inputs: capital, production worker hours, the cost of materials, the cost of electricity, and the cost of fuels. 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 within-industry variation in the ratios of key variables, in some cases quite drastically. For example, for the median industry in 2002, for plants with imputed cost of materials data, the IQR of the ratio of cost of material to total value of shipments is only 21% as large as the IQR of the same ratio for plants with observed data for both variables. 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 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.

4 Multiple Imputation using Classification and Regression Trees

Given the evidence of the impact of imputed data in table 2 and the deficiences of the imputation methods used most frequently in the CM, we replace the Bureau's imputations with multiple imputations created via sequential regression

[17] Note that in this case the ratio imputation method introduces measurement error (in Y) that is correlated with the explanatory variable Z.

trees, as developed by Burgette and Reiter (2010).[18] Before describing the details of the CART method, we provide some intuition for how it improves on the Census Bureau's imputation methods. First, as noted above, the Bureau's univariate regression and average ratio methods put the imputed values on regression lines, thus underestimating the true variability in the data. In contrast, the CART method is designed to approximate the conditional distributions of the variables being imputed. Second, some of the Bureau's imputation methods use very simple models, conditioning on a single variable. As noted above, this can introduce bias in estimates of relationships between the imputed variable and other variables. The CART method is designed to avoid this problem by potentially conditioning on any available variables (as well as interactions of those variables). The CART method has also been shown to perform well in the related problem of generating synthetic data (Reiter (2005), Drechler and Reiter (2011), and Wang and Reiter (2012)). This suggests that the CART method may also produce reasonable imputations for missing data in the CM.

Classification and regression trees (CART) approximate the conditional distribution of a single variable using multiple predictors (see Breiman, Friedman, Olshen, and Stone (1984), Hastie, Tibshirani, and Friedman (2009), and Ripley (2009)). Intuitively, the procedure is designed to classify units (in our application, manufacturing plants) into relatively homogeneous groups. One can think of the algorithm as building a tree from the ground up, where the leaves of the tree contain sets of similar plants. Suppose, for example, we are building an imputation model for plant output, and suppose that we have only one potential predictor: employment. In the each stage of the tree-building process, the goal is to use employment to divide the plants into two subgroups that are more homogeneous in plant output than the group that is being divided. The CART algorithm searches through all the observed values of employment for the threshold such that the variance of *output* within the two subgroups (above and below the employment threshold), is reduced the most. This split results in the first two branches in the tree — plants with employment values below the threshold are put in one branch, and those above the threshold are put in the other branch. The process continues recursively on each branch of the tree until the "leaves" contain some minimum number of plants or until the leaves all meet some criteria for homogeneity.

Of course, in general there will be many potential predictors available. In the general case, at each stage of the tree-building process, the algorithm searches over all observed values (within a given branch) of all the predictors for the split which most reduces the variance of output in that branch. Once the tree for output is built, imputations for output are created by taking draws from the output observations in the appropriate leaves of the tree. Thus imputations for missing output for plant i are drawn from observed output values of plants that are similar to i. A separate tree is built for each variable in the dataset, and the entire process is repeated multiple times to create multiple imputations for each missing value.

The CART multiple imputation algorithm is an example of a chained equations approach to multiple imputation (Raghunathan, Lepkowski, Hoewyk, and Solenberger (2001), Van Buuren and Groothuis-Oudshoorn (2011)). Applications of chained equations approaches often use parametric models at each step, such as generalized linear models (GLMs). CART models have potential advantages over GLMs in that (i) they can capture distributional features that can be difficult to model or even detect with standard GLMs, such as non-linearities and complex interactions, and (ii) they can be applied with minimal tuning even when the data comprise many variables, thus minimizing the need for model selection procedures. However, if the econometrician can specify the correct GLM, the CART algorithm can be comparatively inefficient. Applications of chained equations for continuous data also use predictive mean matching, in which each missing value is replaced with an observed value taken from a record with the most similar predicted value of the fitted GLM. The CART algorithm is similar in spirit to this approach, essentially using a more flexible regression tree to decide the donor pool.

[18] The "mice" software package in R now includes routines for CART imputation.

The CART algorithm, and standard multiple imputation routines in general, are appropriate provided that the data are Missing at Random[19] (Rubin (1976)). If data are not Missing at Random, one should consider nonignorable missing data models, such as selection models and pattern mixture models. When items throughout the dataset are missing (as they are in the Census of Manufactures), using standard selection models on only the complete cases tosses out the information from cases that have partial responses. This can be inefficient if data are not too far from Missing at Random, which may be plausible if the imputation models contain enough variables.

We now describe the CART procedure in more detail. We run the imputation process separately for each industry. We begin the process in any industry by deleting (making missing) any Census Bureau imputations identified by the item-level edit/impute flags and filling in initial guesses for these 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 variance of Y_1 in the leaves. We cease splitting any particular leaf when the variance in that leaf is less than 0.00001 times the variance in the marginal distribution of Y_1 or when we cannot ensure at least k records in each child leaf. For any plant with missing data, we trace down the branches of $\mathcal{Y}^{(1)}$ until we find that plant's terminal leaf. Let L_w be the wth terminal leaf in $\mathcal{Y}^{(1)}$, and let $Y^{(1)}_{L_w}$ 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 $Y^{(1)}_{L_w}$ 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. For the analysis of TFP dispersion in every manufacturing industry, described in at the beginning of section 5, we set M to 100. For the analysis of industries with physical quantities of output, described in section 5.1, we set M to 500.[20] By cycling through the process ten times between completed datasets, we minimize dependence between the completed datasets.[21]

[19] Note that Missing at Random (MAR) and Missing Completely at Random (MCAR) are two different patterns of missing data. The missing data is MCAR only if the missingness does not depend on any of the observed data (or the missing data). The data are MAR if the probability an item is missing depends on the observed data, but does not depend on the missing values themselves. Missing Not at Random (MNAR) or nonignorable missingness means the probability that an item is missing depends on the value of the missing item, even conditional on the observed data.

[20] 500 datasets is probably more than necessary for our analysis. When we estimated within-industry TFP and price dispersion using only 20 CART-completed datasets, most of the estimates were within 1 or 2 percentage points of the estimates in table 4, and all but 3 estimates were within 5 percentage points. The exceptions were in industries with a very small number of observations and relatively large within-industry dispersion (prices for carbon black and prices and productivity for plywood). Rubin (1987) provides some guidance on how many datasets are needed. Intuitively, more datasets are needed when there is more missing information about the estimand of interest.

[21] This independence allows us to use Rubin's (1987) combining formulas to estimate the impact of imputed data on our standard errors, which we cannot do with the Census Bureau's single imputations.

5 Industries and Imputation Models

We first investigate the impact of imputation on within-industry productivity dispersion in each of the roughly 470 industries in the U.S. manufacturing sector in 2002 and 2007. We have plant-level revenue data for all of these industries, but we have physical output data for only a few industries. Following most of the literature on plant-level TFP, here we focus on the revenue TFP measure (equation 2).

We estimate within-industry TFPR dispersion using three different datasets. The first dataset is the Census Bureau completed data (i.e., including both reported data and the Census Bureau's imputations). For each industry-year we compute the ratio of a plant at the 75th percentile of the industry's TFPR distribution to a plant at the 25th percentile.[22] For the second dataset, we drop plants with industry average ratio imputations or imputations from univariate regression using only current-year data, and we recompute the 75-25 TFPR ratios for each industry-year.[23] We call this the non-imputed data. Note, however, that this "non-imputed" data includes Census Bureau imputations constructed using lagged data (e.g., trivariate regression) and imputations from administrative records (e.g., for payroll). These other types of imputations may be more reliable than industry average ratio imputations or imputations from univariate regression using only current-year data. Finally, for the third dataset, we replace each of the Census Bureau's industry average ratio and univariate regression imputations with 100 sequential CART imputations. We compute the 75-25 TFPR ratio separately for each industry-year in each CART-completed dataset and then take the average 75-25 TFPR ratio (across the 100 datasets) for each industry-year.

5.1 Industries with Physical Output and Price Measures

As emphasized by Foster, Haltiwanger, and Syverson (2008), there can be important differences between revenue TFP and TFPQ. To facilitate comparision with the existing literature on plant-level productivity, we focus on the manufacturing industries studied in Foster, Haltiwanger, and Syverson (2008): boxes, white pan bread, carbon black, coffee, ready-mix concrete, hardwood flooring, motor gasoline, ice, plywood, and sugar.[24] According to FHS, "Producers of these products make outputs that are among the most physically homogeneous in the manufacturing sector." In industries that are relatively homogeneous, plants with missing data are likely to be relatively similar to plants with complete data. Thus for homogeneous 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 between productivity and other variables.

For all of these industries, in at least one year we have data on the values and physical quantities of the products the plants produce.[25] This allows us to construct plant-level prices and to ensure that the plants in our sample are specializing in the same product or products. We describe our industries and products in detail in the appendix.

Table 4 shows the sample size and imputation rates for key variables for each of our industries. Except for the concrete industry in 2007, the imputation rates in our sample are significantly lower than in the average manufacturing industry.

