Measuring Cross-Country Differences in Misallocation

Martin Rotemberg[⋆]
T. Kirk White[†]

[⋆]NYU
[†]Center for Economic Studies, U.S. Census Bureau

American Economic Association Meetings
San Diego, CA
January 2-5, 2020

Disclaimer

The research in this presentation was conducted while the first author was a Special Sworn Status researcher of the U.S. Census Bureau at the New York Census Federal Statistical Research Data Center and while the second author was an employee of Census Bureau. The views expressed in this presentation are those of the authors and not the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This presentation includes output approved for release under Disclosure Review Board release numbers DRB-B0115-CDAR-20181016, CBDRB-FY19-CMS-7909, CBDRB-FY20-072, and CBDRB-FY20-074.

Motivation: Within-Industry TFP Dispersion is Important

- What we're doing:
  - Dispersion vs measurement error
- Dispersion in firm outcomes is important for a lot of economic models
  - Determines responsiveness to a variety of shocks, such as trade liberalization (e.g., Melitz 2003)
  - Importance of management / R&D / investments (e.g., Bartelsman and Doms 2000, Bloom and Van Reenan 2007)
  - BLS-Census Collaborative Multifactor Productivity Project (CMP)
  - Misallocation and aggregate productivity (e.g. Hsieh Klenow 2009; Bils, Klenow, and Ruane 2018)

Measuring Misallocation (1): Data and Data Cleaning

1. Census data tends to be self-reported
   1.1 US Census does a lot of editing & imputation of raw data (and pushes forward the frontier of knowledge on these topics).
      1.1.1 Other countries (especially developing countries) do not do this
2. Two major types of changes to raw data
   2.1 Imputation for missing data (both unit and item non-response)
   2.2 Editing
      2.2.1 Compare survey responses to administrative records data — correct response data as needed
      2.2.2 Check records for internal consistency and plausibility

Measuring Misallocation (2): Theory

- Productivity growth from reallocation: reallocate inputs from plants with low marginal products to those with high ones
  - Hsieh and Klenow (2009) and Bils, Klenow and Ruane (2018): plants with large (small) distortions have high (low) marginal products
    - Remove distortions –> markets reallocate resources –> get aggregate TFP growth
- Using the HK/BKR model to quantify misallocation, we focus on the role of measurement:
  - How much does data cleaning affect measured allocative efficiency (and thus measured potential for TFP growth from reallocation)?

Measuring Misallocation: Main results (1)

- Census Bureau's data cleaning has an enormous effect on measures of Allocative Efficiency (AE)
  - AE is on average 8 times higher in US Census-cleaned data vs. raw US data
- Effect of Census Bureau data cleaning on measured AE has increased tremendously over time (2002-2012):
  - Ratio of Allocative Efficiency in U.S. cleaned vs. U.S. raw data increased from 4x in 2002 to 87x in 2012

Measuring Misallocation: Main results (2)

- Cross-country differences in data cleaning also have a big impact on cross-country comparisons of Allocative Efficiency
  - Comparing raw U.S. data to raw Indian data, Allocative Efficiency is 3 to 26 times higher in India than in the U.S.
  - Comparing Census-cleaned U.S. data to raw Indian data (a la Hsieh-Klenow) is 20% higher in the US vs India
  - If we apply a common cleaning method to raw data in both countries, AE is roughly the same in both countries in 2002 and roughly 100% higher in India vs. the US in 2007

Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Bils, Klenow, and Ruane (BKR) Set-Up

- Each intermediate good producer i producing in sector s has Cobb-Douglas production function
- Each producer faces idiosyncratic distortions on their prices of capital $\left(\tau_{k_i}\right)$, labor $\left(\tau_{L_i}\right)$, and intermediates $\left(\tau_{M_i}\right)$
- Producers face CES demand

