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

by
T. Kirk White
Center for Economic Studies, U.S. Census Bureau
Jerome P. Reiter
Duke University
Amil Petrin
University of Minnesota,Twin Cities and NBER

January 3, 2013

*Most of the research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center and the Minnesota Census Research Data Center. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. We thank Randy Becker, Bert Grider, Cheryl Grim, John Haltiwanger, Shawn Klimek, Arnie Reznek, Pat Sullivan and participants in the Census Bureau's Center for Economic Studies seminar for their comments. We thank Lucia Foster for giving us some of her computer code, and Jenny Thompson and Kirk Degler for helpful discussions of the Census Bureau's imputation procedures. Reiter gratefully acknowledges support from National Science Foundation grant SES 1131897. Correspondence to: T. Kirk White, Center for Economic Studies, U.S. Census Bureau, 4600 Silver Hill Rd., Suitland, MD

Abstract

Within-industry differences in measured plant-level productivity are large. A large literature has been devoted to explaining the causes and consequences of these differences. In the U.S. Census Bureau's manufacturing data, the Bureau imputes for missing values using methods known to result in underestimation of variability and potential bias in multivariate inferences. We present an alternative strategy for handling the missing data based on multiple imputation via sequences of classification and regression trees. We use our imputations and the Bureau's imputations to estimate within-industry productivity dispersions. The results suggest that there is more within-industry productivity dispersion than previous research has indicated. We also estimate relationships between productivity and market structure and between output prices, capital, and the probability of plant exit (controlling for productivity) based on the improved imputations. For some estimands, we find substantially different results than those based on the Census Bureau's imputations.

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

20746. email: thomas.kirk.white@census.gov. Phone: (301) 763-1879. Fax: (301) 763-5935.

1 Introduction

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 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)).

In this paper, we explore an explanation for (measured) productivity dispersion which has largely been ignored in the economics literature: imputed data. 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

[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.

imputation using the U.S. Census Bureau's Census of Manufactures (CM) which supports much of the empirical research on plant-level productivity. Although the CM represents the best available data for studying U.S. plant-level total factor productivity, imputations for nonresponse comprise a large percentage of the data; in fact, we show that this percentage is far higher than what is reported in the existing literature. For example, in 2007 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%.[2] 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.

The Census Bureau imputes for missing data using a variety of methods, including ratio imputation and conditional mean imputation. The Bureau's primary goal is to facilitate point estimation of industry aggregates; however, several of the Bureau's imputation methods are not appropriate for multivariate regression analysis of microdata — such as estimating plant-level total factor productivity — because these methods can distort covariances and correlations between variables and lead to underestimates of standard errors

[2]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.

(Schafer and Graham (2002) and Little and Rubin (2002)). We find that functions of key variables in the completed 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 many industries that have been studied previously.

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 imputations 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.

What can be done about this imputed data? One solution, popular among economists, is to drop observations with imputed values, and only analyze the plants with complete (non-imputed) data.[3] Unfortunately, it is well-known that in general, complete case analysis — i.e, using only the plants with no missing or imputed data — sacrifices efficiency and can lead to biased parameter estimates.[4] Even if complete case analysis does not create

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

[4]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.

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. Hence, complete-cases is not a trustworthy solution.

As an alternative to these strategies, we create completed datasets via multiple imputation (Rubin (1987), Reiter and Raghunathan (2007)).[5] 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 conditional mean 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 most of the Census Bureau's imputation methods, 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.

Using the Census Bureau-imputed data, we estimate within-industry pro-

[5]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 conditional mean imputation.

ductivity dispersion for several relatively homogeneous industries that have been studied previously. Then we replace the Census Bureau's imputations with multiple imputations using the CART method, and we re-estimate within-industry productivity dispersion for the same industries. Our results suggest that there may actually be *more* within-industry productivity dispersion than the existing literature suggests.

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 there is more within-industry productivity dispersion in U.S. manufacturing than previous research suggests, productivity losses from misallocation in other countries may not be as large as Hsieh and Klenow (2009) suggests. On the other hand, the extent of distortions and misallocation in U.S. manufacturing may be greater than the existing literature suggests.