[22] We convert 2007 value variables to 2002 dollars.

[23] For this analysis we do *not* drop plants in 2007 if only the capital asset variable's edit/impute flag indicates that it was imputed by univariate (or multivariate) regression. Because of a processing anomaly in the 2007 Census, nearly all capital asset observations were flagged as regression imputes, even if the data was not imputed by regression. We discuss this in more detail in section C of the Appendix.

[24] Some of these industries have also been studied previously by Roberts and Supina (1996), Roberts and Supina (2000), and Davis, Grim, and Haltiwanger (2008).

[25] For two of the industries — motor gasoline and ice manufacturing — we also have consistent measures of product-level physical quantities in both 2002 and 2007.

However, the imputation rate for physical quantity of product shipped, at 45%, is substantially higher than the rates for the other variables. In the ready-mix concrete industry in 2007, for most variables the imputation rates are close to the manufacturing average. As noted above, the missingness pattern in the CM data is nonmonotone, and measuring TFP requires using combinations of all of the variables in table 4. Thus the percentage of plants with missing/imputed data for at least one of the variables is higher than the imputation rate for any given variable.

5.2 Imputation Models

What variables should be included in the imputation models? Little and Rubin (2002) and Schafer and Graham (2002) provide guidance on this point for multiple imputation methods in general. Essentially, the imputer should include as input to the imputation procedure any available variable he thinks is not independent of the other variables, including any variable that will be used in the subsequent analysis. The goal of the sequential CART procedure is to preserve the joint distribution of the data. The CART procedure does not build any relationships into the data that do not exist in the observed (non-imputed) data. The imputer tells the CART algorithm which variables *might* be useful for constructing an imputation model. However, for any given "dependent" variable, the CART procedure only splits on predictors that are useful for characterizing the conditional distribution of that variable.

Since we want to analyze total factor productivity, we include any variable in the CM that is used to calcuate TFP, as well as variables that we expect to be useful predictors of these variables. These considerations lead us to include a rich set of variables as inputs to the CART procedure. For the analysis described at the beginning of section 5, TFPR dispersion in every manufacturing industry, the potential predictors for each tree include — whenever the variable is not the dependent variable — the total value of shipments, changes in inventories, the total cost of materials and energy, the plant-year's ratio of cost of energy (electricity and fuels) over the cost of materials, salaries and wages, employment, production worker hours, the plant-year's ratio of production worker wages to (total) salaries and wages, and the book value of assets.

For each industry in table 4 except concrete, the potential predictors for each tree include the variables mentioned above — except that the costs of electricity and fuels and production worker wages are included in levels — as well as the physical quantity of product shipments. In addition, employment is broken out by the number of production workers and the number of non-production workers. Since previous research (e.g., FHS) has shown a correlation between plant-level productivity and plant survival, we also include as a potential predictor an indicator for whether or not the plant exited between 2002 and 2007.

Syverson (2004a) finds that output and TFP dispersion in the ready-mix concrete industry vary significantly across (geographically segmented) markets, and that market-level demand density is an important predictor of TFP and TFP dispersion within a market. Accordingly, in addition to the input and output variables included as potential predictors in the imputation models for the other industries, for the concrete industry we include the Syverson (2004a) measure of demand density as a potential predictor in the imputation model. On the other hand, we do not observe plant-level physical quantities of output for the concrete industry in 2002 or 2007, so we cannot include them as potential predictors in the imputation model.

6 Results

Table 3 presents summary statistics for within-industry revenue TFP dispersion for all manufacturing industries in 2002 and 2007. The columns of the table show the 75-25 ratios of TFPR for the mean industry and for each quartile of the industry distribution for each year. Rows 1 and 4 show that in the Census Bureau-completed data there is substantial within-industry TFPR dispersion throughout the manufacturing sector, as emphasized by Syverson (2004b). For the average industry in 2002, a plant at the 75th percentile of its industry's TFPR distribution was 44% more productive than a plant at the 25th percentile in the same industry. For the average industry in 2007, the 75-25 difference was 53%. These estimates are similar to Syverson's (2004) estimates of the mean within-industry 75-25 differences of TFPR ranging from 34% to 56% (depending on the TFPR measure). In 2002 the 75-25 percentage differences are 34% for an industry at the 25th percentile of the TFPR dispersion distribution and 53% for an industry at the 75th percentile of the industry distribution. These estimates are also consistent with Syverson's (2004) findings of significant within-industry dispersion across the entire manufacturing sector in 1977. In 2007 the TFPR dispersion estimates in the Bureau-completed data are somewhat higher throughout the manufacturing sector, with a 75-25 percentage difference of 53% for the average industry.

The second and fourth rows of table 3 show the results from the non-imputed data. For these plants, in both 2002 and 2007, throughout the manufacturing sector, within-industry TFPR dispersion is even higher than in the Census Bureau completed data. For the average industries in 2002 and 2007, the within-industry 75-25 percentage differences in TFPR are, respectively, 17 and 12 percentage points higher than in the average industries in the Bureau-completed data. This suggests that there is even more within-industry productivity dispersion in U.S. manufacturing than previously thought. However, these estimates from plants with non-imputed data may suffer from sample selection bias, because smaller (potentially less productive) plants are more likely to have missing data.

For the CART-completed data, we compute the 75-25 TFPR ratio separately for each industry-year in each completed dataset, and then compute the mean 75-25 ratio (for each industry-year) across the 100 implicates. Rows 3 and 6 of table 3 report moments of the industry-year distributions of these 100-implicate means. In the CART-completed data, there is even more within-industry productivity dispersion than in the non-imputed data, suggesting that dropping plants with imputed data does create a sample selection bias. For the average industries, the 75-25 percentage differences in TFPR are 71% and 76% in 2002 and 2007, respectively. The increase in within-industry dispersion in the CART-completed data vs. the non-imputed data or the Bureau-completed data is apparent throughout the manufacturing sector in both years. For example, in 2002, for an industry at the 25th percentile of the industry distribution, the within-industry 75-25 percentage difference is 20 percentage points higher in the CART-completed data than in the Bureau-complete data. In 2007, in an industry at the 75th percentile of the industry distribution, the productivity dispersion measure is 26 percentage points greater in the CART-completed data than in the Bureau-completed data.

Table 5 presents within-industry TFPR, TFPQ, and price dispersion statistics for the industry-years for which we are able to calculate them. For each measure we compute the ratio of the 75th percentile to the 25th percentile. Columns 1, 3, and 5 present these statistics calculated from the Bureau-completed data, which includes both the non-imputed data and the Census Bureau's imputations for missing data. Like FHS, we find more within-industry dispersion in the physical quantity-based productivity measure, TFPQ, than in the revenue-based measure. With the exception of boxes and plywood, product prices in the Bureau-completed data are also less dispersed than either measure of productivity — this is also consistent with FHS's findings.

Columns 2, 4, and 6 of table 5 present estimates of within-industry productivity and price dispersion based on datasets completed with the sequential CART method. We compute each statistic separately from each of our 500 completed datasets, and report the means of the 500 estimates. Comparing columns 1 and 2, for every industry except bread there is more within-industry TFPR dispersion in the CART-completed data than in the Bureau-completed data, and in some industries, there is much more dispersion. Comparing TFPQ (columns 3 and 4) and prices (columns 5 and 6), the differences between the CART-completed data and the Bureau-completed data are even larger. For the average industry-year in our sample, TFPQ dispersion is 30% higher in the CART-completed data and price dispersion is 46% higher. So FHS's result that dispersion in TFPQ exceeds TFPR is strengthened in the CART-completed data.

The impact of imputed data also varies substantially across industries. TFPQ and prices for plywood are impacted the most, perhaps because of the relatively small sample size, and the *relatively* heterogeneous products the industry produces (compared to the other industries in our sample). Productivity and price dispersion estimates for the gasoline industry seem to be the least affected by imputed data, perhaps because the primary products are quite homogeneous.

How do the dispersion measures in the Bureau-completed data and the CART-completed data compare to dispersion measures in the non-imputed data? To answer this question we calculate productivity and price dispersion statistics for the subset of plants in our sample that have no imputed data.[26] For each industry-year statistic in table 5, we compute the ratio of that statistic in the non-imputed data over the same statistic calculated from the Bureau-completed data. A ratio greater than 1 indicates that there is more dispersion in the non-imputed data than in the completed data. For TFPR, on average there is slightly *less* dispersion in the non-imputed data than in the Bureau-completed data — the average ratio is 0.96 — although there is some industry-year variation in this ratio. For the average industry-year, the 75-25 TFPQ ratio is 14% larger and the price dispersion is 22% greater in the non-imputed data than in the Bureau-completed data. Thus, on average the TFPQ and price dispersion measures in the non-imputed data lie about half way between the dispersion estimates from the Bureau-completed data and the CART-completed data.

