Research Paper

Approaches To Analysing Micro-Drivers Of Aggregate Productivity

Australia

2020

1351.0.55.164

APPROACHES TO ANALYSING MICRO-DRIVERS OF AGGREGATE PRODUCTIVITY

Chien-Hung Chien[1,2], A.H. Welsh[3], and Robert Breunig[4]

[1]ABS Methodology Transformation Branch
[2]ANU Mathematical Science Institute
[3]ANU Research School of Finance, Actuarial Studies & Statistics
[4]ANU Crawford School of Public Policy

1 EXECUTIVE SUMMARY

This study describes the patterns of productivity growth across eighteen industries. We examine the components of this productivity growth by estimating the contribution of entry, exit, within-firm growth and re-allocation to productivity growth in Australia in the period 2002–2013.

We use an experimental linked dataset of 10 million workers across 1.5 million firms. We produce industry-level estimates using firm-level data across 18 industries. We estimate worker- and firm-specific effects using a grouping algorithm appropriate for sparse matrices.

We find that firm entry and exit are by far the largest contributors to productivity growth across all industries. In general, firm exit contributes positively to productivity growth whereas firm entry generally contributes negatively. This would suggest that policies which facilitate firm entry and exit are likely to help achieve increased productivity gains. Policies which provide large advantages to incumbent firms are likely to detract from productivity growth.

APPROACHES TO ANALYSING MICRO-DRIVERS OF AGGREGATE PRODUCTIVITY

Chien-Hung Chien[1,2], A.H. Welsh[3], and Robert Breunig[4]

[1]ABS Methodology Transformation Branch
[2]ANU Mathematical Science Institute
[3]ANU Research School of Finance, Actuarial Studies & Statistics
[4]ANU Crawford School of Public Policy

2 ABSTRACT

We estimate the contribution of entry, exit, within-firm growth and re-allocation to productivity growth in Australia in the period 2002–2013. We use an experimental linked dataset of 10 million workers across 1.5 million firms. We produce industry-level estimates using firm-level data across 18 industries. We estimate worker- and firm-specific effects using a grouping algorithm appropriate for sparse matrices. Firm entry and exit are by far the largest contributors to productivity growth across all industries. In general, firm exit contributes positively to productivity growth whereas firm entry generally contributes negatively.

Contents

3 INTRODUCTION

Firms are living beings; they are born, some grow to maturity, and all eventually die. Child mortality is high, the few who survive grow rapidly, but only a handful enjoy old age. New entrants are smaller and less productive on average but more diverse than continuing firms. Although size and productivity diversity diminish over time due to the selection process responsible for early death, differences in factor productivity of those that continue are still large and persistent.

Lentz and Mortensen (2010, p.2)

As Lentz and Mortensen (2010) point out, an efficient market allocates resources from less productive firms to more productive ones. Firm dynamics—that is, how contributions from established, entering and exiting firms affect aggregate productivity—is one of the key microdrivers that influence aggregate productivity (Foster et al., 2001).

The seminal surveys by Bartelsman and Doms (2000) and Syverson (2011) discuss the advantages of using microdata to better understand the determinants of aggregate productivity. Aggregate statistics, which give a good overview of trends in productivity growth, do not show the variability that occurs at micro levels. It is important to develop a good understanding of the degree to which different aspects of productivity growth within and across firms contribute to different productivity growth across industries.

This study describes the patterns of productivity growth across eighteen industries. We examine the components of this productivity growth by looking at firm entry and exit, reallocation across continuing firms and productivity growth within firms. We also examine whether these patterns differ across industries?

Our industry level results are decomposed into contributions from surviving, entering and exiting firms. We apply linear models, estimated separately by industry, using a Cobb Douglas production function as the basis to estimate firm level productivity. Previous studies have shown the importance of correcting for endogeneity in estimating productivity due to strong correlation between inputs and outputs in the production process. We adapt the approaches of Abowd et al. (2002) and Mare et al. (2017) to estimate labour inputs which we use to address endogeneity.

This paper is structured as follows: Section 4 provides the literature review, Section 5 describes the scope of the data and Section 6 presents the statistical models. Section 7 discusses estimation methods and how we ensure unique identification of the estimated indicator variables. Section 8 contains empirical results. The final section gives some conclusions and future directions for further research.

