Measuring Cross-Country Differences in Misallocation

Martin Rotemberg[*] and T. Kirk White[†]

[*]New York University
[†]Center for Economic Studies, U.S. Census Bureau

July 14, 2017

Abstract

We describe differences between the commonly used cleaned version of the U.S. Census of Manufactures and what establishments themselves report. Following the methodology of Hsieh and Klenow (2009), we show that several editing strategies, including industry analysts' manual edits, dramatically lower measured losses in the U.S. data: from around 371% in the collected data to 62% in the Census-cleaned data. Many of these types of edits are infeasible in non-U.S. datasets. We therefore reanalyze the iconic Hsieh and Klenow (2009) result using common data cleaning strategies for the U.S. and for India: a standard trimming-outliers approach and a new Bayesian approach for editing and imputation. Under both methods, there is little evidence that measured misallocation is significantly higher in India than in the United States.

Some of the research in this paper was conducted while the second author was an employee of Census Bureau. 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. Thanks to Mitsu Nishida and Amil Petrin, with whom we started this project, for Hang Kim for his helpful comments, and to Lee Tanenbaum and Weitao Lin for their expert research assistance.

I Introduction

The puzzle of large cross-country differences in productivity has recently been attributed to within-industry misallocation of factors. However, unlike inputs and outputs, distortions cannot be observed directly. Thus, researchers must undertake two steps, in order to measure the extent of misallocation. First, assumptions must be made about firm behavior in the absence of distortions, which allows the use of observed behavior to estimate the size and magnitude of existing distortions. Second, the measured distortions are plugged into a model of consumer behavior to calibrate what the aggregate gains would be under different counterfactuals. The seminal paper of Restuccia and Rogerson (2008) develops a framework under which greater misallocation of resources leads to more dispersion in the distribution of plant-level total factor revenue productivity (TFPR). Hsieh and Klenow (2009) show that under relatively standard assumptions, within-industry variation in the revenue shares of each input is evidence of idiosyncratic firm-input distortions, and provide a simple algorithm for calculating the productivity gains from equalizing those distortions across firms. Taking their model to data, they find that "moving to U.S. efficiency would increase TFP by 30%–50% in China and 40%–60% in India."

However, as Hsieh and Klenow (2009) note, measurement error may look to the researcher like misallocation of resources. Firms that report inaccurate information may only spuriously appear to be using a socially inefficient quantity of resources. The converse is true as well, since firms may report values which are in line with the model but do not reflect reality on the ground. As a result, the confidence we have in our measures of misallocation - either measurements of "true" values for a particular country, or of cross-country differences - depends on the extent of measurement error. In this paper, we discuss two potential sources of measurement error: firms that potentially misreport their own characteristics, and subsequent data cleaning which potentially removes actual distortions.

We show that both editing clear reporting errors (such as when firms report distinct values for the same outcome), as well as more subjective edits (such as edits by industry experts) have large effects on measured misallocation. If instead of using the U.S. Census Bureau's cleaned data we simply trim 1% outliers in the data self-reported by each establishment, we find that moving to the new measured U.S. efficiency would *decrease* measured TFP by around $^2/_3$ for both India and China.

The Hsieh and Klenow (2009) insight is as follows: with CES demand and a constant returns to scale production function, revenue productivity (TFPR) is equalized across firms in the absence of distortions (regardless of any underlying variation in quantity productivity). Nevertheless, in the data there is substantial within-industry variation in TFPR. Hsieh and Klenow (2009) rationalize those differences with idiosyncratic distortions on the firm-specific prices for capital and output. Each firm's distortions can be calibrated using the firm's first-order conditions. They then use the model to generate an elegant expression for the potential gains from reallocation from equalizing TFPR across firms. Other models of misallocation have similar features (Banerjee and Duflo, 2005; Restuccia and Rogerson, 2008; Hopenhayn, 2014).

Rather than contributing to the theory of measuring misallocation in the presence of measurement error (as in Bils et al. 2017), we focus on the efforts undertaken by national statistics agencies when firms do not report (or report unlikely) information. Most statistics agencies initially ask firms to verify (or send in) the information, but the subsequent steps vary across surveys. Unlike its Indian and Chinese counterparts, the U.S. Census Bureau both edits and imputes responses.[1] The exact procedures vary across industries and time (White et al., 2017), but broadly take two forms. First, the Census Bureau *edits*

[1] We have confirmed these contrasts in the documentation for the data, as well as via email communications with the relevant national statistics agencies.

some outliers. If a reported variable fails one or more edit rules, then it may be temporarily replaced with a missing value. Second, the Census Bureau *imputes* missing information, using other information reported by the plant (both in that year and in previous years) and other plants in the same industry.[2] For 2002 and 2007, we have access to the original values reported by firms for plants in the Census of Manufactures,[3] which allows us to know exactly which entries were imputed or entered in the cleaned Census data.[4] In order to focus our attention on the role of measurement, we follow the Hsieh and Klenow (2009) model exactly.