All of the sample sizes in table 5, with possible exception of the boxes industry, are small. Do small sample sizes make it more likely that imputation affects estimates of productivity dispersion? To address this question, we turn to an industry with much larger sample sizes: ready-mix concrete. The Census Bureau last collected physical quantity of shipments data for concrete in 1992, so we cannot compute TFPQ or price dispersion for this industry in 2002 or 2007.[27] However, we can still compute TFPR. Furthermore, the large sample sizes for the concrete industry allow us to report estimates for this industry for Bureau-completed data and non-imputed data separately without violating Census Bureau disclosure rules.

Table 6 presents 75-25 TFPR ratios for the concrete industry in 2002 and 2007 on three different datasets: the Bureau-completed data, plants with only non-imputed data, and CART-completed data. Comparing column 2 to column 4, TFPR dispersion in the non-imputed data is the same (to two decimal places) as in the Bureau-completed data in both 2002 and 2007. Column 6 shows the mean TFPR dispersion estimates from 500 CART-completed datasets. Compared to the Bureau-completed data, the 75-25 TFPR ratio in the CART-completed data is 46 percentage points higher in 2002 and 42 percentage points higher in 2007. Thus even in an industry with thousands of producers, imputations have a substantial effect on productivity dispersion estimates.

[26] For the productivity statistics, these are plants that have non-imputed data for all the variables required to calculate productivity. For the price statistics, we only require that the plants have non-imputed product quantity and product value data. Only about 43% of the plants in the samples for table 4 have fully-observed (non-imputed) data for productivity, and about 62% have non-imputed data for product quantity and value data. Thus the efficiency losses from using only complete cases in this sample could be quite large.

[27] Although we cannot calculate TFPQ or prices for concrete in 2002 or 2007, we do have physical quantity data for concrete from the 1977, 1982, 1987, and 1992 Censuses. Table A6 in the appendix reports within-industry-year 75-25 ratios of concrete TFPR, TFPQ, and prices for these years using the same sample of concrete plants used in FHS (2008). In every year, there is more within-industry dispersion in the CART-completed data for concrete (compared to the Bureau-completed data) for all three measures, especially for prices. For TFPR and TFPQ, the 75-25 ratios are, respectively, 8 to 11 and 8 to 16 percentage points higher in the CART-completed data than in the Bureau-completed data. For prices, the 75-25 ratios are 12 to 30 percentage points higher in the CART-completed data.

If our CART imputation models are correct,[28] these results suggest that in addition to the (substantial) efficiency losses that would result from complete-cases analysis, using only complete cases would also create a sample selection bias. Smaller plants are more likely to have missing data, so the apparent sample selection bias may be the result of correlations between size, productivity and prices.

The lower degree of price dispersion in the Bureau-completed data is not surprising given what we know about the Census Bureau's imputation methods. Although the Bureau uses a variety of imputation methods and models for some variables, for the product physical quantity data, it primarily uses the industry average ratio method. This method imputes for missing quantity data by multiplying the value of the product shipments (for the same plant) by an industry average ratio of product quantity to product value. This implicitly assumes that all plants with imputed physical quantity data for a given product sell the product for the same price. Given the degree of within-industry price dispersion we see in the non-imputed data, this imputation method is at least part of the reason price dispersion is significantly lower in the Bureau-completed data. Since small plants are more likely to have missing data than larger plants, plugging in industry averages could underestimate dispersion even further than if their were no relationship between plant size and missingness.

The method of imputing for missing data clearly affects estimates of within-industry productivity and price dispersion. But do imputations affect key empirical relationships between productivity and other variables? In the next subsection we show that in some important cases they do not.

6.1 Productivity, Demand Shocks, and Plant Survival

Foster, Haltiwanger, and Syverson (2008) build on an important theoretical and empirical literature analyzing and documenting the connection between producers' productivity and survival and its effect on industry aggregates (Jovanovic (1982), Ericson and Pakes (1995), Melitz (2003), and Bartelsman and Doms (2000)). One of FHS's contributions is to estimate the effects of productivity, prices and idiosyncratic demand shocks on plant survival (table 6 in FHS). In table A7 in the appendix we reproduce FHS's table 6, which presents results from probits of plant exit on plant-level productivity, price, demand, and capital stock measures.[29] In all specifications, the marginal effects of traditional TFPR, TFPQ and idiosyncratic demand shocks are negative and statistically significant at the 1% level or 5% level.[30]

To see how FHS's exit probit results are affected by imputations, we matched recently-recovered edit/impute flags from the 1977-1997 Censuses of Manufactures to FHS's sample.[31] Roughly 60% of the plant-year observations in FHS's estimation sample have imputed data for at least one of the variables used to compute productivity.[32] For any

[28] In section D of the Appendix, we present the results of validity of the CART imputation models.

[29] We thank Lucia Foster, John Haltiwanger and Chad Syverson for giving us their computer codes and (with approval from the Census Bureau) for giving us access to the final datasets they used in FHS.

[30] All the specifications include product-year fixed effects (not reported).

[31] For summary statistics of these edit/impute flags for the entire 1977-1997 Censuses of Manufactures, as well as a description of the process of recovering these data, see White (2014).

[32] Since FHS did not have access to these item-level impute flags, they used reverse-engineering methods (e.g., looking at modal ratios of costs of materials over payroll, total value of shipments over payroll, and physical output over payroll) to identify imputed data and remove those plants from their sample. They also conducted some sensitivity analysis using various assumptions about rounding error in the ratios used to identify imputed physical output. That sensitivity analysis is available on Chad Syverson's website at http://home.uchicago.edu/syverson/modalpricerobust.pdf.

CMF variable except payroll and total employment,[33] we replace the imputed data in FHS's estimation sample with multiply-imputed data from our CART models, keeping exactly the same sample of plants as in FHS's estimation sample.[34] Table 7 shows the results of the exit probits run on 500 CART-completed datasets. We estimate each probit separately on each of the CART-completed datasets and report the means of the marginal effect estimates. For each probit, the standard errors are clustered by plant. We combine the 500 sets of standard errors using Rubin's (1987) combining formula.

The results from the CART-completed data follow the same pattern as FHS's results. Traditional TFP, TFPR, TFPQ, prices, and demand shocks are all negatively associated with exit, both with and without controls for plant capital stocks. Although the marginal effect estimates for Traditional TFP, TFPR, and prices (columns 1, 2, and 4) are attenuated compared to FHS's estimates, in most specifications the FHS exit probit results are substantively robust to replacing the Census Bureau's imputations with multiple CART imputations.

Table A9 in the appendix presents results from running the same probit specifications as above after dropping plant-year observations with imputed data for any of the CMF variables (except payroll or total employment) used to compute prices, demand shocks or any of the TFP measures.[35] In most specifications, the marginal effect estimates are not statistically significantly different from FHS's estimates, and the marginal effect estimates are statistically significant at the 1, 5, or 10% levels.

6.2 Validity Checks

The statistical literature on imputation suggests that the CART method should do a better job of capturing the joint distribution of the data than, for example, the Census Bureau's industry average ratio method or univariate regression using only current-year data. However, it is still possible that the CART imputations are distorting the joint distribution of the variables in our data in a way that leads to biased estimates of productivity dispersion or the marginal effect estimates in table 7. To check the validity of our imputation models for the analyses above, we use posterior predictive checks (He, Zaslavsky, Harrington, Catalano, and Landrum (2010)). The idea is that we use the CART method to create many pairs of datasets, and check to see if the sign of the difference between the estimates from each pair is consistently positive or negative. The first dataset in each pair (the *completed* dataset) includes CART-imputed and non-imputed data. These are the same CART-completed datasets that we used for the CART estimates in tables 5, 6, and 7. The second dataset in each pair — called the *predicted* dataset — is almost entirely imputed. For each predicted dataset, using the same CART trees that were used for the corresponding completed dataset, we replace *every* observation for any variable that has *any* imputed observations in the original data. So, for example, if the total value of shipments is imputed for any observation in the original data, then in the predicted datasets we impute the total value of shipments for every observation. Then we re-estimate the productivity dispersion statistics and exit regressions on each CART-completed dataset and each CART-predicted dataset. If, for example, the marginal effect estimate of interest is consistently higher in the predicted datasets, then this is evidence that the imputation model may be leading to upward-biased estimates of that marginal effect.

[33] For payroll and total employment, the Census Bureau has current-year and/or prior year administrative records data for every plant in the Census of Manufactures. These administrative records data are thought to be relatively accurate, and therefore we treat them as observed (i.e., non-imputed) data.

[34] After replacing the Bureau-imputed data with CART data, we reconstructed industry-level price indexes, idiosyncratic demand shocks, Traditional TFP, TFPR, and TFPQ following the procedures that FHS used to construct these variables in their dataset. We describe this variable construction in more detail in the Appendix.

[35] After dropping observations with imputed data, where necessary we reconstructed FHS's variables using the same methodology that FHS used to construct these variables.