4 LITERATURE REVIEW

Developing a good understanding of the determinants of aggregate productivity is challenging because the economy is complex. One factor in aggregate productivity growth is the reallocation of resources from more productive firms to less productive ones. Part of this effect is captured by firm entry and exit. Several studies describe the role of the reallocation of resources between firms. Influential work by Olley and Pakes (1996) and Bartelsman and Dhrymes (1998) developed the principal methods that most economists use to measure the impact of firm dynamics on aggregate productivity. These methods are often used in analyses to better understand the process of creative destruction that can occur within and between sectors of the economy (Foster et al., 2001).

Lafrance and Baldwin (2011) explored the contribution firm turnover has on productivity growth in the Canadian services industries. They found that the market naturally allocates resources from uncompetitive firms to new entrants. Nguyen and Hansell (2014) explored the firm dynamic effects on productivity growth for Australian manufacturing and business services industries. They have found that entering and exiting firms make smaller contributions to overall productivity than established firms.

Economists also consider productivity differences to come from better measures of inputs used in the production process. Labour economists have observed strong correlations between the differences in firm productivity and wage costs per worker (Lentz and Mortensen, 2010). However, this strong correlation can potentially cause endogeneity (Fox and Smeets, 2011). Better labour quality measures for production are important to minimise endogeneity in productivity analysis (Foster et al., 2001).

This study explores the effects of firm dynamics on aggregate productivity by adapting approach of Mare et al. (2017). The labour component is estimated using the approach of Abowd et al. (2002) which takes into account two-sided worker and firm effects. This estimated labour component is then used in a firm production function equation. The contributions to the aggregate industry productivity are derived using the approaches of Griliches and Regev (1995) and Melitz and Polanec (2015) to take into account firm dynamics.

5 DATA DESCRIPTION

The Australian Taxation Office (ATO), Australia Business Register (ABR) and ABS datasets are held in both the Business Longitudinal Analytical Data Environment (BLADE) for firms (ABS and DIIS, 2017) and the prototype Graphically Linked Information Discovery Environment for workers (Chien and Mayer, 2015a). This section describes the ABS confidentiality protocol and the data processing carried out for this study. The sample period is between 2002–03 to 2012–13.

5.1 Data confidentiality

The ATO data is provided to the Australian Statistician under the Taxation Administration Act 1953 and (ABR) data is supplied to the Australian Statistician under A New Tax System (Australian Business Number) Act 1999. These Acts require that these data are only used by the ABS for administering the Census and Statistics Act 1905. The ABS is obliged to maintain the confidentiality of individuals and businesses in these ATO and ABR datasets, as well as comply with provisions that govern the use and release of this information, including the Privacy Act 1988 ABS (2015).

This study uses a strict access control protocol. Access to the datasets includes audit trails and is limited on a need to know basis. All ABS officers are legally bound to secrecy under the Census and Statistics Act 1905. Officers sign an undertaking of fidelity and secrecy to ensure that they are aware of their responsibilities. The ABS policies and guidelines govern the disclosure of information to maintain the confidentiality of individuals and organisations. This study presents only aggregate results to ensure that they are not likely to enable identification of a worker or a firm.

5.2 Data processing

Our experimental worker panel uses data from the ABS prototype Graphically Linked Information Discovery Environment (GLIDE) (Chien and Mayer, 2015a). The worker panel has $130,281,096$ observations containing $1,903,015$ Australian Business Numbers (ABNs) for firms and $13,131,074$ de-identified and encoded Tax File Numbers (DETFNs) for workers. We only include workers whose age is between $(16, 65]$ in the years between 2001–02 and 2012–13. Worker characteristics such as age, sex and occupation come from Personal Income Tax (PIT) filings and wage information comes from Pay-as-You-Go (PAYG) summaries. PAYG contains a longer time series than PIT, so this study backcasts the PIT data to the same length. The earliest available PIT information is used to backcast sex (holding it constant) and age (by subtracting 1 year). Two methods to backcast the skill categories for workers were explored: either using the average or holding it constant for each worker. This study found that it was not appropriate to use average skill because workers tend to become more skilled over time, so using average skill inflates the worker's skill level over the backcast period. The ABS's Australian and New Zealand Standard Classification of Occupations is used to convert occupations into a 5-point skills categorical variable for the analysis (ABS, 2009). We stress that the prototype worker panel data is constructed for research purposes only.