Taking the Hsieh and Klenow (2009) model as given, we have two mains goals in this paper. Our first goal is to describe the data cleaning efforts undertaken by the U.S. Census Bureau. After qualitatively describing the different sets of edit rules, we assess their effects indirectly by measuring how much each edit rule affects measured misallocation in the U.S. Census of Manufactures. We do so in two ways: first by describing how much measured misallocation changes when we *only* use each particular edit rule, and second by each edit's Shapley (1953) value.

The most important edits are "logical" edits and analyst corrections. Logical edits are possible when the Census implicitly asks for the same information in multiple ways, for instance by asking for the total value of shipments as well as the total value of shipments for each product. If the two values for the same outcome diverge, the Census may edit the reported total with the sum of the disaggregated components. Analyst corrections rely on the expertise of full-time industry specialists employed by the U.S. Census Bureau.

[2] Firms that have a variable edited have that variable imputed as if the firm had not reported anything for that variable. Note that the imputed data must also pass the editing rules. For most establishments, at least one of the variables needed to calculate TFP is imputed (White et al., 2017). For payroll and number of employees, the Census Bureau uses administrative records (mainly IRS payroll data) to replace reported data that fails edit rules. The Census Bureau classifies these changes from the reported data as "non-imputes". However, these non-imputes still change plants' measured TFP.

[3] It is worth noting that when Hsieh and Klenow (2009) was written, neither imputation flags nor this data were available for the Census years used in their study (1977-1997).

[4] Researchers have almost exclusively used the cleaned data for studies on manufacturing in the U.S.

Imputation of missing data also lowers measured misallocation.

In many international datasets with microdata on firms (including in India), these types of edits are often infeasible. For instance, when redundant questions are not asked logical edits cannot be implemented. We reanalyze the cross-country comparison after cleaning the raw (plant-level) data reported by firms in a common fashion, using the information which is commonly available in most firm-level datasets.[5] We have two strategies.

Our most parsimonious approach trims the 1% or 2% tails of the distributions of the firms' distortions and TFPR (relative to each plant's industry mean in the corresponding year), with no additional imputation or editing to the self-reported data.[6,7] In both 2002 and 2007, in the raw data (after either not trimming, or trimming the 1% or 2% tails of the distributions of firms' distortions and industry-year demeaned TFPR) the U.S. appears to have substantially more misallocation than India and China. We do not take this result literally - we do not think that we have compelling evidence that the U.S. manufacturing sector is characterized by more misallocation than most other countries. Instead, we consider our results a "smoking gun" that measurement (and data processing in particular) is deeply important to the study of misallocation.

Trimming is a popular data cleaning strategy because it is easy to understand and implement, but there are several reasons to look for alternative data cleaning strategies. Trimming is a blunt instrument, and varies paper to paper: while Hsieh and Klenow (2009) trim static measured distortions, other papers using the exact same data have

[5] Hsieh and Klenow (2009) implicitly use a similar strategy for the measurement of capital. For capital, they use the book value, since that is the only such variable available in their cross-sectional Indian data, and for Census of Manufactures firms that are not in the Annual Survey of Manufacturers sample. We are essentially applying the logic of treating the cross-country data as similarly as possible to the cleaning step.

[6] If a firm does not report one of the variables needed to estimate misallocation, it is not used in the estimation. As a result, we do not need to focus our attention on imputation for missing data for this part.

[7] Hsieh and Klenow (2009) also undertake a trimming strategy within the U.S., Indian, and Chinese datasets they use.

trimmed other characteristics such as growth rates (for instance, Allcott et al. 2016). We propose a more reproducible approach: *one* common and theoretically-motivated data cleaning exercise, which could then be used across firm-level datasets without further need for data pre-processing.

To that end, we adopt the data cleaning strategy proposed by Kim et al. (2015). Unlike trimming, which drops outliers and leaves missing values blank, the Kim et al. (2015) method simultaneously edits and imputes the data. First, we look at the ratios of reported variables, and flag the outliers of the ratios. We then impute entries in order for the cleaned data to pass the edit checks. Unlike the imputation methods the Census Bureau uses most frequently in the Census of Manufactures, the Kim et al. (2015) method tries to preserve the joint distribution of the covariates. Given the ratio outlier flags, we favor making edits that are likely given our model for misreporting, and similarly impute values that are likely given the underlying model for the data. The imputation step works for missing values as well as those which are flagged.

In the next section, we recap the theory of distortions underlying our analysis. Section 3 discusses the extant data collection and cleaning procedures in the United States. Section 4 describes the Kim et al. (2015) method for cleaning plant-level data. In Section 5, we compare commonly-cleaned data for the U.S. and India. Section 6 concludes.

II A Theory of Misallocation

In this section, we briefly describe the Hsieh and Klenow (2009) approach to measuring misallocation that we follow. First, we start from the firm side of the problem, showing how firm behavior is affected by idiosyncratic distortions on capital and output. In the model, variation in those distortions is captured by variation in firm-level revenue productivity. We then turn to the aggregate side, and derive how aggregate productivity would be affected in a counterfactual where the variation in revenue productivity is removed.