We conduct validity checks for each measure of productivity and/or price dispersion in tables 5 and 6 and for the exit probits in table 7. For most of the estimates we find no evidence of bias from the CART imputations. In the few cases where there is evidence of a bias, the biases tend to be small. Together this evidence suggests that the CART model generates plausible data with respect to most of the estimated relationships represented in tables 5, 6, and 7. A formal description of the validity checks and the detailed results are presented in section D in the Appendix.[36]

7 Conclusions and Suggestions for Further Research

Much of the literature on plant-level productivity uses the U.S. Census Bureau's Census of Manufactures (CM). A surprisingly large percentage of the CM data is imputed. Our results suggest that these imputations have an economically significant effect on estimates of within-industry productivity dispersion throughout the manufacturing sector. For a handful of industries for which physical quantity of output data is available, we also show that the Census Bureau's imputations have a significant effect on estimates of within-industry price dispersion.

Using classification and regression trees (CART), we provide a new set of multiple 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 relationships between productivity and other economic 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 is more within-industry productivity dispersion than the previous literature suggests. However, we also find that estimated joint relationships between plant survival, productivity, and demand shocks are similar when we replace Bureau-imputed data with CART-imputed data.

In addition to within-industry differences in prices or demand shocks, the existing literature provides a variety of explanations for measured within-industry productivity differences, including heterogeneity in management practices, the quality of labor inputs, information technology, research and development, international trade, and regulation. None of these variables are part of the Census Bureau's imputation models for the CM. Most existing research using Census Bureau manufacturing data was unable to identify much of the imputed data. Now that item-level impute flags are available for the CM, future research could fruitfully explore whether estimates of correlations between productivity dispersion and these other variables are robust to different methods of imputation for missing data.

For three variables used to compute TFP — payroll, employment, and sales — the Census Bureau has administrative records data. Variables from administrative records data can be included as additional predictors in the CART models for imputation of missing data. Also, for plants which are surveyed in the Annual Survey of Manufactures (ASM), plant-specific lagged values are available for many of the variables in the Census. We did not used lagged values from the ASM in our CART imputations because they are not available for the vast majority of imputed cases in the Census of Manufactures. However, these lagged values are potentially very useful for imputation in studies that focus only on ASM cases.

Our findings also have broader implications. Some of the goals of statistical agencies that collect microdata and publish aggregate statistics are different from the goals of economists and other researchers who use the microdata.

[36] Table 3 summarizes within-industry TFPR dispersion estimates using CART-completed data for every manufacturing industry in 2002 and 2007. It is possible to conduct validity checks of the imputation models for each industry-year. However, given the large number of industries, the rich imputation models we are using, and the large number of datasets required, this would take months of continuous running time on Census Bureau computers. We leave this for future research.

In particular, statistical agencies usually focus less on rich multivariate relationships in the microdata. As a result, statistical agencies' imputations for missing data are often not suitable for multivariate microeconometric analysis. In the past, it was difficult to identify imputed data in many economic microdatasets. Fortunately, in recent years, the U.S. Census Bureau and other agencies have made it easier to identify imputations in their microdata. Our results suggest that using this information and improved imputation procedures like the one presented here would benefit users of these datasets as well as consumers of their research.

References

BAILEY, M., C. HULTEN, AND D. CAMPBELL (1992): "Productivity Dynamics in Manufacturing Plants," in *Brookings Papers on Economic Activity: Microeconomics*, vol. 4, pp. 187–267. Brookings Institute.

BARTELSMAN, E., J. HALTIWANGER, AND S. SCARPETTA (2008): "Cross Country Differences in Productivity: The Role of Allocative Efficiency," Working Paper.

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

BECKER, R., AND V. HENDERSON (2001): "Effect of Air Quality Regulation on Polluting Industries," *Journal of Political Economy*, 108(2), 379–421.

BERNARD, A. B., AND J. B. JENSEN (2004): "Why Some Firms Export," *Review of Economics and Statistics*, 86(2), 561–569.

——— (2007): "Firm Structure, Multinationals, and Manufacturing Plant Deaths," *Review of Economics and Statistics*, 89(2), 193–204.

BERNARD, A. B., S. J. REDDING, AND P. K. SCHOTT (2010): "Multi-Product Firms and Product Switching," *American Economic Review*, 100(1), 70–97.

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. P. 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.

DRECHLER, J., AND J. P. REITER (2011): "An empirical evaluation of easily implemented, nonparametric methods for synthetic datasets," *Computation Statistics and Data Analysis*, 55(2), 3232–3243.

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

ELLISON, G., E. L. GLAESER, AND W. R. KERR (2010): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review*, 100(3), 1195–1213.

ERICSON, R., AND A. PAKES (1995): "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62(1), 53–82.

FOSTER, L., C. GRIM, J. HALTIWANGER, AND Z. WOLF (2015): “Macro and Micro Dynamics of Productivity: Is the Devil in the Details?,” NBER Summer Institute conference paper.

FOSTER, L., J. HALTIWANGER, AND C. KRIZAN (2001): *New Developments in Productivity Analysis*chap. Aggregate Productivity Growth: Lessons from Microeconomic Evidence, pp. 303–372. University of Chicago Press.

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.

GREENESTONE, M. (2002): “The Impacts of Environmental Regulations on Industrial Activiety: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures,” *Journal of Political Economy*, 110(6), 1175–1219.

GRILICHES, Z., AND J. MAIRESSE (1995): “Production Functions: The Search For Identification,” NBER Working Paper 5067.

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.

HSIEH, C.-T., AND P. J. KLENOW (2009): “Misallocation and Manufacturing TFP in China and India,” *Quarterly Journal of Economics*, 74(5), 1403–1448.

JOVANOVIC, B. (1982): “Selection and the Evolution of Industry,” *Econometrica*, 50(3), 649–670.

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 (2002): *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.

MELITZ, M. (2003): “The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity,” *Econometrica*.

OLLEY, S., AND A. PAKES (1996): “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica*, 64(6), 1263–1298.

RAGHUNATHAN, T. E., J. M. LEPKOWSKI, J. V. HOEWYK, AND P. SOLENBERGER (2001): “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models,” *Survey Methodology*, 27(1), 85–95.

REITER, J. P. (2005): “Using CART to generate partially synthetic public use microdata,” *Journal of Official Statistics*, 21(2), 441–462.

REITER, J. P., AND T. E. RAGHUNATHAN (2007): “The multiple adaptations of multiple imputation,” *Journal of the American Statistical Association*, 102, 1462–1471.

RIPLEY, B. (2009): “Tree: classification and regression trees,” cran.r-project.org.

ROBERTS, M. J., AND D. SUPINA (1996): “Output Price, Markups, and Producer Size,” *European Economic Review*, 40(3), 909–921.

——— (2000): “Output Price and Markup Dispersion in Micro Data: The Roles of Producer Heterogeneity and Noise,” in *Advances in Applied Microeconomics, Vol. 9, Industrial Organization*, ed. by M. R. Baye, chap. 4. JAI Press.

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

RUBIN, D. B. (1976): “Inference and Missing Data,” *Biometrika*, 63(3), 581–592.

RUBIN, D. B. (1981): “The Bayesian bootstrap,” *The Annals of Statistics*, 9, 130–134.

SCHAFER, J. L., AND J. W. GRAHAM (2002): “Multiple Imputation for Missing Data: Our View of the State of the Art,” *Pyschological Methods*, 6, 147–177.

SYVERSON, C. (2004a): “Market Structure and Productivity: A Concrete Example,” *Journal of Political Economy*, 112(6), 1181–1222.

——— (2004b): “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.

VAN BIESEBROECK, J. (2004): “Robustness of Productivity Estimates,” NBER Working Paper.

VAN BUUREN, S., AND K. GROOTHUIS-OUDSHOORN (2011): “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, 45(3), 1–67.

WANG, H., AND J. P. REITER (2012): “Multiple imputation for sharing precise geographies in public use data,” *Annals of Applied Statistics*, 6(2), 229–252.

WHITE, T. K. (2014): “Recovering The Item-Level Edit And Imputation Flags In The 1977-1997 Censuses Of Manufactures,” Working Papers CES-WP-14-37, Center for Economic Studies, U.S. Census Bureau.

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

In this online appendix we provide brief descriptions of each imputation method the Census Bureau uses in the Census of Manufactures (tables A1-A2). Tables A4 and A5 present the percentage of observations in the mail samples of the 2002 and 2007 Censuses by type of edit/imputation method for each key variable. This appendix also provides detailed descriptions of the industry and product samples, variable construction, and method for identifying imputed data for the analyses in the main text. The final section of the appendix presents test statistics (table A10) for validity checks of the imputation models described in the main text.

