NBER WORKING PAPER SERIES

IMPUTATION IN U.S. MANUFACTURING DATA AND ITS IMPLICATIONS FOR PRODUCTIVITY DISPERSION

T. Kirk White
Jerome P. Reiter
Amil Petrin

Working Paper 22569
http://www.nber.org/papers/w22569

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2016

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 or the National Bureau of Economic Research. All results have been reviewed to ensure that no confidential information is disclosed. Reiter gratefully acknowledges support from National Science Foundation grant SES 1131897.

Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion
T. Kirk White, Jerome P. Reiter, and Amil Petrin
NBER Working Paper No. 22569
August 2016
JEL No. C80,L11,L60

ABSTRACT

In the U.S. Census Bureau's 2002 and 2007 Censuses of Manufactures 79% and 73% of observations respectively have imputed data for at least one variable used to compute total factor productivity. The Bureau primarily imputes for missing values using mean-imputation methods which can reduce the true underlying variance of the imputed variables. For every variable entering TFP in 2002 and 2007 we show the dispersion is significantly smaller in the Census mean-imputed versus the Census non-imputed data. As an alternative to mean imputation we show how to use classification and regression trees (CART) to allow for a distribution of multiple possible impute values based on other plants that are CART-algorithmically determined to be similar based on other observed variables. For 90% of the 473 industries in 2002 and the 84% of the 471 industries in 2007 we find that TFP dispersion increases as we move from Census mean-imputed data to Census non-imputed data to the CART-imputed data.

T. Kirk White
U.S. Census Bureau
tkirkwhite@gmail.com

Jerome P. Reiter
Duke University
jerry@stat.duke.edu

Amil Petrin
Department of Economics
University of Minnesota
4-101 Hanson Hall
Minneapolis, MN 55455
and NBER
petrin@umn.edu

1 Introduction

Nearly all economic surveys suffer from item non-response. Most statistical agencies impute for the missing values before making data available for analyses and it is well known that the manner of imputation may impact these analyses (Little and Rubin (2002)). We investigate the extent of imputation in the U.S. Census Bureau's Census of Manufactures (CM) and document its impact on the measured dispersion in total factor productivity, which is already thought to be large (see Syverson (2011)). Our results may have implications for the many highly cited studies that use plant-level U.S. Census manufacturing data, including research on 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 (2010)), industry agglomeration (Ellison, Glaeser, and Kerr (2010)), and firm structure and plant exit (Bernard and Jensen (2007)).

Item non-response has been an important issue for the U.S. Census of Manufacturers. In 2002 imputation rates ranged from between 20 and 40 percent for important production variables and 2007 is similar.[1] The Census Bureau primarily imputes missing data using industry average ratios or univariate regressions. Both methods impute towards the mean of the data in the sense that all plants that are missing a value for variable Y (like total value of shipments) have the same imputed value Y^{imp} if they are of the same "type," where type is determined by the value of another single variable X (like total employment). For every variable entering TFP in 2002 and 2007 we find the dispersion is significantly smaller in the Census mean-imputed versus the Census non-imputed data.

As an alternative to mean imputation we show how to use classification and re-

[1]In calculating these imputation rates, we exclude the administrative records as researchers typically do.

gression trees (CART) from Burgette and Reiter (2010) to allow for multiple possible impute values Y^{imp} for any "type," and to allow for *multiple* possible explanatory variables when determining plant type.[2] Manufacturing plants of the same "type" live on the same "leaf" of the classification tree and the distribution of possible impute values is taken from all of the plants on that leaf. Impute values are drawn using sampling with replacement and when all missing values have been filled in the data set is said to be "CART-completed". Repeating this process M times yields M CART-completed data sets. For any statistic of interest like TFP dispersion calculating its value across the M CART-completed data sets serves as a measure of the uncertainty introduced by imputing missing values.