We also investigate how imputed data affects some key empirical relationships between productivity and other economic variables. First, following Foster, Haltiwanger, and Syverson (2008), we estimate the probability of plant exit conditional on plant-level prices, capital stocks, and plant-level productivity. When we use the Census Bureau's imputations, we reproduce FHS's result that (conditional on plant-level productivity and capital stocks) plant-level prices are associated with significant decreases in exit probabilities. However, when we estimate the same regression using only plants with non-imputed price and quantity data, the price coefficient is attenuated

and no longer statistically significant. The Census Bureau's imputations appear to create an artificially strong relationship between exit and plant-level prices, conditional on productivity and capital stocks. When we replace the Bureau's imputations with multiply-imputed CART data, the coefficient on prices is also attenuated and statistically insignificant.

In a second example, following Syverson (2004a), we regress local market-level productivity and size moments on a measure of demand density for the ready-mix concrete industry. Using the Census Bureau's imputations, we largely confirm Syverson's results. However, when we replace the Census Bureau's imputations with multiply-imputed data using CART, the demand density coefficients are significantly affected, and in some cases the sign is reversed.

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 (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 those used by the Census Bureau, the differential results 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 describes our sample selection strategy and how we choose variables for the imputation model. Section 6 shows how the imputations affect estimates of productivity dispersion and some key empirical relationships between productivity and other variables. Section 7 presents validity checks of our imputations. Section 8 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:

$$TFPQ_i = lnQ_i - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \quad (1)$$

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:

$$TTFP_i = ln(P_iQ_i) - \beta_k lnK_i - \beta_l lnL_i - \beta_e lnE_i - \beta_m lnM_i \quad (2)$$

where P_iQ_i is the total value of the plant's output.[6] 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

[6]FHS call this measure "Traditional TFP."

parameters in equations 1 and 2 is the approach used in some of the most influential research in the productivity literature.[7] Thus this choice facilitates comparisons with the existing 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.

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 — is 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 to identify which, if any, items for a given plant were imputed due to item nonresponse, because item-level flags were not made available. Previous researchers developed several clever ways

[7]For example, Bailey, Hulten, and Campbell (1992), Syverson (2004b), Syverson (2004a), and Foster, Haltiwanger, and Syverson (2008), all use this approach. Hsieh and Klenow (2009) also use cost shares, although they estimate two-factor value-added production functions.

to identify some of the imputed values.[8] However, the item-level flags that became available in the 2002 Census show that a much higher percentage of observations are imputed than are identified by 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 — 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.

The Census Bureau uses a variety of methods to impute for missing data, including conditional mean imputation and industry average ratio imputation. The industry average ratio method imputes for missing values of vari-

[8]See for example, Dunne (1998), Foster, Haltiwanger, and Syverson (2008), Roberts and Supina (1996), and Roberts and Supina (2000).

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

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

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 $\overline{\left(\frac{Y}{X}\right)}$, 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 conditional mean 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 Z is a strong predictor for Y even after conditioning on X. 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.[9] The same logic applies to conditional mean 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

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

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. 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 Census Bureau's imputation methods, we replace the Bureau's imputations with multiple imputations created via sequential regression trees, as developed by Burgette and Reiter (2010).[10] 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 conditional mean 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 is also likely to

[10]The code for implementing the sequential CART method is available on the internet at http://www.burgette.org/software.html.

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.

We now describe the 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 $Y_{L_w}^{(1)}$ be the n_{L_w} values of Y_1 in leaf L_w. For all records whose terminal leaf is L_w, we generate replacement values of Y_{ij} by drawing from $Y_{L_w}^{(1)}$ using the Bayesian bootstrap (Rubin (1981)). Repeating the Bayesian bootstrap for each leaf of $\mathcal{Y}^{(1)}$ results in an initial set of plausible values.