A Industry Definitions and Products

This section provides detailed descriptions of our industry definitions and products. For the industries and years analyzed in tables 4-6, to the extent possible, we follow the industry definitions of Foster, Haltiwanger, and Syverson (2008). We choose industries that produce products for which the Census of Manufactures collects physical quantities of products. For industries for which physical quantity data is collected in both 2002 and 2007, we also require that the data is collected for the same products in both years. We exclude all plants flagged as Administrative Records (AR) cases, since virtually all of the data for these plants is imputed. Following FHS, we also limit the sample to plants for which at least 50% of the plant's revenue is from the product or products that we use the define the plant's industry. As described in FHS, the Census Bureau uses balancing codes to correct for cases where the sum of the values of the plant's products do not sum to the value that the plant reports as it's total value of shipments. Following FHS, we exclude these balancing records when we calculate the plant's specialization. For the industries in tables 4-6, one way in which our samples differ from FHS is in how we deal with imputed data. We use the item-level edit/impute flags to identify imputed data, but we include these plants in our sample. FHS delete plants with data that they identify as imputed. They attempt to identify imputed data by finding plants for which certain ratios are the same as the within-industry-year mode of that ratio. They use the ratios of materials costs over payroll (i.e., salaries and wages), total value of shipments (TVS) over payroll, and product physical quantity over payroll to identify imputed items. This method does not identify most of the data that are imputed in the 2002 and 2007 CM data.

For tables 4-6 we define our industries as follows.

Boxes manufacturing plants in 2002 produce one or more of the following 12 products: corrugated shipping containers for food and beverages (NAICS product code 3222110111), corrugated carryout boxes for retail food (3222110114), corrugated shipping containers for paper and allied products (3222110221), corrugated shipping containers for metal products, machinery, equipment (3222110341), corrugated shipping containers for electrical machinery, equipment (3222110345), corrugated shipping containers for glass, clay, and stone products (3222110431), corrugated shipping containers for chemicals and drugs, including paints,varnishes, cosmetics, and soaps (3222110433), corrugated shipping containers for lumber and wood products, including (3222110435), corrugated shipping containers for all other end uses (3222110437), corrugated paperboard in sheets and rolls, lined and unlined (3222110551), corrugated solid fiber containers (3222110661), and corrugated and solid fiber pallets, pads, and partitions (3222110665). The physical quantity measure for boxes is thousands of square feet.

Bread plants produce white pan bread, not frozen (NAICSPC 3118121111) and/or frozen white pan bread (3118121121). White pan bread is measured in thousands of pounds.

Carbon black plants products carbon black (NAICSPC 3251820100), which is measured in thousands of pounds.

Coffee manufacturing plants produce whole bean roasted coffee (3119201111), ground roasted coffee (3119201211), or ground roasted coffee mixtures (3119201331), all of which are measured in thousands of pounds.

Concrete manufacturing plants in our sample produce ready-mix concrete (3273200100). The Census of Manufactures last collected physical quantity data for this product in 1992. However, the CM does still collect product-level *value* of shipments data for concrete plants. The concrete plants in our sample are highly specialized, with over 90% of the revenue of each plant coming from ready-mix concrete shipments. Syverson (2004a) finds that market-level demand density is correlated with TFP dispersion, so we want to include demand density as a potential predictor in our imputation model for concrete. Following Syverson (2004), we restrict the concrete sample to plants in markets with at least 5 concrete plants.

Hardwood **flooring** plants in our sample produce hardwood oak flooring (3219187111), hardwood oak parquetry flooring(3219187121), other hardwood oak flooring (3219187131), and/or hardwood maple flooring (3219187141). The physical measure is thousands of board feet.

Gasoline plants in our sample produce motor gasoline (3241101121) in 2002. In 2007, this product is disaggregated into regular grade motor gasoline (3241101122), mid-premium grade motor gasoline (3241101123), and premium grade motor gasoline (3241101124). The physical measure in both years is thousands of barrels.

Ice plants produce manufactured can or block ice (3121130111) or manufactured cubed, crushed, or other processed ice (3121130121) in 2002. In 2007 these two products are classified as one product: manufactured ice, (cubed, crushed, etc.), including can or block (3121130100). The physical measure in both years is short tons.

Plywood manufacturing plants in our sample produce hardwood plywood, veneer core (3212113111), hardwood plywood, particleboard core (3212113221), hardwood plywood, medium density fiberboard core (3212113231), and/or hardwood plywood, other core (3212113291). Plywood is measured in thousands of square feet.

Sugar manufacturing plants in our sample produce raw cane sugar (3113110111), which is measured in short tons.

For the analysis presented in table 7, we use the estimation sample from table 6 of FHS, so we are using exactly the same set of industries and products described in detail in FHS's online Appendix.

B Variable Construction

This section provides detailed descriptions of the variables we use in the main text.

B.1 Variables for Tables 3-6

For physical output, we use the physical quantity shipped for that product. For plants in our sample that produce more than one of the products that define our industries, following Foster, Haltiwanger, and Syverson (2008), we aggregate the physical quantities for those products.

We compute prices by dividing the total product value (i.e., the reported revenue from the given product or products) by the physical quantity for that product.

In tables 4-6, for the "Revenue TFP" measure, our output measure is the sum of the plant's total value of shipments, deflated by the shipments deflator for the corresponding industry from the NBER Productivity Database. For the output measure used table 3, changes in inventories are added to the plant's total value of shipments before deflating.

Energy is the sum of the cost of fuels and the cost of purchased electricity. For materials, we use the total cost of intermediate inputs less energy costs. To construct real values for these inputs, we deflate the nominal measures by the energy and materials deflators for the corresponding industry deflators from the NBER Productivity Database.

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.

The 2002 and 2007 Censuses of Manufactures have data on the plant's total book value of assets. We construct real capital stocks by deflating the nominal book values to 2002 levels using sector-specific deflators from the Bureau of Economic Analysis, using the procedure described in Foster, Haltiwanger, and Krizan (2001).

To estimate output elasticities, we use industry-level cost shares. For labor, energy, and materials, we using the industry-level costs for the corresponding industry-year from the NBER Productivity Database. To construct capital costs, we multiply the industry-level capital stocks for equipment and structures in the NBER Productivity Database by the corresponding 3-digit NAICS sector-level rental rates for capital equipment and structures. The capital rental rates are from unpublished data used to construct the Bureau of Labor Statistics' multifactor productivity index.

We construct the market-level demand density variable used in the imputation model for concrete just as described in Syverson (2004a). Concrete demand density within a market is defined as construction sector employment per square mile. We use the Bureau of Economic Analysis's Component Economic Areas (CEA) in 2007 as the market definition. We obtain county-level construction sector employment from the Census Bureau's 2002 and 2007 County Business Patterns published data, and county-level areas from the Census Bureau's City County Data Book. For each year, 2002 and 2007, we compute the average construction sector employment per square mile within each CEA.

B.2 Variables for Tables 7, A7, and A9

For the regressions in tables 7, A7, and A9, we follow the variable construction in Foster, Haltiwanger, and Syverson (2008). For the results in table A7, we use the same final dataset that FHS used, and thus the variable construction is exactly the same as described in FHS.[37] For the results in table A9 (FHS sample without imputes), we had to reconstruct two of FHS's variables. First, after dropping plants in FHS's sample with imputed data, following FHS, we adjust plant-level unit prices to a common 1987 basis using the revenue-weighted geometric mean of the product prices across all of the plants producing the product in the remaining sample. After adjusting prices to a 1987 basis, we follow the procedure described in FHS to construct plant-specific demand shocks. As described in FHS, one can think of these as residuals from IV estimation of product-specific demand systems, with estimated local income affects added back in.

For the results presented in table 7, we use exactly the same sample of plants as in table A7. However, since we are replacing imputed items in the FHS sample, we reconstruct all the variables in the FHS data that are constructed from imputed data. For example, since unit prices are constructed from plant-level product quantity shipped (PQS) and plant-level product value shipped, after we replace the Bureau's imputations for PQS, we reconstruct the unit prices for the plants with imputed PQS. Before calculating TFPQ, FHS also scale the plant's PQS by dividing by the plant's specialization ratio. So after replacing the Bureau's PQS imputations with CART imputations, we use the

[37] We thank Lucia Foster, John Haltiwanger, and Chad Syverson for giving us their computer code as well as providing access to their data.

plant-level specialization ratio to reconstruct the plant's physical output measure, which we use for recalculating the plant's TFPQ. Importantly, after replacing the Bureau-completed data with CART imputations, we re-adjust the plant-level prices to a 1987 basis using the revenue-weighted geometric mean of the product prices across all of the plants producing the product in the sample. Using the quantities and adjusted prices, we recompute the idiosyncratic demand shocks following the procedure described in FHS. We adjust the prices and compute the idiosyncratic demand shocks separately for each CART-completed dataset.

C Identifying Imputed Data

In this section 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 the most important variables in the Census of Manufactures. We use the item's edit/impute flag variables to determine whether or not an item was imputed. We define an observation as imputed if it meets the criteria below based on the edit/impute flags.

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. Each variable 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 and retirements. 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 identify these "non-imputed" items as imputes.