Ex ante it is not obvious how the significant reduction in dispersion we observe from mean-imputation affects total factor productivity (TFP) because TFP is a ratio of output over an input index. We examine dispersion in TFPR, where output is defined as deflated revenue, and TFPQ, where output is quantity produced, and in unit prices across three variants of the Census data: only non-imputed data, (mean imputed) Census-completed data currently used by researchers, and CART-completed data. For 90% of the 473 industries in 2002 and 84% of the 471 industries in 2007 the 75-25 percentile ratio increases as we move from Census-completed to non-imputed to CART-completed data. In the CART-completed data 66% (2002) and 51% (2007) of industries have 75-25 TFPR ratios that are at least 10 log points higher than in the Census-completed data, suggesting TFPR has *more* dispersion than has been currently thought. For the small collection of industries where we observe quantities we find on average TFPQ dispersion is 27% higher and price dispersion is 58% higher in the CART-completed data relative to the mean-imputed Census data, and the non-imputed data

[2]See Little and Rubin (2002) for discussion of the potential benefits of multiple imputation over mean imputation.

lie approximately halfway between the dispersion estimates from the mean-imputed Census data and the CART-completed data.

We also revisit Foster, Haltiwanger, and Syverson (2008), who report negative and significant relationships between plant exit and TFPR, TFPQ, prices, and idiosyncratic demand shocks. These findings are important in part because they are very much in the spirit of an important theoretical literature on firm dynamics analyzing the connection between producers' productivity, demand, product quality, and survival (e.g. Jovanovic (1982), Ericson and Pakes (1995), Melitz (2003)). We show FHS's results are very robust to CART-imputation.

Our results have implications for the many highly cited studies that use plant-level U.S. Census manufacturing data. For questions related to average effects - like regression coefficients - the reduction in dispersion caused by mean imputation may not be problematic, although there is no way to tell without also trying an alternative. There can also be an issue with bias in the estimated standard errors (Little and Rubin (2002)). For questions related to dispersion in total factor productivity - like the result from Hsieh and Klenow (2009) that both India and China would experience an increase of over 30% in growth if they could move to U.S. sized "gaps" - researchers may be better served by using a method like CART instead of mean-imputation.

The next section examines the extent of imputation in the U.S. Census of Manufactures. Section 3 discusses the CART method and Section 4 contains the results. Section 5 concludes.

2 Imputed Data in the U.S. Census of Manufactures

The Census of Manufactures is taken every five years and includes data on over 200,000 manufacturing plants in the United States.[3] Historically item non-response has been an issue for Census data.[4] For the 2002 and 2007 censuses Table 1 presents the means and standard deviations of the within-industry imputation rates for several variables in all 6-digit NAICS industries, the most detailed level of industry classification in the Census data. In 2002 imputation rates for these variables range from a low of 19% for production worker hours to a high of 42% for the cost of materials. In 2007 imputation rates range from a low of 27% for the value of shipments to a high of 42% for the cost of materials. If output is measured using the total value of shipments adjusted for inventory changes and inputs in production include capital, labor, energy and materials, then a researcher wanting to only use non-imputed data would lose 79% and 73% of plant-year observations in 2002 and 2007 respectively.[5]

The Census Bureau primarily uses industry average ratios and univariate regressions to impute missing data.[6] Both methods impute towards the mean of the data in the

[3]There are over 300,000 plants in the survey but the smallest 100,000 plants have data that is almost entirely imputed and so are routinely excluded from Census data analysis. These plants are known as the "administrative records" plants

[4]Prior to the 2002 census researchers did not have access to the item-level imputation flags. Researchers interested in figuring out which data were imputed developed several approaches (see Roberts and Supina (1996), Roberts and Supina (2000), or Foster, Haltiwanger, and Syverson (2008)). White (2014) uses these recently recovered item-level impute flags to show the complete extent of imputation in the Census data.

[5]Foster, Grim, Haltiwanger, and Wolf (2015) redid this calculation excluding inventories and the book value of assets and report 68.8% (2002) and 69.2% (2007) of observations would have to be omitted.