We next move to the second variable in the ordering with missing data, say Y_2. We fit the tree of Y_2 on all other variables, which we call $\mathcal{Y}^{(2)}$, using the newly completed values of Y_1. We run observations down $\mathcal{Y}^{(2)}$ to create plausible values for Y_2. The process continues for each Y_i in the ordering, each time using the newly imputed values of Y_{-i} to fit the tree and in locating leaves. We then cycle through this process ten times to help move the trees away from the initial starting values. The end result is one completed dataset. We repeat this entire process 20 times to generate 20 completed datasets. By cycling through the process ten times between completed datasets, we minimize dependence between the completed datasets.[11]

5 Sample Selection and Imputation Model

In order to directly investigate the impact of imputation on estimates of plant-level productivity dispersion, we select a few detailed industries. Our industry selections are motivated by two factors. First, we want to choose

[11]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.

industries that have been studied previously, to facilitate comparision with the existing literature on plant-level productivity. Second, we want to select industries that are relatively homogeneous. 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.

To satisfy both of our criteria we select 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.[12] According to FHS, "Producers of these products make outputs that are among the most physically homogeneous in the manufacturing sector." For all of these industries except concrete, in at least one year we have data on the values and physical quantities of the products the plants produce.[13] 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 sample selection procedure in detail in the appendix.

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

[13]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.

Table 3 shows the sample size and imputation rates for key variables for each of our selected industries. Except for the concrete industry, the imputation rates in our sample are significantly lower than in the average manufacturing industry. 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, 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 3. 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. In the next section we describe the sequential CART method, which we use to create multiple imputations to replace each of the Census Bureau's imputations.

What variables should be included in the imputation model? 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. 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 predic-

tors of these variables. These considerations lead us to include a rich set of variables as inputs to the CART procedure. For each industry in table 3 except concrete, the potential predictors for each tree include — whenever the variable is not the dependent variable — the value of product shipments, the physical quantity of product shipments, the book value of assets, the cost of purchased electricity, the cost of fuels, the total cost of materials, the number of production workers, production worker hours, total salaries and wages, production worker wages, the total value of shipments, and the number of non-production workers. Since we also plan to analyze the relationship 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. For the ready-mix concrete industry, we have a larger sample size and slightly different set of available variables, so we build a somewhat different imputation model. We describe the variable selection for the concrete industry in section 6.2.

6 Results

Table 4 presents within-industry productivity and price dispersion statistics for the industry-years for which we are able to calculate them. For each measure — TTFP, TFPQ, and prices — 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 4 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 20 completed datasets, and report the means of the 20 estimates. Comparing columns 1 and 2, for every industry except bread there is more within-industry TTFP 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.

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.[14] For each industry-year statistic in table 4, 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 TTFP, 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.

[14]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 required that the plants have non-imputed product quantity and product value data. Only about 43% of the plants in our sample in 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.

If our CART imputation models are correct,[15] 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

[15]We check the validity of the CART imputation models in section 7.

within-industry productivity and price dispersion. But do imputations affect key empirical relationships between productivity and other variables? In the next subsections we show that they do.

6.1 Productivity, Prices 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 affect on industry aggregates (Jovanovic (1982), Ericson and Pakes (1995), Melitz (2003), and Bartelsman and Doms (2000)). One of FHS's contributions is to separately measure the effects of productivity and prices on plant survival. Specificallly, they find that, controlling for physical productivity (TFPQ) and plant capital stock, plants with higher output prices are less likely to exit. FHS interpret the price coefficient as a demand effect.

FHS use data from the Censuses of Manufactures for 1977-1997, and it is difficult to identify imputed data for these years. Here we attempt to replicate FHS's results using the 2002 and 2007 CM data, so that we can take advantage of the item-level impute flags to identify imputed data. To the extent possible we use FHS's sample selection strategy. The main difference is that we exclude concrete plants, since the Census Bureau stopped collecting physical quantity data for this industry in 1992.[16] Following FHS, we use

[16]We describe our industry definitions and sample selection strategy in detail in the appendix.

probit regressions to estimate the probability that a plant exits between 2002 and 2007, conditional on the plant's productivity, output price, and capital stock.[17]