Inferring Plant-level Distortions

- With Cobb-Douglas production functions, efficiency implies that each input's share of revenue = their share of costs (= production function elasticity)
- With no frictions: marginal product of labor = wage
- Implied distortions are the ratio of the revenue share to the cost share (in real data, the revenue share tends to be lower)

$$MRPL_{si} = w\left(1 + \tau_{L_{si}}\right)$$

Inferring Plant-level Distortions

- HK/BKR insight: in the model, with no distortions, all plants in same sector have same $\frac{Y_{si}}{L_{si}^{\alpha_{L_s}} K_{si}^{\alpha_{K_s}} M_{si}^{\alpha_{M_s}}} = TFPR_{si} = \overline{TFPR}_s$
- Can measure the distortions from observed within-industry $TFPR_{si}$ dispersion
- Given the assumed CES demand structure (constant markups), can back out $TFPQ_{si}$ from measured revenue
- HK derive expression for aggregate TFP losses from misallocation (due to within-industry distortions) using value added measures
- BKR (and us) use gross output production functions and add a distortion to intermediates

Misallocation vs. Measurement Error

$$MRPL_{si} = w\left(1 + \tau_{L_{si}}\right)$$

- What could lead to $\tau_{Lsi} \neq \tau_{Lsj}$?
  - Actual distortions: within-industry differences in markups, taxes, labor market frictions, or...
  - Measurement error:
    - Plant has undistorted optimal labor/output ratio, but reports the wrong thing
    - Plant reports optimal labor/output ratio, but Census edits change reported values
    - Plant doesn't report fully and Census imputes the missing values

Whole Economy Measures

- Hard to compare TFPQ and TFPR across sectors
- So we normalize within sector to create aggregate measures

Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Editing/Imputation

- The Census Bureau imputes data in the CM for several reasons
  - Unit non-response
  - Item non-response
  - Response data fails edit checks (e.g., payroll/employee=$1 billion per employee)
- In this paper we focus on imputation for output, labor, and materials
  - Capital is known to be hard to measure
  - Census uses simple imputation models to replace missing/faulty data on value of shipments, cost of materials
  - Employment and payroll edits mostly come from administrative records — many significant changes to reported values

U.S. Census Bureau Imputation Strategies

- For many key variables, the most frequently-used imputation methods in the Census of Manufactures are not designed to reproduce the within-industry dispersion we see in the non-imputed data
- Industry-specific regression model to impute input Y given observed Xs (plant i, industry s, year t):

$$Y_{ist}^{impute} = \beta_j X_{ist}$$

or

$$Y_{ist}^{impute} = \beta_{s1} X_{ist} + \beta_{s2} Y_{is,t-1} + \beta_{s3} X_{is,t-1}$$

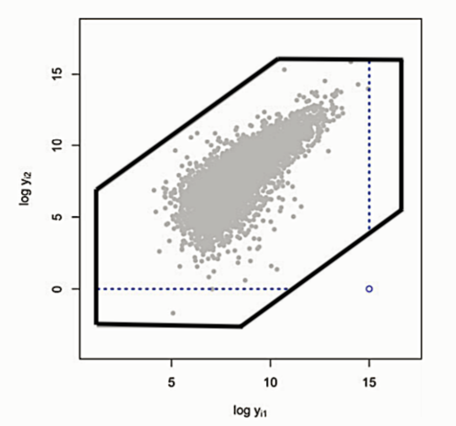
- Industry Average Ratio models:

$$Y_{ist}^{impute} = X_{ist} \left(\frac{Y_{is}}{X_{is}} \right)$$

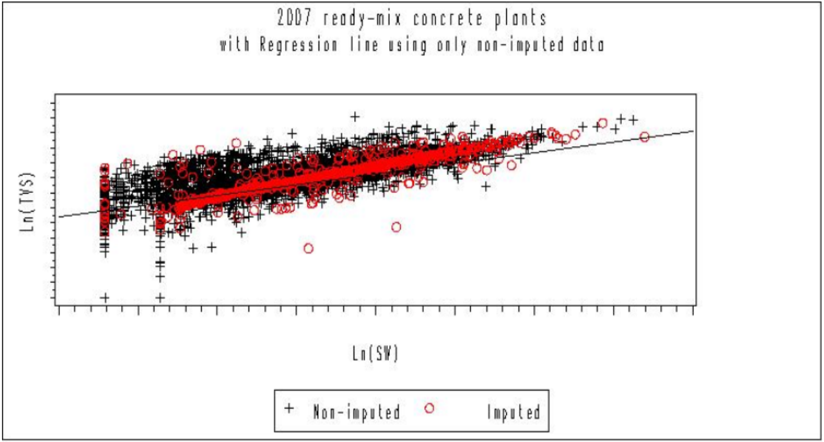
Important Types of Editing in US Census of Manufactures

- Logical edits (aka balance edits)– Example: TVS
- Units errors
- Analyst corrections
- Check against administrative records
- Ratio edits
  - based on within-industry IQRs.

Fellegi-Holt (1976): Combining Edit Rules results in Feasible Region $\mathcal{D}$



Census Bureau imputation methods are not designed for microdata research

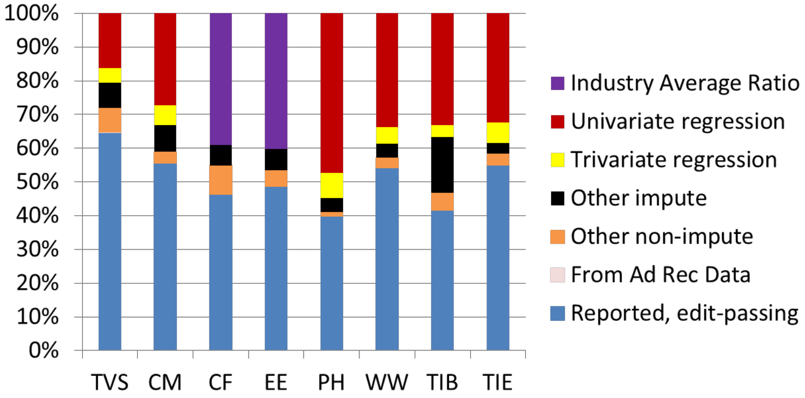


Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Frequencies of Editing/Imputation

2007 Census of Manufactures. Note: Swiss Cheese Missingness



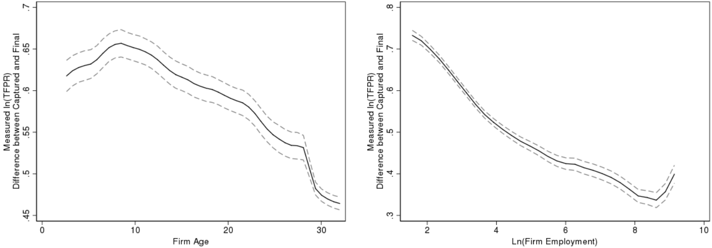
Effect of Imputation on TFPR disperion

	Captured Data			Census-Cleaned Data
	Outcome			Outcome
Year	St. Dev	90/10	75/25	St. Dev
2002	0.889			0.401
2007	0.955			0.442
2012	1.089			0.421

Effect of Imputation on TFPR disperion

	Captured Data			Census-Cleaned Data		
	Outcome			Outcome		
Year	St. Dev	90/10	75/25	St. Dev	90/10	75/25
2002	0.889	1.337	0.577	0.401	0.783	0.331
2007	0.955	1.716	0.902	0.442	0.87	0.356
2012	1.089	1.888	1.031	0.421	0.831	0.346

Who Gets Edited?



Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Quantifying effect of editing/imputation on BKR measure of Allocative Efficiency: Census (CMF)

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.00005	0.109	0.176	0.14	0.461	0.554
2007	0.000005	0.012	0.024	0.042	0.302	0.425
2012	0.00000038	0.004	0.024	0.059	0.349	0.455

- India 1% trimming: ≈ 0.387

Quantifying effect of editing/imputation on BKR measure of Allocative Efficiency: Representative Sample (ASM)

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.003	0.209	0.415	0.16	0.458	0.555
2007	0.00004	0.026	0.058	0.085	0.294	0.416
2012	0.00007	0.004	0.074	0.077	0.34	0.457

- India 1% trimming: ≈ 0.387

Cross-Country Differences in Misallocation

For Countries with Census(ish) data, using Value Added

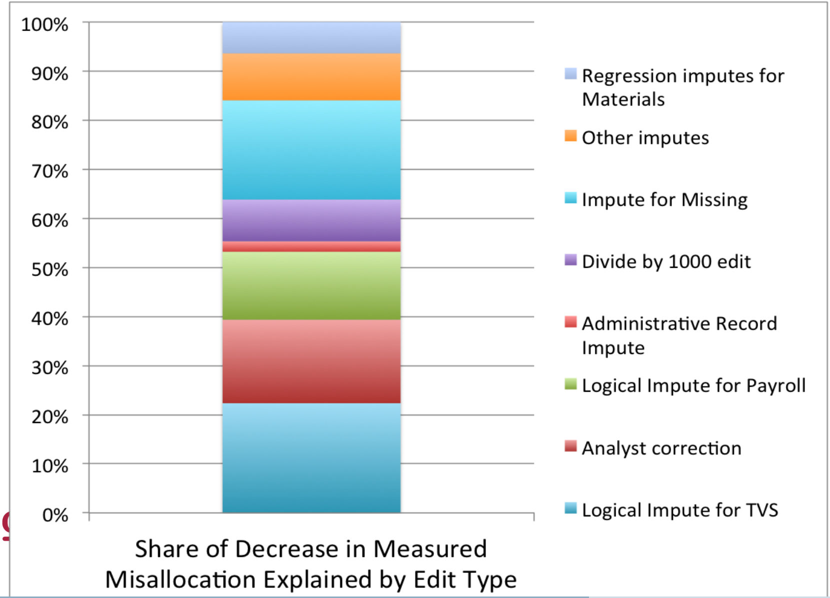
Country	Gains in Most Recent Year	Gains Relative to: Cleaned US	Raw US
India	100%	32%	-56%
Mexico	95%	32%	-57%
China	87%	26%	-59%
Chile	77%	19%	-61%
Indonesia	68%	13%	-63%
Venezuela	65%	11%	-64%
Bolivia	61%	8%	-65%
Uruguay	60%	8%	-65%
Argentina	60%	8%	-65%
Ecuador	58%	6%	-65%
Slovenia	57%	6%	-65%
El Salvador	57%	6%	-65%
Colombia	49%	1%	-67%
Brazil	41%	-5%	-69%

How should we do cross-country comparisons?

- For cross-country comparisons, we would like to use same data cleaning methods as in U.S.
  - Problem for us: U.S. Census Bureau has an entire staff cleaning the data for months
  - Can we replicate just the “important” parts of what Census Bureau does?
  - Which Census Bureau edits have big impact on measured allocative efficiency?