[6]In 2007 these two methods were used to impute for the 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. See tables A1-A5 in White, Reiter, and Petrin (2015) for a complete discussion of Census imputation

sense that all plants that are missing a value for variable Y have the same imputed value Y^{imp} if they are of the same "type," where type is determined by the value of another single variable X. Letting i index plants the industry average ratio for the n_j plants in industry j for which both Y_i and X_i are observed is given as $\frac{1}{n_j}\sum_i \frac{Y_i}{X_i}$. If observation Y_l is missing for plant l in industry j the industry average ratio method used by Census imputes Y_l^{imp} by multiplying that plant's X_l by the industry average ratio:

$$Y_l^{imp} = X_l * \frac{1}{n_j}\sum_i \frac{Y_i}{X_i}. \tag{1}$$

Their univariate regression imputation uses a no-intercept regression of Y_i on X_i for the plants i in industry j for which both Y_i and X_i are observed to predict Y_l^{imp} for plants with missing values of Y by using the estimated no-intercept regression model and the value of X_l.

We investigate the extent to which the Census imputation leads to a reduction in the variance of the imputed variables. For each industry j and for any input X we separate the plant-year observations into those X_i^I that are imputed and those that are not X_i^N. To control for size differences we divide each input by the plant's total value of shipments Y_i. We then compare the distributions of (I)mputed $\frac{X_i^I}{Y_i}$ to (N)on-imputed $\frac{X_i^N}{Y_i}$ by calculating the interquartile range of each distribution and then taking the ratio

$$R_j^X = \frac{IQR(\frac{X_i^I}{Y_i})}{IQR(\frac{X_i^N}{Y_i})}. \tag{2}$$

A ratio of R_j^X well below one suggests the imputed data has significantly less variation than the non-imputed data for variable X in industry j.

Table 2 summarizes the distribution of R_j^X across the $j = 1, \ldots, J$ industries. In 2002 the median value of R_j^X for hours worked is 0.29, for electricity is 0.11, for cost of

methods.

fuels is 0.17 and for cost of materials is 0.20. Moving up to the 75th percentile of the distribution of R_j^X the ratio for hours worked is 0.52, for electricity is 0.21, for cost of fuels is 0.35 and for cost of materials is 0.45. The results are similar for 2007 and they suggest the mean-imputation approach leads to a significant reduction in measured dispersion relative to the non-imputed data.

3 Multiple Imputation using Classification And Regression Trees

In this section we discuss how to use classification and regression trees (CART) to allow for multiple possible impute values Y_l^{imp} for any "type". We follow Burgette and Reiter (2010) and use the CART algorithm to classify plants into different types using *multiple* explanatory variables.[7] They describe the specifics of the classification method as:

> (CART) partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure with leaves corresponding to the subsets of units. The values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define the leaf.

Once the tree is constructed plants of the same "type" live on the same "leaf" of the tree and sampling with replacement from that leaf is used to fill in missing values.

[7]See also Breiman, Friedman, Olshen, and Stone (1984), Hastie, Tibshirani, and Friedman (2009), and Ripley (2009)). The "mice" software package in R includes routines for CART imputation. 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)).

When all missing values have been filled the data set is said to be "CART-completed". Repeating this process M times yields M CART-completed data sets. For any statistic of interest (like R_j^X) calculating its value across the M CART-completed data sets serves as a measure of the uncertainty introduced by missing values.

Figure 1 illustrates the use of CART in constructing an imputation model for total value of shipments (Y) conditional on the single covariate total employment (TE). The algorithm begins by searching for the level of total employment such that splitting plants into those below and above it minimizes the total variance of TVS across the two split branches. Figure 1 shows this split occurs at TE equal to 250. The process continues recursively on each branch of the tree until either the branch contains some minimum number of plant-year observations or the variance in the branch meets some minimum variance criterion for homogeneity. The branch with TE less than 250 satisfies one of these criteria but the other branch does not. CART splits the other branch one more time at total employment equal to 500. These last two branches now also satisfy the stopping criteria and the classification tree is done. Each branch is now synonymous with a leaf. The multivariate CART is similar in that at each stage of the tree-building process the algorithm searches for splits over multiple observed predictor variables within a given branch.