II.A Firm-level Distortions

Overall utility Y is a Cobb-Douglas aggregate over sectoral output Y_s,

$$Y = \prod Y_s^{\theta_s},$$

so normalizing the price of the final good to 1, expenditure for each sector is a fixed proportion

$$P_s Y_s = \theta_s Y$$

where P_s is the price index for sector s.

Within each sector, output takes a CES form over output of each variety Y_{si}:

$$Y_s = \left(\sum_{i=1}^{M} Y_{si}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}$$

and each firms produces value added using capital and labor, with Cobb-Douglas production-function elasticities which vary across sectors:

$$Y_{si} = A_{si} K_{si}^{\alpha_s} L_{si}^{1-\alpha_s}.$$

The wage and rental rate are constant in the economy, but firms face idiosyncratic distortions on output and capital. As a result, each firm's profits are:

$$\pi_{si} = \left(1 - \tau_{Y_{si}}\right) P_{si} Y_{si} - w L_{si} - \left(1 + \tau_{K_{si}}\right) R K_{si}.$$

Marginal revenue productivity for each input is

$$\begin{aligned} MRPL_{si} =& \frac{\sigma-1}{\sigma}\left(1-\alpha_s\right)\frac{P_{si}Y_{si}}{L_{si}} \\ MRPK_{si} =& \frac{\sigma-1}{\sigma}\left(\alpha_s\right)\frac{P_{si}Y_{si}}{K_{si}}. \end{aligned}$$

and each firm's revenue productivity is

$$TFPR_{si} = \frac{P_{si}Y_{si}}{K_{si}^{\alpha_s}L_{si}^{1-\alpha_s}} = P_{si}A_{si} \propto MRPL_{si}^{1-\alpha}MRPK_{si}^{\alpha}. \tag{1}$$

II.A.1 Optimization Behavior

Profit maximization implies that:

$$\begin{aligned} P_{si} =& \frac{\sigma}{\sigma-1}\left(\frac{R}{\alpha_s}\right)^{\alpha}\left(\frac{w}{1-\alpha_s}\right)^{1-\alpha}\frac{1}{A_{si}}\frac{\left(1+\tau_{K_{si}}\right)^{\alpha_s}}{\left(1-\tau_{L_{si}}\right)} \\ A_{si} \propto& \frac{\left(P_{si}Y_{si}\right)^{\frac{\sigma}{\sigma-1}}}{K_{si}^{\alpha_s}L_{si}^{1-\alpha_s}} \end{aligned} \tag{2}$$

$$\begin{aligned} wL_{si} &= \left(1-\tau_{Y_{si}}\right)\frac{\sigma-1}{\sigma}\left(1-\alpha_s\right)P_{si}Y_{si} \\ \Rightarrow MRPL_{si} &= \frac{w}{\left(1-\tau_{Y_{si}}\right)} \end{aligned} \tag{3}$$

$$\begin{aligned} \left(1+\tau_{K_{si}}\right)RK_{si} &= \left(1-\tau_{Y_{si}}\right)\frac{\sigma-1}{\sigma}\left(\alpha_s\right)P_{si}Y_{si} \\ \Rightarrow MRPK_{si} &= \frac{R\left(1+\tau_{K_{si}}\right)}{\left(1-\tau_{Y_{si}}\right)} \end{aligned} \tag{4}$$

As a result, combining Equations 1, 3, and 4 gives

$$TFPR_{si} \propto \frac{\left(1+\tau_{K_{si}}\right)^{\alpha}}{\left(1-\tau_{Y_{si}}\right)}, \tag{5}$$

so revenue productivity is only a function of the distortions, and not directly a function of firm TFP. As a result, in the absence of distortions, TFPR would be equalized across firms. In the next subsection, we show how variation in TFPR affects aggregate productivity

II.B Aggregate Distortions

Aggregate productivity in each sector is

$$TFP_s = \frac{Y_s}{K_s^{\alpha_s} L_s^{1-\alpha_s}} = \frac{\overline{TFPR}_s}{P_s}, \tag{6}$$

where, given cost-minimization, the price index for sector s is:

$$P_s = \left(\sum_{i=1}^{M} P_{si}^{1-\sigma} \right)^{\frac{1}{1-\sigma}}.$$

From Equation 1, we can rewrite the price index as

$$P_s = \left(\sum_{i=1}^{M} \left(\frac{A_{si}}{TFPR_{si}} \right)^{\sigma-1} \right)^{\frac{1}{1-\sigma}},$$

and plugging back in to Equation 6 gives the core Hsieh and Klenow (2009) expression for productivity

$$TFP_s = \left(\sum_{i=1}^{M} \left(A_{si} \frac{\overline{TFPR}_s}{TFPR_{si}} \right)^{\sigma-1} \right)^{\frac{1}{\sigma-1}}, \tag{7}$$