Table 5 presents the results of the probit regression run on different samples. All the regressions include industry fixed effects (not reported). Column 1 shows the results from the Bureau-completed data. Like FHS, we find that a higher capital stock, higher productivity, and higher prices are associated with a statistically significantly lower probability of exit.[18] Column 2 reports the results when we run the same regression on only the plants with imputed product physical quantity data. The coefficient on physical productivity is 36% larger in magnitude, and the coefficient on prices is almost 3 times larger in magnitude compared to column 1. Column 3 shows the results when we repeat the same regression on the sample of plants with observed (non-imputed) product physical quantity data. Compared to column 2 the coefficient on physical TFP is about 40% smaller in magnitude, the price coefficient is about 80% smaller, and neither coefficient is statistically significant. The Census Bureau's imputations seem to strongly affect the estimated relationship between prices and the probability of exit, conditional on plant-level productivity and capital stock.

[17]The exit rate in our estimation sample is 18.4%, which is quite similar to the average exit rate of 19.6% in FHS's sample.

[18]FHS also estimate a probit, but report the marginal effects. For comparison, our estimated marginal effects for physical TFP, prices, and capital (evaluated at the sample mean) are, respectively, -0.099, -0.092, and -0.042. These are similar to FHS's estimates.

Column 4 of table 5 shows the results of the same probit regression run on the CART-completed dataset. We estimate each probit separately on each of the 20 CART datasets and report the means for each coefficient. We compute the standard errors using Rubin's (1987) combining formula. The coefficients on physical TFP and capital are still significant, but the price coefficient is no longer significant. The relationship between plant-level prices and plant survival, conditional on physical productivity and capital, does not seem to be as strong as the Bureau-completed data suggests.

6.2 Another Concrete Example: Productivity and Market Structure

The ready-mix concrete industry has been studied by a number of economists — Syverson (2008) reviews some of this research. Perhaps most notably, Syverson (2004a) uses data from the 1982-1992 Censuses of Manufactures to study the industry. Because ready-mix concrete is subject to high transport costs, the industry's output markets are geographically segmented. The product is also relatively homogeneous. Syverson takes advantage of these features of the industry to assess the impact of market-level demand shocks on market structure and market-level moments of the productivity distribution. Here we attempt to replicate some of Syverson's findings using the 2002 and 2007 Censuses, and we assess the impact of imputation on the results.

Following Syverson, we use the Bureau of Economic Analysis's component

economic area (CEA) as our market definition, and we select producers in markets with at least 5 concrete plants. For each plant-year observation we calculate traditional TFP as in equation 2. For each market-year, we calculate the median TFP, the interquartile range (IQR) of TFP, output-weighted mean TFP, the 10th percentile of TFP, the log of average plant output, and Syverson's measure of concrete demand density — construction sector employment per square mile.[19] Table A3 in the appendix presents descriptive statistics for our sample, computed from the Bureau-completed data.[20]

We regress each market-year moment of productivity or output on the log of market demand density. Table 6 presents estimates and heteroskedasticity-robust standard errors for the coefficient on demand density. Columns 1 and 2 show the results of two specifications run on the Bureau-completed data. With the exception of the output-weighted mean TFP regression, the magnitudes of the coefficients are roughly the same as Syverson's results and the sign pattern is the same: within-market TFP dispersion is negatively associated with demand density, and median TFP, 10th percentile TFP, and average output are positively associated with demand density. The demand

[19]We describe the variable construction in more detail in the appendix.

[20]Compared to Syverson's sample, in the average market we have about 3 more plants with TFP data, and the interquartile range of the number of plants per market is exactly the same. In our sample there is more across-market variation in within-market TFP dispersion, median TFP, and weighted mean TFP, less variation in the 10th percentile of TFP and demand density, and about the same variation in average output.

density coefficient is statistically significant at the 5% or 1% level in the latter three regressions.

The across-market output and TFP variation that we see in table A3 and the results in the first two columns of table 6 suggest that market-level demand density is a useful predictor of plant size and productivity. 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 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, so we cannot include them as potential predictors in the imputation model.