3.1 CART Implementation

In this subsection we describe the details of implementing CART and in the next we discuss posterior predictive checks that check for model misspecification. Readers not interested in these details can skip directly to the results in Section 4.

We start by setting to missing all Census values that were imputed using either industry average ratios or univariate regression. We collect all variables in the data matrix $Y = (Y_P, Y_C)$, where Y_P are the p_1 are the columns of variables that are not fully

observed and Y_C includes the variables that are completely observed. Y_P is arranged from left to right in order of greatest to least number of missing values. Let the conditional distributions $p(Y_l|Y_{-l})$ denote the CART-based prediction model for Y_l, the lth column of Y_P, conditional on Y_{-l}, columns of Y with Y_l removed.

The first step to CART-completing the data provides the initial guess at a completed Y. Let the matrix $Z = Y_C$ and start with Y_1, the first column of Y. We use CART to fit the tree of Y_1 on all other variables Z using observations for which Y_1 and Z are observed. We fit the tree by finding the successive "splits" in the covariates Z that minimize the variance of Y_1 in the leaves. We cease splitting any particular leaf when the variance in that leaf is less than $10e$-5 times the variance in the marginal distribution of Y_1 or when we cannot ensure at least 5 manufacturing-plant year observations are in the leaf. We impute all missing values for Y_1 and append Y_1 to Z sampling from the CART tree using the Bayesian Bootstrap (BB) of Rubin (1981), who shows "the Bootstrap and the BB ...operationally they are very similar." We repeat this process for Y_2 through Y_{p_1} appending each column after all missing values have been imputed. After this initial step Y no longer has any missing values.

The second step iterates over the columns of Y many times. For $l = 1, \ldots, p_1$ impute missing values for original missing values in Y_l conditional on Y_{-l}. This process yields another new Y matrix. We repeat this process ten times. The resulting Y is one CART-completed data set. We repeat both steps one and two M times to yield M CART-completed data sets on which we perform our analysis.

3.2 Posterior Predictive Checks

After the M CART-completed data sets have been created we can carry out posterior predictive checks to check for model misspecification. We do so by seeing whether results from the CART-completed data are similar to results from data sets where *all*

observed and missing values of Y_P are imputed. We call these data sets *predicted*.[8]

We generate M completed data sets and then we generate M fully predicted data sets by setting all values of variables included in Y_P to missing and filling in all the values using the estimated CART conditional prediction model and Y_C. We compare the statistic θ across pairs of CART-completed and CART-predicted data sets by computing a two-sided posterior predictive P-value:

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

where $I(x)$ equals one if $x > 0$ and equals zero otherwise, $\widehat{\theta}_{imp,i}$ is the estimate of parameter θ from the ith completed dataset, and $\widehat{\theta}_{pred,i}$ is the estimate from the ith predicted dataset. A P-value close to zero indicates that the $\widehat{\theta}_{pred,i}$ consistently differs from $\widehat{\theta}_{imp,i}$ in one direction suggesting possible model misspecification. A P-value close to one suggests the differences in the statistic are not systematically too high or too low across the pairs of data sets.

4 Productivity Dispersion Across Imputation Methods

We allow production function parameters to vary by 6-digit NAICS industry code and we use industry cost shares to estimate these parameters.[9] We define TFPR as

$$TFPR_i = ln(\frac{R_i}{P_j}) - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \tag{4}$$

[8]See e.g. He, Zaslavsky, Harrington, Catalano, and Landrum (2010)).