Since we know from Equation 5 that $TFPR_{si}$ would only be different from $\overline{TFPR}_s$ in the presence of distortions, the "efficient" counterfactual TFP is $\overline{A}_s = \left(\sum_{i=1}^{M} A_{si}^{\sigma-1} \right)^{\frac{1}{1-\sigma}}$, and so (aggregating over all sectors)

$$\frac{Y_s}{Y_{s(efficient)}} = \prod_{s=1}^{S} \left[\sum_{i=1}^{M_s} \left(\frac{A_{si} \overline{TFPR}_s}{\overline{A}_s TFPR_{si}} \right)^{\sigma-1} \right]^{\frac{\theta_s}{\sigma-1}}. \tag{8}$$

Equation 8 can be calculated from observed data. Instead of measuring how sensitive our calculation of productivity gains are to different underlying assumptions, which has been the primary focus of much of the recent methodological literature on misallocation, we instead calculate Equation 8 using different cuts of the data, which we describe in the next section.

III Data Cleaning in the United States

We primarily use micro-data from the United States, from the 2007 U.S. Census of Manufactures. The quinquennial survey covers roughly 300,000 manufacturing plants, although information for the smallest plants - roughly one third of the sample - are almost entirely imputed. The standard is to exclude these so-called administrative records plants, and we follow that standard throughout our analysis.

As in most surveys, not all respondents answer all of the questions, and for some plants some responses seem inconsistent with each other. The Census Bureau has created imputation and edit rules for this data, which are described in Grim (2011). However, until the 2002 Census, it was difficult for researchers to identify which, if any, responses for a given plant were imputed. We go beyond the imputation flags and use the newly available actual responses from the establishments themselves (the "reported" data).

The reported data differs from the final ("cleaned") data in two respects.[8] First, missing values due to non-response in the reported data are imputed in the cleaned data, using a variety of industry-specific regression-based and other imputation strategies. Second, actual responses which fail edit rules in the reported data are also imputed or changed in some way in the final data. The most important edit rules are balance edit rules (certain variables have to add up) and ratio edit rules (ratios of certain variables must be within certain bounds). Edit-rule-failing responses are replaced using a variety of methods, described in Table 1, and the ones that affect measured misallocation the most are described

[8] Another convention for naming the data is "captured" data for the reported information, and "completed" for the cleaned data.

more fully in Subsection III.B.

III.A Measured Misallocation in the Raw U.S. data

First, we consider the effects on measured misallocation of replacing cleaned with raw data in the U.S. manufacturing sector. In 2007, The potential gains from removing misallocation are 371% with the only data cleaning being trimming the 1% extremes of the wedges and TFPQ, and 62% in the cleaned data. Table 2 shows the results of calculations across a range of similar datasets. The largest measured misallocation comes from the raw data with no trimming, where the measured losses are 4293%. The smallest values come from trimming the 2% extremes in the Census-cleaned data, which leads the measured losses to be only 43%.[9]

III.A.1 Measured Cross-Country Differences in Misallocation

We now turn to discussing cross-country differences in measured misallocation. While we have measured misallocation in the United Sates for a large set of data choices, in order to avoid tedium we only describe cross-country differences in misallocation for two extremes: the 2002 and 2007 average for Census-cleaned data, and the corresponding average in the Census-reported data.[10] Our measures of misallocation internationally come from a variety of published sources discussed in Appendix A.I. The results are shown in Table 3. While estimated misallocation in almost every developing country is higher than that for the cleaned U.S. data, in its reported data the U.S. has a higher level of measured misallocation than for any other country. Taking the results literally would imply that for, e.g., Argentina moving to the U.S. level of misallocation would decrease manufacturing TFP by around $^2/_3$.

[9] The values in 2002, the other year for which we also have both cleaned and reported data, are similar. For instance, the measured losses after trimming the 1% extremes are 333% in the raw data and 44% in the cleaned data.

[10] For all values, when possible, we use the reported values from trimming the 1% extremes for TFPQ and the distortions for output and capital.

In a more speculative approach, we build on Kalemli-Ozcan and Sorensen (2012) and calculate the measured gains within country-years in the World Bank Enterprise Analysis Unit's Enterprise Surveys.[11] The surveys are relatively small, and for the countries for which we have both enterprise surveys and data from national statistics agencies, the former have higher measured losses from misallocation. Nevertheless, Figure 1 shows that the gains from the U.S. are larger than the corresponding gains for around 60% of the enterprise survey.[12]

III.B The effect of edits on measured misallocation in the raw U.S. data

There are many different data cleaning steps on the road from measured misallocation of 371% in the 2007 raw data to 62% in the corresponding clean data. In order to determine if changes to the reported data are needed, the Census Bureau primarily uses balance edit rules and ratio edit rules. An example of a balance edit rule is that the number of production workers plus the number of non-production workers must equal the total number of employees. For ration edit rules, the Bureau uses ratios of reported values (for instance, one of the ratios is the total value of shipments over annual payroll). The Census determines industry-specific upper and lower bounds (either by looking at the percentiles of reported outcomes or by relying on additional industry-specific knowledge). If a plant's reported values violate one or more of the balance or ratio edit rules, one ore more of the reported values is replaced using one of types of edits described in table 1.