We create 20 CART-completed datasets for the concrete industry and run the regressions separately on each dataset. Columns 3 and 4 of table 6 show the means of the 20 estimates for each regression. We combine the heteroskedasticity-robust standard errors from each of the 20 regressions using the combining formulas in Rubin (1987). The imputations have a strong impact on some of the results. The demand density coefficients are now much larger in the TFP dispersion regressions and 10th percentile of TFP regressions, and the coefficients are now statistically significant in the former. In the median TFP regressions, the demand density coefficient changes sign and increases in magnitude.

Why do the imputations affect the results in this way? It turns out that in the concrete industry, demand density is correlated with the market-

level imputation rate: in a regression of log demand density on the market-level fraction of concrete TFP data that is imputed, the coefficient on the imputation rate is -0.51. Markets with lower levels of demand density have higher percentages of imputed concrete data in the CM. Since the Census Bureau's imputation methods tend to affect the overall distribution of TFP we see in the observed data, these imputations also affect the estimated relationships between market-level TFP and demand density.

7 Validity Checks

Intuition and the statistical literature on imputation suggest 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. 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 regression coefficients in tables 5 or 6. To check the validity of our imputation models for the analyses above, we use posterior predictive checks (He, Zaslavsky, Harrington, Catalano, and Landrum (2010)). Intuitively, we use the CART method to create many pairs of datasets, where the first dataset in each pair includes imputed and non-imputed data, while the second dataset is (almost) entirely imputed. Then we re-estimate the regressions and productivity dispersion statistics on each dataset. If, for example, the regression coefficient of interest is consistently

higher in the almost entirely imputed datasets, then this is evidence that the imputation model may be leading to upward-biased estimates of that coefficient.

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 regression coefficient — 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:

$$P_j = \frac{2}{500} min\{\sum_{i=1}^{500} I(\widehat{\theta}_{imp,ij} - \widehat{\theta}_{pred,ij}), \sum_{i=1}^{500} I(\widehat{\theta}_{pred,ij} - \widehat{\theta}_{imp,ij})\} \quad (5)$$

where $I(x)$ equals one if $x > 0$ and equals zero otherwise. Here, $\widehat{\theta}_{imp,ij}$ is the estimate of parameter θ_j — a regression coefficient or 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 price dispersion in table 4 and for the regressions presented in tables 5 and 6For 24 of the 33 dispersion moments estimated in table 4, the associated P probabilities are greater than 0.05, and 20 of them are greater than 0.10.[21] In the few cases where there is evidence of a bias, the biases tend to be small. For example, the traditional 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

[21]We provide the full set of P values in table A4 in the appendix.

dispersion statistics for the boxes industry.

For the exit probits in table 5, the P probabilities associated with the coefficients on physical TFP, prices, and capital are, respectively, 0.24, 0.24, and 0.06. Thus we find no evidence that the CART imputations are leading to biased estimates of the coefficients that are most affected by the imputation method — the coefficients on physical TFP and prices. Together this evidence suggests that the CART model generates plausible data with respect to most of the estimated relationships represented in tables 4 and 5.

In table 6, for the regressions with productivity dispersion as the dependent variable, the validity checks provide no evidence that the CART imputations lead to a bias in the demand density coefficients.[22] In the other regressions in table 6, the validity checks suggest there is a positive bias in the demand density coefficients. However, in the median TFP and weighted-mean TFP regressions, this bias works *against* finding that the imputation method matters. In the median TFP regressions, we find negative, statistically significant coefficients in the CART data despite evidence of a positive bias.

[22]We report the P values for the regressions in 6 in table A5 in the appendix.

8 Conclusions and Suggestions for Further Research

Much of the literature on plant-level productivity uses the 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 as well as relationships between productivity and other important economic variables.

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. We also find that estimated relationships between productivity and demand density, and between prices, capital, and plant survival are not robust to replacing Bureau-imputed data with CART-imputed data.

In addition to demand density, prices, and capital stocks, 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. These variables are not 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. This suggests that many existing estimates of the relationships between U.S. manufacturing productivity and other key variables may be biased.

