Discussion of “Measuring Cross-Country Differences in Misallocation” by Rotemberg and White

Allan Collard-Wexler
Duke University

July 18, 2017 NBER Macro Productivity
When statistical agencies pull in data to produce GDP, there is a serious amount of processing that gets done before final statistics are produced.

Much of these procedures are difficult to find — these are internal protocols. Moreover, they frequently have little echo in the academic economic literature (in contrast to say computation).

Novelty in this paper: use of the U.S. Census Bureau’s pre and post imputation samples.

Integrating data cleaning and analysis in the same research unit.
Why is imputation so important in firm datasets?

- A lot of data is missing — administrative records, sampling procedures rather than the complete population, missing variables.
- As well, as the paper documents, much of the measured data is clearly wrong: thousands versus millions for instance, product line data that does not add up to total sales.
- Collard-Wexler (2011) imputations to fill out data for a dynamic game.
- The goals of the imputes done at Census might not be the one the researcher is interested in: variance of growth rates with age.
Cross-Country Analysis

From Asker, Collard-Wexler and De Loecker (2014), a lot of heterogeneity in where the data comes from.

- **US**: Manufacturing only, Census questionnaire, sent out to all non-small firms (say more than 10 employees), some data comes from tax records.
- **Mexico**: Manufacturing only, Census only for firms above 100 employees.
- **Slovenia**: data based on value added taxes.
- **World Bank Enterprise Survey**: all firms, with unclear guidance on the sampling frame.

Having worked with the ASI (India) and Census of Manufacturers (US), these are some of the best questionnaire based datasets out there.

→ Consistent data handling is a real concern here, also over time.
A hard part of working on firms is the wide distribution of firm size (Gibrat, 1931, Gabaix, 2011).
- In Korea, 2 firms account for 22% of sales.
- In the US the top 50 firms are 24% of sales.

The literature on allocation, or any output weighted statistic, very large, or very big and productive firms, play an outsized role.

Therefore, unlike say a conditional median or mean estimator, measurement error in high percentiles of TFP does not wash out for the statistics of interest.

Hsieh Klenow (2009) misallocation here is simply an exemplar of aggregate welfare statistics.
Statistical Sampling Theory with Aggregation

- Integrating data cleaning and analysis opens the door for taking sources of statistical error seriously in aggregate welfare measures.
- In Collard-Wexler and De Loecker (2015), we look at error in the estimation of a production function and its effect on decompositions.
- Need a better idea of the Data Generating Process for these errors.
Across the entire sample period, over which productivity increased by 22 percent, the plant improvement component accounted for a 9.5 percent increase in aggregate productivity (or a 43 percent share), while reallocation and net entry are responsible for the remainder. Thus, the total share of reallocation in aggregate productivity growth, including both the reallocation induced by market-share reallocation across incumbents and the net-entry process, is two-thirds.

A clear picture emerges when we move to the decomposition by technology. The main driver of productivity growth for MMs is the plant improvement component of 11.8 percent, capturing the technological change in MMs. This is suggestive of the substantial learning by doing that took place in MM production—in particular, learning how to produce higher quality steel—over the sample period. The reallocation component is negligible.

The same analysis of VI producers yields substantially different results: The plant improvement component, of 9.3 percent, is smaller than that of MMs (11.8 percent), and the net entry term of 3.8 percent is almost 17 percent of total industry productivity growth over the sample. Most noteworthy is that the reallocation term of 11.3 percent is responsible for 23 percent of industry-wide productivity growth.

In the last row of Table 8, we restate the distinct role of the net-entry process across technologies. We present the productivity premium of entrants, compared to the set of exiting plants. Across the entire sample period, VI entrants were 4.4 percent more productive than those VI plants that exited the industry. New MMs, on the other hand, entered with no specific productivity advantage.

To summarize, we find a drastic difference in the role of reallocation between technologies. The productivity growth of MMs was entirely due to common within-plant productivity growth, whereas integrated producers’ productivity growth came from the reallocation of resources across producers. In the next section, we focus on the role of reallocation among VI producers, which was a key driver of productivity growth among producers relying on the old technology, and consequently, triggered productivity growth for the industry as a whole.

### Table 8—Dynamic Decomposition of Productivity Growth (percent)

<table>
<thead>
<tr>
<th>Component</th>
<th>All</th>
<th>Minimill</th>
<th>Integrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total change</td>
<td>22.1%</td>
<td>9.6</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.28)</td>
<td>(0.49)</td>
</tr>
<tr>
<td>Plant improvement</td>
<td>9.5%</td>
<td>11.8</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.34)</td>
<td>(0.19)</td>
</tr>
<tr>
<td>Reallocation</td>
<td>9.3%</td>
<td>-0.3</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.03)</td>
<td>(0.23)</td>
</tr>
<tr>
<td>Net entry</td>
<td>3.3%</td>
<td>-2.0</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.03)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Entry-exit premium</td>
<td>0.0</td>
<td>4.4</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The share of each component in the total aggregate productivity growth is listed in parentheses. See equation (17) for definitions of various terms. For example, the share of minimill productivity growth (9.6 percent) in aggregate productivity growth is given by: 9.6/17.7 × 0.77 = 0.28—i.e., we compute the share of the minimill productivity growth in the unweighted aggregate productivity growth term, which we know from the top panel is 0.77.
1. It would be nice to have an idea of which variables, output, material, labor, capital, get mismeasured the most.

2. I am bit skeptical about the imputation algorithms: these are only as good as the variables that you think are important for your application, and do not always work well for getting all the correlations right.

3. There is clearly some data cleaning and correcting going on in India, it is just not straightforward to document.

4. I think that the paper should focus a bit more on the best practice for dealing with data cleaning and imputing, rather erroneous misallocation.
1. It would be nice to have an idea of which variables, output, material, labor, capital, get mismeasured the most.

2. I am bit skeptical about the imputation algorithms: these are only as good as the variables that you think are important for your application, and do not always work well for getting all the correlations right.

3. There is clearly some data cleaning and correcting going on in India, it is just not straightforward to document.

4. I think that the paper should focus a bit more on the best practice for dealing with data cleaning and imputing, rather erroneous misallocation.
1. It would be nice to have an idea of which variables, output, material, labor, capital, get mismeasured the most.

2. I am bit skeptical about the imputation algorithms: these are only as good as the variables that you think are important for your application, and do not always work well for getting all the correlations right.

3. There is clearly some data cleaning and correcting going on in India, it is just not straightforward to document.

4. I think that the paper should focus a bit more on the best practice for dealing with data cleaning and imputing, rather erroneous misallocation.
1. It would be nice to have an idea of which variables, output, material, labor, capital, get mismeasured the most.

2. I am bit skeptical about the imputation algorithms: these are only as good as the variables that you think are important for your application, and do not always work well for getting all the correlations right.

3. There is clearly some data cleaning and correcting going on in India, it is just not straightforward to document.

4. I think that the paper should focus a bit more on the best practice for dealing with data cleaning and imputing, rather erroneous misallocation.
Concluding Comments

- Paper is concise, very nicely written: worth 20 minutes of reading time.
- We should appreciate the effort into thinking about data cleaning and processing — just because it is technical, it does not mean it is not important.
- Need to bring in statistical sampling theory into our micro aggregation work, which will be difficult given skewed firm size distribution.