In this subsection, we characterize the effect of each edit. For eight edit supercategories, we measure misallocation in the U.S. using raw data for everything but those

[11] The raw data is available at `http://www.enterprisesurveys.org/data`. Our version is from August 1 2016, and we use the most recent sample from the 18 countries where we do not have access to an actual census, and there are at least 250 firms that report sales, labor, materials, and the replacement value of capital. We also drop Turkey in 2013 and Nigeria in 2014, since the measured gains from removing distortions are over 58000% in those countries, which is implausibly high. Further details of data construction are in Appendix A.I

[12] When the World Bank Enterprise Group uses the data to calculate firm TFP, they undertake a careful process to clean the data and remove outliers. We, however, use the raw data.

edits (and using the cleaned data for the plants affected by the given edit). The eight supercategories are: logical edits (separately for shipments, materials, and payroll), administrative edits, regression imputes, rounding imputes, analyst imputes, and the rest, which we describe in turn.

Logical edits are done when there are many survey questions which ask for the same information. For instance, total value of shipments shows up in three different parts of the survey: (1) there is a question that asks for the total value of the plants shipments; (2) there are many questions about the value of shipments for specific products that a given industry produces (these values can be summed by the U.S. Census Bureau); and (3) there is a question – separate from (1) – that asks the respondent to total the values of the products in (2). If these values differ by more than a certain amount (the same tolerance is used for all industries within a year, but has varied over time), then the Census compares each of them to annual payroll for the same plant and then takes the "best" one. The "best" one is selected in a form similar to the regression imputes, described below.

Administrative edits are similar to logical edits, but differ in that the alternative source of information is from administrative records. For instance, for payroll the administrative records come from IRS payroll tax records. Again, if the administrative data differ from the reported values, the reported values may be replaced.

Regression imputes are used to edit data when alternative sources of information are not available. The U.S. Census Bureau uses a variety of industry-specific regression-based imputation strategies. Since they do not require any observed alternative value, regression imputes are also used to impute missing values when no other information is available for a given variable. In general, regression imputes create predictions using one other variable, and (for plants surveyed in the Annual Survey of Manufactures), one-year lags of the imputed variable as well. Unlike administrative and logical edits, there is not necessarily any direct evidence that the reported firm value may be incorrect.

In order to measure the importance of each edit, we undertake the following exercise. For each type of edit, we replace *all* of an establishment's information with the clean data if it was affected by the edit. For example, to understand the importance of logical edits for total value of shipments, we use the cleaned outcomes for firms whose total value of shipments has a logical edit flag, and the raw outcomes for the other firms. We show how much measured misallocation in the U.S. is affected by each edit by showing (a) the change in misallocation for each edit if it is the only one applied (this always decreases measured misallocation), and (b) its Shapley value. For this context, the Shapley value is the value of the following thought exercise: we first consider all possible combinations of edits, done one at a time. We credit each edit for its marginal contribution within each (ordered) combination, and calculate the Shapley Value as the average of those marginal contributions. The results are in Table 4, with the third column reporting the share of the total decrease from all the edits that the Shapley procedure credits to each edit type.

The most important edits are logical imputes for total value of shipments, which is responsible for a fifth of the decrease in measured misallocation, as well as imputes for the missing values, which are collectively responsible for another fifth. Analyst corrections (for any TFPR variable) and logical imputes for payroll are also each responsible for over 10% of the decline.

The results of Table 3 are unsatisfying - cross-country comparisons of measured misallocation in datasets which have been cleaned differently will pick up differences due to both underlying cross-country differences and cross-country differences in data cleaning. However, while comparing raw data solves the latter problem, it does so at the expense of introducing new errors. The natural solution is to compare datasets which have been commonly cleaned. While one strategy may be to use the approach of the U.S. Census Bureau everywhere, Table 4 shows that around ⅔ of the changes - the logical imputes, analyst corrections, and administrative record edits - are difficult if not impossible to replicate in other contexts (depending on the availability of alternative reports for the same outcome and industry specialists). As a result, in the next two sections we describe and then implement an algorithm for editing and imputing raw firm-level (or plant-level) data.

IV A Common Approach to Cleaning Data

Suppose that establishment i reports p characteristics, $\boldsymbol{y}_i = \left\{y_{i1}, y_{i2} \ldots y_{ip}\right\}$ (where items could be missing). The corresponding true values are $\boldsymbol{x}_i = \left\{x_{i1}, x_{i2} \ldots x_{ip}\right\}$, with s_{ij} indicating if response j for establishment i is incorrect. Given the dataset of reported values $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_n\}$ the goal of data cleaning is to find the dataset of true values $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$. In order to do this, we follow the approach of Kim et al. (2015).