Our findings also have broader implications. Although imputation for missing data is a well-developed field in statistics, economists have ignored many of its findings. Complete-cases analysis is known to lead to biased parameter estimates except under restrictive assumptions (Little and Rubin (2002)), but economists still routinely drop imputed data from their samples without correcting for sample selection bias. The goals of statistical agencies that collect microdata and publish aggregate statistics are typically different from the goals of economists and other researchers who use the microdata the agencies collect. 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 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., 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.

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. (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.

——— (2008): "Markets: Ready-mix Concrete," *Journal of Economic Perspectives*, 22(1), 217–233.

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

Van Biesebroeck, J. (2004): "Robustness of Productivity Estimates," NBER Working Paper.

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.

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

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

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: Imputation Rates for Key Variables, Selected Industries

	FHS industries, except concrete	concrete, 2002	concrete, 2007
Sample Size	1453	4845	5512
Value of Shipments	12%	24%	20%
Quantity of Product	45%	NA	NA
Book Value of Assets	13%	25%	32%
Production Worker Hours	8%	39%	27%
Cost of Electricity	11%	35%	36%
Cost of Fuels	10%	33%	34%
Cost of Materials	19%	33%	36%

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 4: Productivity and Price Dispersion, Selected Industries

			75-25 TTFP 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	2.03
bread	2002	71	1.79	1.61	2.10	2.11	1.09	1.53
carbon black	2002	21	1.45	1.56	1.45	1.83	1.09	1.82
coffee	2002	98	1.15	1.67	1.32	2.38	1.07	1.71
flooring	2002	40	1.35	1.40	1.81	2.09	1.26	1.95
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	1.08
ice	2002	169	1.48	1.61	1.67	2.10	1.15	1.72
ice	2007	237	1.68	1.78	1.93	2.71	1.11	2.34
plywood	2002	36	1.26	1.29	1.89	3.96	1.50	3.11
sugar	2002	21	1.31	1.64	1.40	1.61	1.02	1.10

The table shows ratios of the 75th percentile to the 25th percentile of within-industry-year distributions of total factor productivity (TFP) and prices. TTFP is a traditional 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 20 CART-completed datasets.

Table 5: Plant-level Productivity, Prices, and Exit

Specification	(1)	(2)	(3)	(4)
Physical TFP	-0.386***	-0.526***	-0.314	-0.326**
	(0.136)	(0.192)	(0.196)	(0.162)
Prices	-0.358**	-0.988***	-0.178	-0.281
	(0.171)	(0.307)	(0.223)	(0.176)
Capital Stock	-0.162***	-0.148***	-0.165***	-0.137***
	(0.030)	(0.042)	(0.047)	(0.029)
Sample Size	1155	434	721	1155

*The table shows coefficient estimates (and heteroskedasticity-robust standard errors) for exit probit regressions estimated on four different samples. All regressions include industry fixed effects. The samples consists of plants in the industries in table 4. Columns 1-3 report results from, respectively, the Census Bureau-completed data, the sample of plants with imputed product quantity data, and plants with non-imputed product quantity data. The final column shows the means of coefficient estimates from 20 CART-completed datasets and robust standard errors combined using Rubin's (1987) combining formulas. *,**,and *** indicate significance at the 10, 5, and 1 percent levels, respectively.*

Table 6: Concrete Productivity, Size, and Demand Density

	Bureau-completed data		CART-completed data	
	(1)	(2)	(3)	(4)
TFP interquartile range	-0.019	-0.019	-0.147***	-0.146***
	(0.012)	(0.012)	(0.048)	(0.048)
Median TFP	0.015**	0.013**	-0.054**	-0.054**
	(0.007)	(0.006)	(0.022)	(0.022)
Output-weighted mean TFP	0.001	0.000	0.006	0.006
	(0.014)	(0.014)	(0.027)	(0.027)
10th percentile TFP	0.025***	0.023***	0.141*	0.140*
	(0.007)	(0.006)	(0.079)	(0.078)
Mean output	0.202***	0.202***	0.208***	0.207***
	(0.023)	(0.023)	(0.023)	(0.022)
Year dummies?	no	yes	no	yes

The table shows estimated coefficients on demand density when market-level productivity and output statistics are regressed on demand density. The sample consists of 444 market-year observations. Heteroskedasticity-robust standard errors are in parentheses. The first two columns show estimates from Census Bureau-completed data. Columns 3 and 4 show the means of coefficient estimates from 20 CART-completed datasets and robust standard errors combined using Rubin's (1987) combining formulas.