Finally, in 2007, the edit/impute flag for the end of year assets variable (TAE) indicates that 98.7% of the TAE observations were imputed using a "Beta Cold Deck Statistical" method. This flag was set as part of a mass edit of the capital asset variables during 2007 Economic Census processing. Due to a misinterpretation of the depreciable asset questions on the 2007 questionnaire, many respondents incorrectly totaled the end-of-year assets, which are supposed to be equal to beginning of year assets plus capital expenditures less capital retirements and depreciation. To correct this, the Census Bureau ran a mass edit which correctly totaled the assets variables, but in the process set the second character of the edit/impute flags for the asset variables to 'B'. As a result, nearly every TAE observation in the 2007 mail sample has the second character of the edit/impute flag set to 'B'. We cannot distinguish between asset flags that were set by the mass edit versus those that were imputed by normal processing. However, as noted above, the first character of the edit/impute flag tells us whether or not the item was reported on the survey form ('R'), or not reported (indicated by a blank). For the imputation rates reported in Table 1 we only identify a capital asset observation as imputed if it was not reported on the survey form and the second character in the edit/impute flag was set to 'B'. End of year asset imputations identified this way account for 31% of observations in the mail sample in 2007.

The Census Bureau uses different imputation methods for different variables. Appendix tables A4 and A5 summarize the percentage of observations that are imputed/non-imputed by type of imputation method for key variables in the 2002 and 2007 Censuses of Manufactures. Items flagged as "Beta Cold Deck Statistical" imputes are constructed from one of two regression models. The first regression model uses a single current-year explanatory variable and an industry-specific regression parameter:

$$Y_{it}{}^{imputed} = \beta_{1j} X_{it} \tag{5}$$

where Y_{it} is the (imputed) observation of variable Y for plant i in year t, β_{1j} is a regression parameter for industry j, and X_{it} is the observation of variable X for the same plant i in year t. These imputations are labeled "univariate regression" in tables A4 and A5. They use only current-year data. For example, to impute for the total cost of materials, the Census Bureau uses the plant's total value of shipments as the explanatory variable in equation 5. The second type of regression model the Census Bureau uses includes three explanatory variables — one current-year variable, the prior-year value of the same variable, and the prior-year value of the variable being imputed:

$$Y_{it}{}^{imputed} = \beta_{2j} X_{it} + \beta_{3j} X_{i,t-1} + \beta_{4j} Y_{i,t-1} \tag{6}$$

For example, to impute for year t cost of materials using the model in equation 6, the Census Bureau uses the plant's year t total value of shipments, year $t-1$ total value of shipments, and year $t-1$ total cost of materials.

The item-level edit/impute flags in the Census of Manufactures do not tell us *which* regression model was used — the flags only tell us that one of the two models was used. However, the 3-variable model in equation (6) can be used only if prior-year data for the relevant variables are available for the same plant. For example, to impute total cost of materials using equation (6), prior-year ASM data on total cost of materials and total value of shipments for the same plant must be available. In its "Beta Cold Deck Statistical" regression imputations for the CM, the Census Bureau always tries to use the model in equation (6) before resorting to the model in equation 5. Accordingly, for a given item with a "Beta" edit/impute flag, if the relevant prior-year data is available for that plant, then we assume that model (6) was used. If the relevant prior-year data is not available for that plant, then we assume model (5) was used. This approach is conservative in the sense that we only identify a regression-imputed item as being imputed with model (5) if it was impossible to impute it using model (6).[38]

In 2002, 60% to 89% of *all* imputations for total value of shipments, the cost of materials, production worker wages, beginning-of-year inventories, and end-of-year inventories, were imputed using a univariate regression model with only current-year data (i.e., no lagged values). In 2007, for these variables, the percentage of all imputations using univariate regression with only current-year data ranges from 58% to 78%.

C.1 Imputed data in the Foster, Haltiwanger, and Syverson (2008) Sample

Using ratios of certain variables (the cost of parts and materials over payroll, total value of shipments over payroll, and physical quantity shipped over payroll), FHS identified plants with imputed values for materials and total values of

[38] Using a slightly different approach, Foster, Grim, Haltiwanger, and Wolf (2015) (available at http://conference.nber.org/confer/2015/SI2015/PRCR/Foster_Grim_Haltiwanger_Wolf.pdf) estimate that 53% of TVS imputations, 65% of cost of material imputations, and 76% of production worker hours imputations in the 2007 Census of Manufactures were imputed using univariate regression models (see table A18 in that paper).

shipments and dropped those plants from their sample. To facilitate comparison, we use the same sample of plants that FHS use, so we are effectively also dropping those plants. The Census Bureau also imputes for missing product-level data on physical quantities (PQS). Regarding these PQS imputes, the Appendix of FHS (section A.3) says:

> The Census Bureau imputes physical quantities when product-level data are not fully reported. Unfortunately, imputed data are not explicitly identified. To distinguish and remove imputed product-level data from the sample, we use techniques similar to those employed by Roberts and Supina (1996, 2000).

To identify imputed product-level data (PQS), Roberts and Supina (1996, 2000) looked at the modal ratio of product value (PV) over product quantity shipped (PQS) — the unit price — in each industry. Under the assumption that plants with the modal price have imputed PQS data, they dropped these plants from their sample. Lucia Foster, John Haltiwanger, and Chad Syverson also conducted some sensitivity analysis of the main results in FHS (2008). Using several different assumptions (including several assumptions regarding rounding errors) to identify imputed PQS values, and they find that the main results in table 6 of FHS (2008) are robust to dropping these cases.[39]

Before we became aware of the Census Bureau data files with product-level edit/impute flags, we also tried identifying imputed PQS data using the within-product-year modal prices. However, as mentioned in the sensitivity analysis conducted by FHS, rounding errors introduce some uncertainty into this reverse-engineering approach to identifying imputed data. First, the Bureau rounds the product-level average price to the fourth decimal place. Second, the imputed PQS value is rounded to the nearest unit value (zero decimal places).[40] The implied price for a plant with a rounded, imputed PQS value may not be close to the modal price. Because of these rounding errors, the reverse-engineering approach fails to identify some imputed cases. Rounding error is greater for smaller values of PQS, and thus these imputations are less likely to be identified by looking at the modal price in an industry. This is important, because we know from the 2002-2007 data that smaller plants (which tend to have smaller PQS values) are more likely to have missing data. The product-level edit/impute flags allow us to identify imputed data without the uncertainty associated with the reverse-engineering approach used in previous research (i.e., looking for modal ratios).

Like Roberts and Supina (1996, 2000), FHS looked at the modal unit price, but FHS did not actually drop these plants from their estimation sample.[41] We used plant-level numeric identifiers in both files to match the product-level edit/impute flags to FHS's sample. As expected, 100% of the plants in the FHS sample matched to the flags datasets. Using the edit/impute flags we found that roughly 50% of the plants in the FHS sample have imputed PQS data. In the ready-mix concrete industry (which accounts for about two thirds of their sample), in 1977 roughly 40% of the concrete PQS observations' in FHS's sample were imputed and in 1982, 1987, and 1992 roughly 55% were imputed.

How accurate are reverse-engineering approaches in identifying imputes in this sample? To answer this question we calculated the percentage of imputed PQS observations (as identified by the edit/impute flags) that are correctly identified by reverse-engineering using the assumptions about rounding that are used in the FHS sensitivity analysis. To adhere to Census Bureau disclosure rules regarding sample sizes, here we focus on the concrete industry. Table A8 presents the results. When PQS imputes are identified as plants with the product-year mode of the plant-level unit price (without rounding), in 1977, 1982, 1987, and 1992, respectively 5%, 1%, 3%, and 2% of the PQS imputes are correctly identified. If we first round the unit price to the nearest 0.002 or 0.005 before finding the mode, one percentage point

[39] The sensitivity analysis conducted by Foster, Haltiwanger, and Syverson is available on the internet at http://home.uchicago.edu/syverson/modalpricerobust.pdf.

[40] Richard Williamson of the Census Bureau's Manufacturing and Construction Division explained this in an email to one of the authors on January 8, 2014.

[41] In an email to one of the authors on January 6, 2014, John Haltiwanger confirmed our finding that FHS did not drop plants with imputed PQS values in the published version of the paper.

more of the PQS imputes are correctly identified in 1987 and 1992. Finally, if we use 1 over the product-year median PQS as the rounding factor, in 1977, 1982, 1987, and 1992, respectively 7%, 3%, 4%, and 3% of the PQS imputes are correctly identified. One can increase the number of imputes that are identified by increasing the rounding factor, but of course this will also increase the number of false positives.

D Validity Checks

To check the validity of our imputation models for the analyses above, we use posterior predictive checks (He, Zaslavsky, Harrington, Catalano, and Landrum (2010)). We now provide a formal description of the validity checks. 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 dataset in each pair be called the predicted datasets. We then estimate the parameter of interest — the within-industry productivity or price dispersion or a marginal effect — 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:

$$P_j = \frac{2}{500} min\{\sum_{i=1}^{500} I(\widehat{\theta}_{imp,ij} - \widehat{\theta}_{pred,ij}), \sum_{i=1}^{500} I(\widehat{\theta}_{pred,ij} - \widehat{\theta}_{imp,ij})\} \tag{7}$$

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 marginal effect or within-industry 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_j 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 data, and thus estimates based on the imputed data may be biased.