First, we define the feasible region $\mathcal{D}$ of plausible reports. This region limits possible values due two rules: balance rules which require entries to add up (for instance, non-production wages + production wages = total wages, or more generally $\left(x_{iT_\ell} - \sum_{j\in\beta_\ell} x_{ij} = 0\right)$ for x_{iT_ℓ} as the total for the ℓth balance rule for the set of component variables β_ℓ), and a set of ratio edit rules which bound the ratios of any two variables. While the balance rules are *a priori* clear, the ratio edit rules can come either from industry specific knowledge, or from outliers in the data itself. Fellegi and Holt (1976) note that the set of explicit ratio edit rules can imply additional ones as well.[13] While $\boldsymbol{s}_i$ is not directly observed, A_i indexes the failed ratio & balance edit rules.

After cleaning the data, we want our cleaned data to be likely given a model for reporting error, likely given a model for error indicators, and likely given a model for the underlying data. More formally,

$$f\left(\boldsymbol{x_i}, \boldsymbol{s_i} | \boldsymbol{y_i}, A_i\right) \propto f\left(\boldsymbol{y_i} | \boldsymbol{x_i}, \boldsymbol{s_i}, A_i\right) f\left(\boldsymbol{s_i}, A_i | \boldsymbol{x_i}\right) f\left(\boldsymbol{x_i}\right). \tag{9}$$

[13] For instance, rules $x_1 \leq x_2$ and $x_2 \leq x_3$ imply $x_1 \leq x_3$.

For the model of reporting error, we maintain U.S. Census Bureau's (implicit) approach: data reported with error provides no information on the true value. Therefore, $f\left(\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{s_i}, A_i\right)$ is uniform over the support of feasible values if $y_{ij} \neq x_{ij}$.

However, unlike the Census Bureau, we also assume a uniform distribution for the errors. That is to say, we do not use weights on which variables are more likely to be reported with error, so all candidates $\boldsymbol{s_i}$ that result in feasible solutions are equally likely.

For the model for the underlying data, we assume that each establishment belongs to one of K mixture components (z). After assuming K[14], we need to estimate the probability of membership in each component (π), and within each mixture the mean vector $(\boldsymbol{\mu})$ and covariance matrix $(\boldsymbol{\Sigma})$. In order to ensure that all of the draws will pass both the balance and ratio edits, we impose that the distribution of $\boldsymbol{x_i}$ conditional on μ, Σ, z_i, given feasible region $\mathcal{D}$ is

$$f\left(\boldsymbol{x}_i|\theta_i\right) = \mathcal{N}\left(\boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right) \prod_{l=1}^{q} \delta\left(x_{iT_\ell} - \sum_{j \in \beta_\ell} x_{ij}\right) \mathbb{1}\left[\boldsymbol{x}_i \in \mathcal{D}\right]$$

where $\delta\left(\cdot\right)$ is that Dirac delta function with the point mass at zero and $\boldsymbol{x}_{i,NT}$ is the set of reported values which are themselves totals of other reported values.

We run a single chain of Markov Chain Monte Carlo for 1000 iterations for the final cleaned dataset. This consists of first proposing $\boldsymbol{s_i}$ which are consistent with $\boldsymbol{A_i}$, and then editing values $\boldsymbol{y_i}$ given the draw of $\boldsymbol{s_i}$ and the underlying probability distributions for the responses which were not reported with error.

V Cross Country Differences in Measured Misallocation

The main choice associated with the model in Section IV is defining the feasible region $\mathcal{D}$. We do so by following the resistant fences method, which is the starting point for how Census chooses its ratio bounds (Thompson and Sigman, 1999). Within each industry,

[14] In practice we set K=50, which is large enough that no data are in the lowest probability components.

for each log ratio $r_{jk} = \ln\left(\frac{y_j}{y_k}\right)$, we calculate its 25th and 75th percentiles, Q_{jk}^{25} and Q_{jk}^{75}, and therefore the interquartile range IQR_{jk} . We then flag all ratios that are either smaller than $Q_{jk}^{25} - C \times IQR_{jk}$ or larger than $Q_{jk}^{75} + C \times IQR_{jk}$ where C is a pre-specified threshold. The variables we use are the total cost of materials, total value of shipments, number of blue and white collar workers, blue and white collar wages, total capital, inventory at the beginning and end of the year, total benefits, and (in India) the sampling weight, and we run the estimation separately by industry (6 digit in the U.S. and 2 digit in India). In Table 5, we report results from $C = 1.5$ for both the Indian (2002) and U.S. (2007) data.[15]

The gap between the raw U.S. and raw Indian data shrinks after applying the common data editing procedure. In the commonly-cleaned data, misallocation is slightly higher in the U.S. than in India (65% vs 63%). The multiplier of 1.5 also results in measured misallocation that is similar to that in the Census Bureau-cleaned U.S. data with an additional 1% trimming of outliers (65% vs. 62%).