Appendices

In this online appendix we provide detailed descriptions of the sample selection strategy, variable construction, and method for identifying imputed data for the analyses in the main text. We also provide descriptive statistics (table A3) for the concrete sample, and P-values (tables A4 and A5) for the validity checks described in the main text.

A Sample Selection

This section provides detailed descriptions of our sample selection strategy. For the industries analyzed in tables 4-6, to the extent possible we follow the sample selection strategy of Foster, Haltiwanger, and Syverson (2008). We select 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. One way in which our sample selection strategy differs 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.

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.

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.

B Variable Construction

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

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.

For the "traditional TFP" measure, our output measure is the plant's total value of shipments deflated by the shipments deflator for the corresponding industry from the NBER Productivity Database.

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 corre-

sponding 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 table 6 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. In the regressions presented table 6 we use the natural log of this measure.

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. However, in 2002 we find that for many plants the flags indicate that *none* of these capital variables was reported on the survey form and all of them were "raked." Since it is impossible to adjust to a total that was not reported, we treat these items as imputed.

The Census Bureau uses different imputation methods for different variables. For example, the "industry average" ratio method is used frequently for the energy input cost variables (cost of fuels, cost of purchased electricity), and almost all of the imputations for the product-level product quantity shipped (PQS) data are created with this method. On the other hand, for total value of shipments and the total cost of materials, the "Beta (Cold Deck Statistical)" method is the most common method. Note that although the edit/impute flags tell us what general method was used to impute each data element, for most variables we still do not know exactly how each element was imputed. For example, if the edit/impute flag for a plant's cost of purchased electricity is set to ' V', we know that the plant's electricity costs are set to the industry average by ratio imputation, but we do not know what the denominator of the ratio is. Similarly, a flag set to ' B' ("Cold Deck Statistical") means that the item was imputed using a regression model based on historical data. However, in general we do not know what sample was used for this regression or even what explanatory variables are in the regression model. One exception to this rule are the imputations for the product

quantity variable (PQS). For this variable, almost all of the imputations are constructed by multiplying the plant's product value shipped (PV) by an industry average ratio of product quantity over product value.

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: Concrete Markets: Productivity, Size, and Demand Density

	Mean	Std. Dev.	75th-25th Percentile Range	90th-10th Percentile Range
TFP interquartile range	0.29	0.27	0.19	0.42
Median TFP	1.49	0.16	0.22	0.37
Output-weighted mean TFP	1.60	0.27	0.25	0.52
10th percentile TFP	1.24	0.18	0.19	0.38
Number of plants with TFP	17.4	19.6	10.0	36.0
ln(mean plant output)	8.14	0.55	0.74	1.42
Construction employment/square mile	3.21	3.95	2.77	7.04

The table shows descriptive statistics for 444 market-year observations for the ready-mix concrete industry in 2002 and 2007. A market is defined as the Bureau of Economic Analysis's Component Economic Area (CEA).

Table A4: Validity Checks of the CART Imputation Models:

Within-industry Productivity and Price Dispersion

		P value for:		
		75-25 TTFP	75-25 TFPQ	75-25 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
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 traditional 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 A5: Validity Checks of the CART Imputation Model

For Concrete: Regressions on Demand Density

Dependent variable	(1)	(2)
TFP interquartile range	0.164	0.164
Median TFP	0.000	0.000
Output-weighted mean TFP	0.000	0.000
10th percentile TFP	0.024	0.024
Mean output	0.000	0.000
Year dummies?	no	yes

The table shows P probabilities (see equation 5 in main text) for checks of the validity of the CART imputations model for a regression of market-level moments of concrete productivity or output on concrete market demand density. A probability close to zero is evidence that the CART imputations are distorting the joint distributions in a way that leads to biased estimates of the demand density coefficient.