We calculate P for each measure of productivity and/or price dispersion in tables 5 and 6 and for the exit probit results presented in table 7. For 24 of the 33 dispersion moments estimated in table 5, the associated P probabilities are greater than 0.05, and 20 of them are greater than 0.10.[42] In the few cases where there is evidence of a bias, the biases tend to be small. For example, the revenue TFP dispersion for the boxes industries tends to be only about 5 percentage points higher in the predicted data than in the completed data, and there is no evidence of bias in the physical TFP or price dispersion statistics for the boxes industry. For the two concrete productivity dispersion moments estimated in table 6, we also find evidence of a small upward bias from the CART model. In 500 CART-predicted datasets, the maximum bias in the TFPR dispersion estimate is 18 percentage points (in both 2002 and 2007), and the mean biases are 13 percentage points (in the 2002 data) and 14 percentage points (2007). Thus the upward bias created by the CART models does not explain most of the 42 or 46 percentage point difference between TFPR dispersion in the CART-completed data versus TFPR dispersion in the Bureau-completed data.

[42] We provide the full set of P values in table A10 in the appendix.

For the exit probits in table 7, the P probabilities associated with the marginal effect estimates are all between 0.6 and 0.9. Thus we find no evidence that the CART imputations are leading to biased estimates of the marginal effects. Together this evidence suggests that the CART model generates plausible data with respect to most of the estimated relationships represented in tables 5, 6, and 7.

Table A1: Edit/Impute Flags in the 2002 and 2007 Census 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)

Table A3: Definitions of Key Variables in the Census of Manufactures

Variable Name	Definition
SW	Total Salaries and Wages (all employees)
SW-NL	Total Salaries and Wages, non-leased employees only (2002 only)
TE	Total Number of Employees (March 12)
TE-NL	Number of Non-leased Employees (March 12) (2002 only)
TVS	Total Value of Shipments
CM	Total Cost of Materials
CF	Cost of Fuels
EE	Cost of Purchased Electricity
PH	Total Number of Production Worker Hours
PH-NL	Production Worker Hours of Non-leased Employees (2002 only)
WW	Production Worker Wages
WW-NL	Wages of Non-leased Production Workers (2002 only)
TIB	Total Inventories, Beginning of Year
TIE	Total Inventories, End of Year
TAE	Book Value of Assets, End of Year

Table A4: Percentages of Observations by Type of Edit/Impute Flag,
2002 Census of Manufactures Mail Sample

	SW-NL	TE	TE-NL	TVS	CM	CF	EE
From Ad Rec Data	2.5	0.0	0.1	0.5	0.0	0.0	0.0
Univariate regression	0.0	15.5	15.3	17.8	30.4	0.0	0.0
Trivariate regression	0.0	4.6	4.1	4.3	5.3	0.0	0.0
Industry Average Ratio	0.0	0.0	0.0	0.3	0.0	43.3	43.3
Other impute	15.6	0.1	23.5	7.0	7.6	4.3	5.2
Other non-impute	2.1	0.0	1.5	3.6	2.3	8.0	5.0
Edit/impute flag missing	18.9	79.8	0.0	0.5	0.7	0.0	0.0
Reported, edit-passing	60.9	0.0	55.5	66.1	53.7	44.5	46.6
Total % imputed	18.1	20.2	43.0	29.9	43.3	47.6	48.5

	PH	PH-NL	WW	WW-NL	TIB	TIE	TAE
From Ad Rec Data	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Univariate regression	15.2	38.7	15.4	41.0	32.0	30.3	0.0
Trivariate regression	4.0	5.1	4.4	5.0	4.2	5.7	0.0
Industry Average Ratio	0.0	0.0	0.0	0.0	0.2	0.1	0.0
Other impute	22.2	19.4	0.0	0.0	6.8	6.2	0.0
Other non-impute	0.0	1.7	0.2	2.3	2.0	1.5	41.1
Edit/impute flag missing	0.0	0.2	79.9	0.3	2.9	2.1	0.0
Reported, edit-passing	58.6	35.0	0.0	51.3	52.0	54.2	58.9
Total % imputed	41.4	63.2	19.9	46.1	43.2	42.3	0.0

The table shows the percentages of observations by type of edit/impute flag for key variables in the mail sample of the 2002 Census of Manufactures. In the 2002 Census, the impute flag for salaries and wages of all employees (not shown) is blank for almost all plants. See table A3 for variable definitions.

Table A5: Percentages of Observations by Type of Edit/Impute Flag,
2007 Census of Manufactures Mail Sample

	SW	TE	TVS	CM	CF	EE
From Ad Rec Data	13.8	0.1	0.3	0.0	0.0	0.0
Univariate regression	0.0	12.1	16.3	27.3	0.0	0.0
Trivariate regression	0.0	4.1	4.2	5.8	0.0	0.0
Industry Average Ratio	0.0	0.0	0.2	0.0	39.1	40.2
Other impute	16.7	34.1	7.4	7.9	5.9	6.2
Other non-impute	5.5	1.2	7.1	2.8	1.2	1.7
Edit/impute flag missing	0.0	0.0	0.0	0.6	7.4	3.1
Reported, edit-passing	64.0	48.3	64.4	55.6	46.4	48.8
Total % imputed	30.5	50.4	28.5	41.0	45.1	46.4

	PH	WW	TIB	TIE	TAE
From Ad Rec Data	0.0	0.4	0.0	0.0	0.0
Univariate regression	47.3	33.8	32.9	31.6	98.7
Trivariate regression	7.5	4.9	3.5	6.0	0.0
Industry Average Ratio	0.0	0.0	0.0	0.0	1.0
Other impute	4.0	4.1	16.4	3.1	0.0
Other non-impute	1.4	2.3	5.1	3.4	0.2
Edit/impute flag missing	0.0	0.3	0.8	2.3	0.0
Reported, edit-passing	39.8	54.2	41.3	53.6	0.1
Total % imputed	58.8	43.3	52.8	40.7	99.7

The table shows the percentages of observations by type of edit/impute flag for key variables in the mail sample of the 2007 Census of Manufactures. See table A3 for variable definitions. Note: in 2007 nearly all observations for TAE were flagged as regression imputes due to a reporting anomaly. See text for details.

Table A6: Within-industry-year Productivity and Price Dispersion
Foster, Haltiwanger, and Syverson (2008) Concrete Sample

	(1)	(2)	(3)	(4)	(5)	(6)
Dataset	*Bureau-completed*		*Non-imputed only*		*CART-completed*	
	Sample	75-25	Sample	75-25	Sample	75-25
Year	Size	Ratio	Size	Ratio	Size	Ratio
			Revenue TFP			
1977	2184	1.25	743	1.24	2184	1.33
1982	3316	1.26	1145	1.25	3316	1.31
1987	3236	1.22	1305	1.33	3236	1.35
1992	3427	1.26	1374	1.31	3427	1.37
			Physical TFP			
1977	2184	1.32	743	1.33	2184	1.41
1982	3316	1.28	1145	1.31	3316	1.36
1987	3236	1.23	1305	1.35	3236	1.39
1992	3427	1.29	1374	1.37	3427	1.42
			Prices			
1977	2184	1.08	743	1.18	2184	1.20
1982	3316	1.05	1145	1.19	3316	1.24
1987	3236	1.01	1305	1.25	3236	1.34
1992	3427	1.00	1374	1.22	3427	1.30

The table shows ratios of the 75th percentile to the 25th percentile of within-industry-year distributions of productivity and prices and sample sizes for three different datasets. Columns 2, 4, and 6 show, respectively, the ratios calculated from Census Bureau-completed data, plants with non-imputed data, and data in which the Bureau's imputations are replaced with CART imputations.

Table A7: Selection on Productivity or Profitibility, 1977-1992

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Traditional TFP	-0.073						
	(0.015)						
Revenue TFP		-0.063					
		(0.014)					
Physical TFP			-0.040			-0.062	-0.034
			(0.012)			(0.014)	(0.012)
Prices				-0.021		-0.069	
				(0.018)		(0.021)	
Demand shock					-0.047		-0.047
					(0.03)		(0.003)
	Controlling for plant capital stock						
Traditional TFP	-0.069						
	(0.015)						
Revenue TFP		-0.061					
		(0.013)					
Physical TFP			-0.035			-0.059	-0.034
			(0.012)			(0.014)	(0.012)
Prices				-0.030		-0.076	
				(0.018)		(0.021)	
Demand shock					-0.030		-0.029
					(0.004)		(0.004)
Capital Stock	-0.046	-0.046	-0.046	-0.046	-0.023	-0.046	-0.023
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)

This table replicates table 6 in Foster, Haltiwanger, and Syverson (2008). The table shows marginal effects evaluated at the median for probits of plant exit by the next census (presented by column) on plant-level productivity, price, demand, and capital stocks measures. All regressions include product-year fixed effects. Standard errors (clustered by plant) are in parentheses. The sample is FHS's pooled sample of 17,314 plant-year observations.