[9]While cost shares do not address the simultaneity issue raised in Marschak and Andrews (1944), we use them because that is what most research with Census data has used. In an earlier version of the paper we showed that our findings do not change if we address the simultaneity issue using the control function approaches of Olley and Pakes (1996), Levinsohn and Petrin (2003), or Wooldridge (2009), although the production function estimates do appear more sensitive across imputation methods relative to cost shares.

where R_i is the nominal total value of shipments adjusted for changes in inventories, P_j is industry j's output price deflator, 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.[10] Similarly we define TFPQ as

$$TFPQ_i = lnQ_i - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \tag{5}$$

where Q_i is the quantity of physical output. We calculate unit prices for the plants with measured output by dividing their value of total shipments by their physical quantity of product shipped.

For the CART imputation model we want good predictors of the variable to be imputed especially if the predictors have low imputation rates. For this reason when we carry out the CART-completion for TFPR we include all of the variables used in estimation as predictors. We also add changes in inventories, the plant-year's ratio of cost of energy over the total cost of materials and energy, salaries and wages, employment, and the plant-year's ratio of production worker wages to (total) salaries and wages. When carrying out CART-completion for the industries with quantity data we also include as predictors the physical quantity of shipments, the ratio of product-level value of shipments (for the main product) to plant-level total value of shipments, and an indicator for plant exit before the next Census.[11] In both the TFPR dispersion exercise and the TFPQ exercises we allow for a different imputation model for each variable and each industry. We start with the TFPR results.

[10] We deflate the 2007 dollars to 2002 dollars.

[11] In the concrete industry we include a measure of demand density as it was important in predicting productivity in Syverson (2004).

4.1 TFPR Dispersion

For each of 473 industries in 2002 and the 471 industries in 2007 we compare within-industry TFPR dispersion across the Census-completed data, the Census non-imputed data, and the CART-completed data using the 75-25 TFPR ratio. We replace industry average ratio and univariate regression imputations with CART imputations to create 100 CART-completed data sets. For the CART-completed dataset we take the average 75-25 TFPR ratio across the M=100 data sets for each industry-year. For each industry-year we calculate (i) the log of the 75-25 TFPR ratio in the non-imputed minus the log of the 75-25 TFPR ratio in the Census-completed data, and (ii) the log of the 75-25 TFPR ratio in the CART-completed data minus the log of the 75-25 ratio in the Census-completed data.

Table 3 presents the results for each year. Dispersion increases as we move from Census-completed data to Census non-imputed data to CART-completed data by any measure. For example, for the average industry, the dispersion measure is 11.3 and 7.3 log points higher in non-imputed data in 2002 and 2007 respectively. Moving from Census-completed to CART increases TFPR dispersion by 16.2 log points in 2002 and 12.3 log points in 2007. The increase in dispersion is apparent throughout the manufacturing sector. In 2002 and 2007, respectively 47% and 33% of industries have a 75-25 TFPR ratio that is at least 10 log points higher in the non-imputed data than in the Census-completed data. For the CART-completed vs. Census-completed comparison, the analogous percentages are 66% of industries in 2002 and 51% in 2007. Our results suggest mean imputation in the Census data leads to a compression of the TFPR distribution meaning there is *more* TFPR dispersion than has been currently thought (see Syverson (2011)).

4.2 TFPR, TFPQ, and Price Dispersion

We focus on a subset of the manufacturing industries studied in Foster, Haltiwanger, and Syverson (2008) for which we have at least 100 observations in an industry-year: ready-mix concrete, boxes, and ice. Table 4 presents within-industry TFPR for concrete and TFPR, TFPQ, and price dispersion statistics for boxes and ice.[12] Columns 1, 3, and 5 report the statistics calculated from the Census mean-imputed data and columns 2, 4, and 6 use the CART-completed data. We compute each statistic separately from each of our 500 CART-completed datasets and report the mean across the 500 estimates.