An alternative (simpler) approach to common data cleaning is to just trim outliers in the raw data. We report going from no trimming to 1% trimming to 2% trimming. This has an unsurprisingly large effect on measured misallocation in the US, which falls from 4,293% to 319% to 264%, and India (for which the measured values fall from 147% to 91% to 76%), but the relative magnitudes are similar in the datasets with 1% or 2% trimming. A similar pattern is found if we both clean the data using the Kim et al. (2015) approach and also trim the 1% or 2% tails of the wedges and productivity, although in these cases misallocation is now somewhat higher in India than in the U.S.

VI Discussion

In this paper, we use previously unexplored versions of the United States Census of Manufacturers for 2002 and 2007 in order to investigate the role that measurement plays for

[15] We are running alternative cut-offs as well, and are hoping to clear those results soon.

estimating misallocation. We have two complementary goals. The first is to quantify the importance of data cleaning. We show that in the data that is reported directly to the Census Bureau by American establishments, measured misallocation in the United States is substantially higher than for any other country for which we have data from an official statistics agency. We do not take this result literally: there are many reasons to believe that comparing the raw U.S. data to its counterparts in other countries is not like-for-like.[16] Furthermore, we show that many of the important edits undertaken in the U.S. are infeasible in other settings, because they either use multiple responses for the same information or because they rely on industry experts. When we use common data cleaning strategies, we find little or no evidence that measured misallocation is significantly higher in India than in the United States.

While we demonstrate that there is a large scope for different measurement choices to affect the estimation of misallocation in manufacturing, our moral is not nihilistic. We suggest an alternative approach for cleaning firm-level data. In addition to being crucial for comparing cross-country differences in outcomes, the method is more broadly useful for data cleaning instead of more traditional ad-hoc approaches such as winsorizing.

References

Allcott, H., A. Collard-Wexler, and S. D. O'Connell (2016). How do electricity shortages affect industry? evidence from india. *The American Economic Review 106*(3), 587–624.

Banerjee, A. V. and E. Duflo (2005). Growth Theory through the Lens of Development Economics. *Handbook of Development Economics 1*(05), 473–552.

Bartelsman, E. J. and W. Gray (1996). The NBER Manufacturing Productivity Database.

[16] For no country do we know if measured misallocation in the raw data is larger or smaller than it is in reality, nor do we have a way of comparing the relative precision of self-reported information across countries. We do know that the vast majority of Indian firms are unable to fill out their survey forms on a computer.

Bils, M., P. J. Klenow, and C. Ruane (2017). Misallocation or mismeasurement? Technical report, Working Paper.

Busso, M., L. Madrigal, and C. Pages (2013). Productivity and Resource Misallocation in Latin America. *B.E. Journal of Macroeconomics 13*(1), 903–932.

Fellegi, I. P. and D. Holt (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association 71*(353), 17–35.

Grim, C. (2011). User notes for 2002 census of manufactures. *Unpublished Technical Note*.

Hopenhayn, H. A. (2014). On the Measure of Distortions. *Working Paper*.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and Manufacturing Tfp in China and India. *Quarterly Journal of Economics 124*(4), 1–55.

Kalemli-Ozcan, S. and B. Sorensen (2012). Misallocation, Property Rights, and Access to Finance: Evidence from Within and Across Africa. *Working Paper*.

Kim, H. J., L. H. Cox, A. F. Karr, J. P. Reiter, and Q. Wang (2015). Simultaneous edit-imputation for continuous microdata. *Journal of the American Statistical Association 110*(511), 987–999.

Nishida, M., A. Petrin, M. Rotemberg, and T. K. White (2015). Are We Undercounting Reallocation's Contribution to Growth? *Working Paper*.

Randy Becker , Wayne Gray, J. M. (2016). NBER-CES Manufacturing Industry Database: Technical Notes. *National Bureau of Economic Research Technical Working Paper Series*.

Restuccia, D. and R. Rogerson (2008, oct). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–720.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2*(28), 307–317.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality*, 1–28.

Thompson, K. J. and R. S. Sigman (1999). Statistical methods for developing ratio edit tolerances for economic data. *Journal of Official Statistics 15*(4), 517.

White, T. K., J. P. Reiter, and A. Petrin (2017). Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion. *The Review of Economics and Statistics*.

Yang, M.-J. (2012). Micro-level Misallocation and Selection: Estimation and Aggregate Implications. *Working Paper*.

A Appendix

A.I Cross-Country Estimates of Misallocation

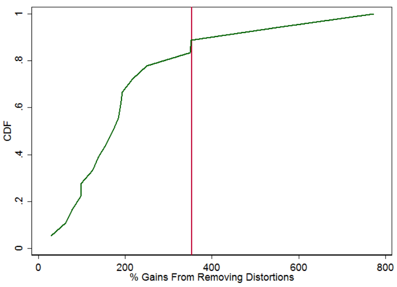
We use a variety of published sources to report the measured gains from reallocation in Table 3. With the exceptions of Chile and Columbia, our estimated potential gains from reallocation for South America come directly from Busso et al. (2013), and in Indonesia from Yang (2012). Both of those sources use information from national firm censuses. In addition to those values, we report information from Chile, Columbia, India and Slovenia, using the same micro-data described in Nishida et al. (2015).