Table A8: Percentage of Physical Quantity Imputes Correctly Identified by Reverse-Engineering in the Foster, Haltiwanger, and Syverson (2008) Concrete Sample

Reverse-engineering Method	1977	1982	1987	1992
Modal Price (no rounding)	5%	1%	3%	2%
Mode of Rounded Price (nearest 0.002)	5%	1%	4%	3%
Mode of Rounded Price (nearest 0.005)	5%	1%	4%	3%
Mode of Rounded Price (nearest 1/(Median(PQS))	7%	3%	4%	3%

The table shows the percentage of imputes for physical quantity of shipments (PQS) correctly identified in each year of the sample by reverse-engineering using various rounding factors. For the "no rounding" method, PQS imputes are identified as plants with the mode of the plant-level unit price (product value shipped divided by PQS). For each of the other methods, the unit price is rounded before finding the mode. Imputes identified using these methods are then compared to the impute flags for the same plants.

Table A9: FHS Exit Probits Using Only Non-imputed Data

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Traditional TFP	-0.036						
	(0.021)						
Revenue TFP		-0.036					
		(0.019)					
Physical TFP			-0.027			-0.039	-0.020
			(0.019)			(0.019)	(0.015)
Prices				-0.004		-0.026	
				(0.019)		(0.024)	
Demand shock					-0.041		-0.041
					(0.004)		(0.004)
	Controlling for plant capital stock						
Traditional TFP	-0.046						
	(0.020)						
Revenue TFP		-0.048					
		(0.019)					
Physical TFP			-0.027			-0.049	-0.022
			(0.015)			(0.019)	(0.015)
Prices				-0.008		-0.046	
				(0.018)		(0.024)	
Demand shock					-0.028		-0.027
					(0.006)		(0.006)
Capital Stock	-0.035	-0.035	-0.035	-0.035	-0.016	-0.035	-0.016
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)

The table shows marginal effects evaluated at the median for probits of plant exit by the next census (presented by column) using the same specifications as in tables A7 and 7. The sample is the 7,404 plant-year observations in the table A7 sample which use only non-imputed data for any variable except payroll or employment. Standard errors (clustered by plant) are in parentheses.

Table A10: Validity Checks of the CART Imputation Models:
Within-industry Productivity and Price Dispersion

		P value for:		
		75-25	75-25	75-25
		TFPR	TFPQ	Price
industry	Year	Ratio	Ratio	Ratio
boxes	2002	0.000	0.176	0.348
bread	2002	0.280	0.468	0.148
carbon black	2002	0.552	0.204	0.056
coffee	2002	0.272	0.508	0.440
concrete	2002	0.000	N/A	N/A
concrete	2007	0.000	N/A	N/A
flooring	2002	0.164	0.404	0.532
gasoline	2002	0.008	0.000	0.000
gasoline	2007	0.076	0.000	0.000
ice	2002	0.012	0.076	0.088
ice	2007	0.192	0.552	0.416
plywood	2002	0.008	0.476	0.664
sugar	2002	0.168	0.196	0.000

The table shows P probabilities (see equation 5 in the main text) for revenue TFP, physical TFP, and price dispersion measures by industry-year. A probability close to zero is evidence that the CART imputation model distorts the joint distribution of the data for that industry-year such that the given dispersion estimate may be biased.

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	27%	31%	19%	38%	37%	42%
	s.d.	9%	10%	7%	14%	14%	10%
2007	Mean	27%	32%	31%	37%	35%	42%
	s.d.	9%	10%	13%	13%	12%	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 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: Industry Distributions of Within-industry Productivity Dispersion, 2002 and 2007 Censuses of Manufactures

	25th percentile	median	mean	75th percentile
			2002	
Census Bureau-completed data	1.34	1.42	1.44	1.53
Non-imputed data	1.46	1.57	1.61	1.72
CART-completed data	1.54	1.65	1.71	1.81
			2007	
Census Bureau-completed data	1.39	1.48	1.53	1.61
Non-imputed data	1.49	1.60	1.65	1.79
CART-completed data	1.55	1.68	1.76	1.87

The table shows ratios of the 75th percentile to the 25th percentile of within-industry-year distributions of revenue TFP for industries at the 25th, 50th and 75th percentiles and means of the industry distributions of revenue TFP dispersion in the 2002 and 2007 mail samples of the Censuses of Manufactures, using three different datasets: (i) the Census Bureau-completed data, in which missing or faulty data was imputed by the Census Bureau using a variety of methods; (ii) a "non-imputed" sample, which excludes plants for which any variable needed to calculate TFPR was imputed using the industry average ratio method or univariate regression on current-year data; (iii) the CART-completed data, in which variables in (i) that were imputed by industry average ratio or univariate regression are replaced by CART imputations.

Table 4: Imputation Rates for Key Variables, FHS Industries

	FHS industries, except concrete	concrete, 2002	concrete, 2007
Sample Size	1453	3294	4961
Value of Shipments	12%	10%	21%
Quantity of Product	45%	NA	NA
Book Value of Assets	13%	9%	42%
Production Worker Hours	8%	3%	33%
Cost of Electricity	11%	20%	39%
Cost of Fuels	10%	19%	37%
Cost of Materials	19%	18%	39%

The imputation rate is the percentage of cases in the sample that are imputed by the Census Bureau. Excluding concrete, the FHS industries are boxes, white pan bread, carbon black, coffee, hardwood flooring, motor gasoline, ice, plywood, and sugar.

Table 5: Productivity and Price Dispersion, FHS Industries

			75-25 TFPR Ratios		*75-25 TFPQ Ratios*		*75-25 Price Ratios*	
			(1)	(2)	(3)	(4)	(5)	(6)
		Sample	Census		Census		Census	
industry	Year	Size	Bureau	CART	Bureau	CART	Bureau	CART
boxes	2002	626	1.17	1.18	1.90	2.13	1.86	2.04
bread	2002	71	1.79	1.59	2.10	2.10	1.09	1.53
carbon black	2002	21	1.45	1.54	1.45	1.84	1.09	1.60
coffee	2002	98	1.15	1.67	1.32	2.40	1.07	1.72
flooring	2002	40	1.35	1.39	1.81	2.13	1.26	2.01
gasoline	2002	73	1.12	1.15	1.15	1.18	1.08	1.10
gasoline	2007	61	1.18	1.19	1.16	1.21	1.05	1.07
ice	2002	169	1.48	1.61	1.67	2.11	1.15	1.73
ice	2007	237	1.68	1.78	1.93	2.75	1.11	2.37
plywood	2002	36	1.26	1.29	1.89	3.57	1.50	3.00
sugar	2002	21	1.31	1.62	1.40	1.62	1.02	1.09

The table shows ratios of the 75th percentile to the 25th percentile of within-industry-year distributions of total factor productivity (TFP) and prices. TFPR is a revenue-based TFP measure. TFPQ is based on the physical quantity of output. Columns 1, 3, & 5 show estimates from the Census Bureau-completed data. Columns 2, 4, & 6 show the means of estimates from 500 CART-completed datasets.

Table 6: Imputation and Measured Productivity Dispersion, Ready-Mix Concrete

	(1)	(2)	(3)	(4)	(5)	(6)
Dataset	*Bureau-completed*		*Non-imputed only*		*CART-completed*	
	Sample	75-25 TFPR	Sample	75-25 TFPR	Sample	75-25 TFPR
Year	Size	Ratio	Size	Ratio	Size	Ratio
2002	3294	1.33	2272	1.33	3294	1.79
2007	4961	1.30	2309	1.30	4961	1.72

The table shows ratios of the 75th percentile to the 25th percentile of within-industry-year distributions of Revenue TFP and the respective sample sizes for three different datasets. Column 2 shows the ratios calculated from Census Bureau-completed data. Column 4 show the ratios when plants with imputed data are dropped from the sample. Column 6 shows the ratios when the Bureau's imputations are replaced multiple imputations using the sequential CART method described in the text.

Table 7: FHS Exit Probits Using CART-completed Data

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Traditional TFP	-0.036						
	(0.015)						
Revenue TFP		-0.042					
		(0.014)					
Physical TFP			-0.025			-0.047	-0.024
			(0.012)			(0.015)	(0.012)
Prices				-0.005		-0.036	
				(0.013)		(0.016)	
Demand shock					-0.054		-0.054
					(0.003)		(0.003)
	Controlling for plant capital stock						
Traditional TFP	-0.035						
	(0.015)						
Revenue TFP		-0.033					
		(0.013)					
Physical TFP			-0.024			-0.040	-0.024
			(0.011)			(0.014)	(0.011)
Prices				0.001		-0.025	
				(0.012)		(0.015)	
Demand shock					-0.041		-0.041
					(0.005)		(0.005)
Capital Stock	-0.046	-0.045	-0.046	-0.046	-0.014	-0.045	-0.014
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.003)	(0.005)

The table shows marginal effects evaluated at the median for probits of plant exit by the next census (presented by column) on plant-level productivity, price, demand, and capital stocks measures. All regressions include product-year fixed effects. The regressions are run separately on each of 500 datasets, where the imputed data in the FHS sample used in table A7 are replaced by multiple imputations using the sequential CART method described in the text. The marginal effects shown are the means of the 500 estimates. Standard errors (clustered by plant) from each regression are combined using Rubin's (1987) combining formulas.