The 75-25 ratios for TFPR, TFPQ, and unit prices across the columns show uniformly more dispersion as one moves from Census mean-imputed to non-imputed to CART-completed with both TFPQ and unit price dispersion exceeding TFPR dispersion. For specific industry cases like prices for ice in 2007 (2.37 vs. 1.11) the differences can be very large. FHS reported that dispersion in TFPQ exceeds TFPR, and this result is further magnified when CART-completion is used to impute missing values.

Table 4 reports two-sided posterior predictive P-values for each measure of productivity and price dispersion in table 4 to check for model misspecification. For each dispersion measure, we also calculate the mean of the differences between the CART-predicted estimate and the CART-completed estimate for the 500 pairs of datasets. These means are presented in columns 1, 3, and 5 of table 5. Means corresponding to a P value less than 0.05 – cases where there is possible evidence of model misspecification – are indicated by an asterisk. To put these differences in perspective, in columns 2, 4, and 6 we show the ratio of the mean difference over the CART-completed mean for each measure, and most of them are small. For example, the TFPR dispersion for the concrete industry in 2002 has P-value of 0 (the CART-predicted estimate is al-

[12]The Census Bureau last collected physical quantity data for concrete in 1992.

ways higher than the CART-completed estimate) but on average the CART-predicted estimate is only 0.13 (about 7%) higher than the CART-completed estimate.

4.3 Correlates of Plant Survival

Foster, Haltiwanger, and Syverson (2008) (FHS) relate plant exit to TFPR, TFPQ, prices, and idiosyncratic demand shocks. In their Table 6 they report negative and significant relationships between plant exit and all four of these measures. This finding is broadly consistent with the predictions from many models of dynamic competition between plants. While FHS were able to identify and drop much of the imputed data, the subsequent release of the item-level impute flags showed some of their remaining data was imputed.[13] In this section we test whether their results are robust to CART-completion.

We replace the imputed data in FHS's estimation sample with multiply-imputed data from CART keeping exactly the same sample of plants as in FHS. We then rerun their probit exit regressions. Table 6 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 estimated marginal effects. For each probit, the standard errors are clustered by plants.[14] FHS's results are very robust to imputation as traditional TFP, TFPR, TFPQ, prices, and demand shocks all continue to be significantly and negatively associated with exit on CART-completed data.

[13]They used reverse-engineering methods and were able to identify some of the imputed data and remove those plants from their sample (see Foster, Haltiwanger, and Syverson (2008) and the robustness analysis associated with it). We thank Lucia Foster, John Haltiwanger and Chad Syverson for sharing their computer codes and (with approval from the Census Bureau) access to their datasets.

[14]We also combine the 500 sets of standard errors using Rubin's (1987) combining formula.

5 Conclusion

Much of the literature on plant-level productivity uses the U.S. Census Bureau's Census of Manufactures. We show that the recent availability of imputation flags for the 2002 and 2007 U.S. Census data implies that over 70% of observations in both years have imputed data for at least one variable used to compute total factor productivity.

The Bureau imputes for missing values using mean-imputation methods which are known to reduce the true underlying variance of the imputed variables. For every variable entering TFP in 2002 and 2007 we show the dispersion is significantly smaller in the Census mean-imputed versus the Census non-imputed data. *Ex ante* it is not obvious how the significant reduction in dispersion we observe from mean-imputation affects total factor productivity (TFP) because TFP is a ratio of output over an input index.

Using classification and regression trees (CART), for 473 industries in 2002 and 471 industries in 2007 we provide a new set of multiple imputations that seek to better preserve dispersion and the joint distribution of key variables. We find TFP dispersion increases as we move from Census mean-imputed data to Census non-imputed data to CART-imputed data, suggesting TFP has *more* dispersion than previously believed and making the amount of within-industry TFP dispersion in plant-level data even more puzzling. For the small collection of industries where we observe quantities we find even starker increases in dispersion for TFPQ and unit prices as we move across the three data sets. In contrast, when we revisit FHS, who report negative and significant relationships between plant exit and TFPR, TFPQ, prices, and idiosyncratic demand shocks, we show FHS's results are very robust to CART-imputation

References

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.