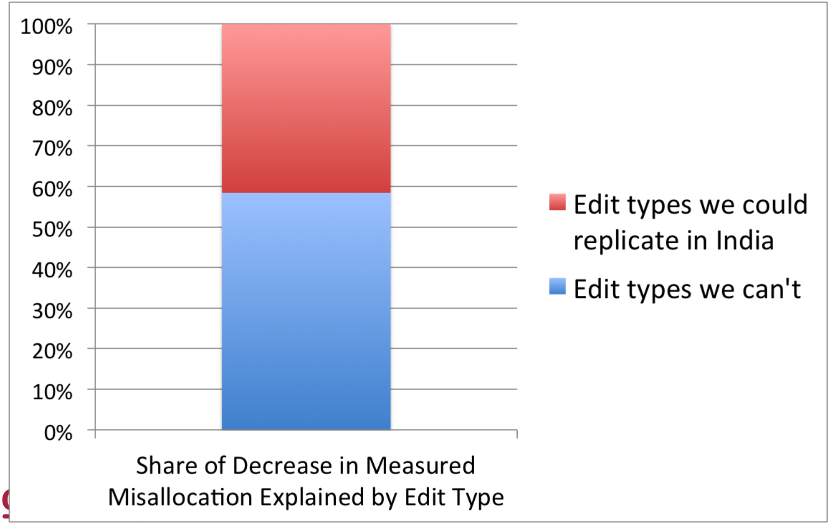
Effect of Census Bureau Edits (Shapley Shares)

on Measured Misallocation in U.S. data, 1% trimming



Effect of Census Bureau Edits (Shapley Shares)

on Measured Misallocation in U.S. data, 1% trimming



Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Motivation for doing our own data cleaning

- For cross-country comparison of misallocation want to clean firm-level data in India like the U.S. data
- Problem:
  - Not feasible for us to replicate US Census Bureau's data cleaning in India
- So...try a fully data-driven approach, following Kim et al. (2015)

Definitions

- $\boldsymbol{y_i}$ is reported firm behavior
- A_i indexes the failed ratio & balance edits
- $\boldsymbol{x_i}$ is (unobserved) the true firm behavior
- $\boldsymbol{s_i}$ is a vector of indicators for the items to be edited/imputed

Simultaneous Editing and Imputation

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

- Favor final values that are
  - Likely under the model for reporting error
  - Likely under the model for error indicators
  - Likely under the model for the underlying data

Reporting Error Model

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

Error Indicator Model

- Assume a uniform distribution for the indicators
  - So do not have weights on which variables are more likely to be reported with error - all candidates s_i that result in feasible solutions are equally likely
  - For missing items can set s_{ij} =1

True Data Model

- Each firm belongs to one of K mixture components (z)
- So need to estimate
  - probability of membership in each component (π)
  - mean vector $(\boldsymbol{\mu})$ and covariance matrix $(\boldsymbol{\Sigma})$ within each mixture

True Data Model

- Distribution of $\boldsymbol{x_i}$ conditional on μ, Σ, z_i, given feasible region $\mathcal{D}$

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

- This ensures that all of the draws will pass both the balance and ratio edits

Advantages of Method

- Imputation model approximates the joint distribution of the edit-rule-passing data
- Imputes automatically satisfy all the edit rules
- Can estimate uncertainty of misallocation estimates due to editing/imputation (although we don’t do this yet)
- Allows us to do cross-country comparisons using a common editing/imputation method

Common Data Cleaning for US and India

- Starting with raw reported data and edit rules:
  - Replace edit-rule-failing reported values with imputes from model
  - Use same model to impute for missing (item) values, which satisfy all edit rules
- We apply this method to “clean” the raw data for India and the US for every manufacturing industry

New Measures of Allocative Efficiency (1% tail trimming)

Country	Year(s)	Raw	Census	Our Cleaning
US	2002	0.109	0.461	0.499
US	2007	0.012	0.302	0.161
US	2012	0.004	0.349	0.231
India	2000-2011	0.393	n/a	0.521

Outline

1. Static Misallocation
2. Editing in the US
3. Imputation in the US
4. Effect on Measured misallocation
5. Data Cleaning
6. Wrap-Up

Conclusions

- Data cleaning done by Census Bureau has huge effect on dispersion in Census of Manufactures
  - The effect of this cleaning has increased tremendously from 2002 to 2007 to 2012
- Cross-country differences in data cleaning also may have big effect on cross-country comparisons
- For consistent cross-country comparisons, use the same data cleaning methods in both countries