The Chilean and Colombian data are annual and we use 1995 and 1991, respectively. The Chilean data, provided by Chile's Instituto Nacional de Estadistica (INE), cover all manufacturing plants with at least 10 employees. The Colombian data from the Annual Manufacturing Survey, provided by Colombia's Departamento Administrativo Nacional de Estadistica (DANE), cover all plants with at least 10 employees.

In India, we use the Annual Survey of Industries (the ASI). Factories with over 100 workers are surveyed every year, while smaller establishments are surveyed every few years (the ASI is designed to be representative at the State by Industry level, so firms without local competitors are more likely to be surveyed). Hsieh and Klenow (2009) use the same dataset, and we follow standard practice in generating measures of value added, capital, and payroll. Industries are grouped using India's NIC (National Industrial Classification) codes, and we report the value of reallocation for 2009. For Slovenia we rely on annual accounting data provided by the Slovenian Statistical Office which covers all manufacturing firms. The Slovenian data, unlike its counterparts in most other countries, is at the firm (not establishment) level, and we report the estimates for 2004. For the U.S. and India, we use cost-shares from the NBER-CES Manufacturing Industry Database as our measures of industry production elasticities, and multiply the book value of capital by 10% in order to impute the cost of capital.

The countries we use in the World Bank Enterprise Surveys (used for Figure 1) are Bangladesh, Colombia, Egypt, Iraq, Jordan, Kenya, Malaysia, Mozambique, Peru, Philippines, Russia, South Africa, Sri Lanka, Sweden, Thailand, Tunisia, Vietnam, and Zimbabwe. We use "Total annual cost of labor" as the measure of labor, the sum of the cost of materials, electricity, communications services, fuel, transport, and water for materials, the sum of the (self-reported) replacement costs for machinery and land for capital, and total sales for gross output. For each manufacturing sector we assume that the capital elasticity is $^1/_3$. Across all surveys, we drop firms that report non-positive values for any of those four variables.

Figure 1: Potential Gains from Reallocation in the World Enterprise Surveys



Notes: This figure plots the gains from removing distortions for 19 countries in the World Bank Enterprise Surveys, described in Appendix A.I. The vertical line corresponds to the average gains in the United States in the raw data for 2002 and 2007.

Table 1: Edits Made to the U.S. Census of Manufacturers

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).
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.)
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.

Notes: Edit rule descriptions are from Grim (2011) and White et al. (2017).

Table 2: Mesaured Misallocation in the 2007 U.S. Census of Manufacturers

	Trimming		
	0%	1%	2%
Census-Cleaned	165%	62%	43%
Raw	4293%	371%	263%

Notes: The values follow Hsieh and Klenow (2009). Each cell represents a different starting point: either the Census-cleaned or raw data, and trimming the 0,1, or 2 percent extremes for TFPQ, the capital wedge, and the output wedge.

Table 3: Measured Cross-Country Differences in Misallocation

	Gains in	Gains Relative to 2002/2007 average:	
Country	Most Recent Year	Clean US	Reported US
Mexico	95%	32%	-57%
India	91%	29%	-58%
China	87%	26%	-59%
Chile	77%	19%	-61%
Indonesia	68%	13%	-63%
Venezuela	65%	11%	-64%
Bolivia	61%	8%	-65%
Uruguay	60%	8%	-65%
Argentina	60%	8%	-65%
Ecuador	58%	6%	-65%
Slovenia	57%	6%	-65%
El Salvador	57%	6%	-65%
Colombia	49%	1%	-67%
Brazil	41%	-5%	-69%

Notes: Each cell shows measured misallocation and the "gains" from moving to measured-US levels. Data sources are discussed in Appendix A.I

Table 4: Changes in Measured Misallocation due to Edits

Edit	Direct Effect on Measured Misallocation	Shapley Value	Shapley Share
TVS logical impute	-166%	-64%	0.21
Impute for Missing	-155%	-58%	0.19
Analyst correction	-130%	-48%	0.16
Payroll (SW) logical impute	-140%	-40%	0.13
Divide by 1000 edit	-89%	-24%	0.08
Other imputes	-85%	-27%	0.09
Regression imputes for Materials	-73%	-19%	0.06
Logical imputes for Materials	-59%	-21%	0.07
Administrative Record Impute	-6%	-5%	0.02

Notes. This table shows the effect of each type of edit on measured misallocation in the United States. Column 1 reports the difference in measured misallocation between the original raw data and the raw data, but with the cleaned final data for the firms affected by that row's edit. Since the sum of the changes is not equal to the difference in measured misallocation between the raw and clean data, Column 2 reports each row's Shapley Value and Column 3 the share of the change.

Table 5: Measured Misallocation After Common Data Cleaning

	Raw data			Clean Data: Multiplier = 1.5		
	Trimming			Trimming		
Country	0%	1%	2%	0%	1%	2%
U.S.	4293%	371%	264%	65%	48%	40%
India	147%	91%	76%	63%	58%	53%