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.

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. 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.

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.

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.

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

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

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

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

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.

White, T. K., J. P. Reiter, and A. Petrin (2015): "Imputation in U.S. Manufacturing Data and Implications for Within-Industry Productivity Dispersion," Federal Committee on Statistical Methodology Research Conference Proceedings.

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

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

Statistic	Total Value of Shipments	Production Worker Hours	Cost of Purchased Electricity	Cost of Fuels	Cost of Materials
			2002		
Mean	27%	19%	38%	37%	42%
s.d.	9%	7%	14%	14%	10%
			2007		
Mean	27%	31%	37%	35%	42%
s.d.	9%	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 Across Industries of Ratios of Within-Industry Interquartile Ranges: Imputed vs. Non-Imputed Data

percentile	Production Worker Hours	Cost of Purchased Electricity	Cost of Fuels	Cost of Materials
			2002	
25th	0.159	0.062	0.088	0.036
50th	0.293	0.112	0.174	0.208
75th	0.522	0.219	0.356	0.456
			2007	
25th	0.353	0.088	0.152	0.089
50th	0.486	0.179	0.370	0.262
75th	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 A value well-below one signifies there is much more variance in the variable in the non-imputed data vs. the imputed data.*

Table 3: Changes in Within-industry Productivity Dispersion Across Imputation Methods

	median	mean	25th percentile	75th percentile
			2002	
Non-imputed vs. Bureau-completed	0.096	0.113	0.045	0.159
CART-completed vs. Bureau-completed	0.130	0.162	0.082	0.212
			2007	
Non-imputed vs. Bureau-completed	0.059	0.076	0.020	0.121
CART-completed vs. Bureau-completed	0.103	0.123	0.055	0.160

The table shows how different imputation methods change measures of within-industry dispersion in revenue-based TFP in the 2002 and 2007 mail samples of the Censuses of Manufactures (CMF). The first and third rows show moments of the distribution of $log(TFPR7525_{j,NI}/log(TFPR7525_{j,CB})$*, where* $TFPR7525_{j,CB}$ *is the ratio of revenue-based TFP (TFPR) at the 75th percentile in industry j over TFPR at the 25th percentile in the same industry in the Census Bureau-completed data, in which missing or faulty data was imputed by the Census Bureau using a variety of methods;* $TFPR7525_{j,NI}$ *is the 75-25 TFPR ratio for industry j in 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. Rows 2 and 4 show moments of the distribution of* $log(TFPR7525_{j,CART}/TFPR7525_{j,CB})$*, where* $TFPR7525_{j,CART}$ *is the mean of the 75-25 TFPR ratios for industry j from 100 implicates of CART-completed data, in which variables in the Bureau-completed that were imputed by industry average ratio or univariate regression are replaced by CART imputations.*

Table 4: TFPR, TFPQ, and Unit Price Dispersion: Census vs. Cart Imputation

		75-25 TFPR Ratios		*75-25 TFPQ Ratios*		*75-25 Price Ratios*	
		(1)	(2)	(3)	(4)	(5)	(6)
	Sample	Census		Census		Census	
industry	Size	Bureau	CART	Bureau	CART	Bureau	CART
				2002			
concrete	3294	1.33	1.79	n/a	n/a	n/a	n/a
boxes	626	1.17	1.18	1.90	2.13	1.86	2.04
ice	169	1.48	1.61	1.67	2.11	1.15	1.73
				2007			
concrete	4961	1.30	1.72	n/a	n/a	n/a	n/a
ice	237	1.68	1.78	1.93	2.75	1.11	2.37

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 5: Posterior Predictive Checks of the CART Imputation Models for TFPR, TFPQ, and Unit Price Dispersion

		75-25 TFPR Ratios		*75-25 TFPQ Ratios*		*75-25 Price Ratios*	
		(1)	(2)	(3)	(4)	(5)	(6)
		Mean Difference	Mean Diff/ CART mean	Mean Difference	Mean Diff/ CART mean	Mean Difference	Mean Diff/ CART mean
				2002			
concrete	3294	0.13*	0.07	n/a	n/a	n/a	n/a
boxes	626	0.05*	0.04	0.16	0.08	0.10	0.05
ice	169	0.27*	0.17	0.63	0.30	0.65	0.38
				2007			
concrete	4961	0.14*	0.08	n/a	n/a	n/a	n/a
ice	237	0.10	0.06	0.2	0.07	0.17	0.07

*Columns 1, 3, and 5 show the means of the differences between 500 pairs of CART-predicted estimates and CART-completed estimates for the dispersion measures in table 4. Columns 2, 4, and 6 show the ratio of the mean difference over the CART-completed mean for each industry-year. * indicates a P probability less than 0.05 (see equation 5 in the text) for the associated statistics in table 4. 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 6: 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.

Table A1: Selection on Productivity or Profitibility, 1977-1992 Industries

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
					(0.003)		(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.

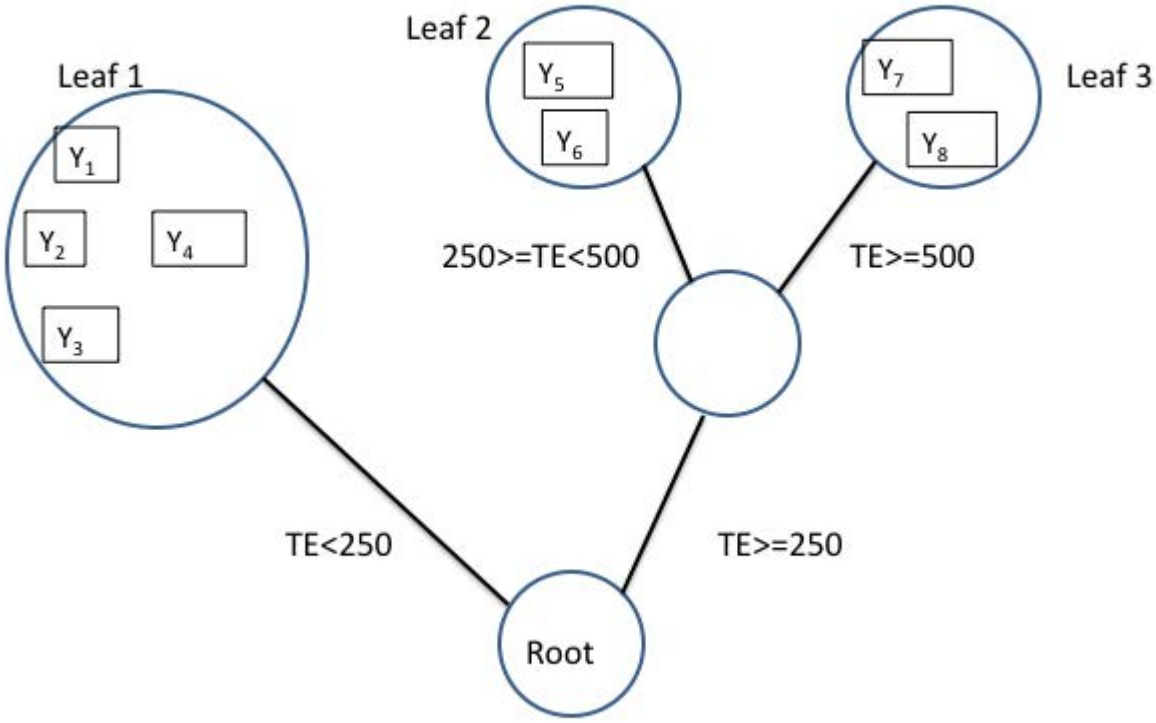

Figure 1: A Simple Classification and Regression Tree