The experimental worker panel is aggregated to the firm level to derive worker-level variables for each ABN. The employee counts come from PAYG records, which contain both ABNs and DETFNs. We count the total number of DETFNs for each ABN in each year. One key challenge in analysing integrated administrative datasets is the different coverage of the firms. For example, at the firm level, there is only a 6% difference in scope between PAYG and PIT files between 2001–02 and 2012–13. This difference in scope can be caused by the timing of the processing or including only working age workers.

The information on firm industry classification comes from both the ATO and the ABS Business register (ABSBR). A majority of firms have valid industry classification. The industry classifications for these ABNs do not change over the sample period between 2001–02 and 2012–13. We impute industry classifications for 45,961 ABNs using the method discussed in Section *A*. We ensure that the imputed industry classifications also do not change for the re-entered firms (e.g., firms drop in and out due to processing errors or late processing). This is important to minimise bias in decomposition analysis at the industry level. We also use the following heuristic rules for the data processing. First, 335 ABNs have missing or invalid year of incorporation variable and a majority of these firms are in 2001–02. We assume that these firms are incorporated in the year when the ABN was first introduced in 2000–01. Secondly, the information on firm entry and exit is from both ABS ABSBR for ABNs in year 2001–02 only and our derivation for ABNs between 2002–03 and 2012–13. We do not classify re-entered firms as entry firms or exit firms during the missing spells. For example, firm A has observations for 2001–02, 2003–04, 2004–05 and 2005–06. Firm A is classified as an entry firm in 2001–02, a continuing firm in 2003–04 and 2004–05 and an exit firm in 2005–06. We also use SAS `proc expand` procedure to longitudinally interpolate multifactor productivity and industry weights for these re-entering firms.

Figure 11 in Appendix *F* shows the firm entry and exit rates over the sample period. It is interesting to note that the exit rates were generally lower between 2002–03 and 2009–10 except for the finance industry in 2007–08. The higher firm exit rates in the finance industry could have been caused by the global financial crisis.

5.3 Data linking and summary

The study uses a similar linking strategy to ABS (2015) and Chien and Mayer (2015b) to assemble the developmental firm panel using an experimental BLADE. The firm records were deterministically linked using ABNs and worker records were deterministically linked using DETFNs. As the linking variable is encrypted, it is not possible to identify individuals in the datasets. The experimental BLADE contains firm data sourced from the ATO, the ABR and the ABS. The sample period is from 2002–03 to 2012–13. BLADE contains detailed firm characteristics data from Business Income Tax (BIT) , Business Activity Statements (BAS) and the ABR (Hansell and Rafi, 2018). The experimental firm panel has 43,191,403 observations with 6,846,067 ABNs in the sample period between 2001–02 and 2012–13. The firm panel contains non-employing firms. Most firm-level variables such as firm sales and materials costs etc. come from Business Income Tax or Business Activities Statements. We include all firms with valid records. This study uses experimental version of BLADE and therefore statistical issues discussed here may not exist in the production version of BLADE. Our experimental BLADE contains firm characteristics data. It does not contain any data about worker characteristics beyond the number of employees and total wages.

Tables 16 and 17 in Appendix E show the summary statistics for the firm and worker panels. The summary statistics for the worker panel show that we focus on workers who have at least 2 years in the labour market (aged 17 years and older) and are at most 65 years old. The proportion of workers aged between 15 and 16 years is small. The summary statistics for the firm panel are broadly consistent with balanced and imputed datasets. This is in line with our observation with the correlation analysis in Tables 2 and 3 below. The summary statistics for the firm panel show that, in the sample, the youngest firm is 1 year old and the 99^{th} percentile is around 19 years old. It is interesting to note that at the 99^{th} percentile, the real WAGES costs is higher than the logarithm of estimated labour components in real terms. Table 1 shows firm sizes and firm years in sample. Large firms, i.e., employee size ≥ 200, are more likely to be in the sample for a longer period.

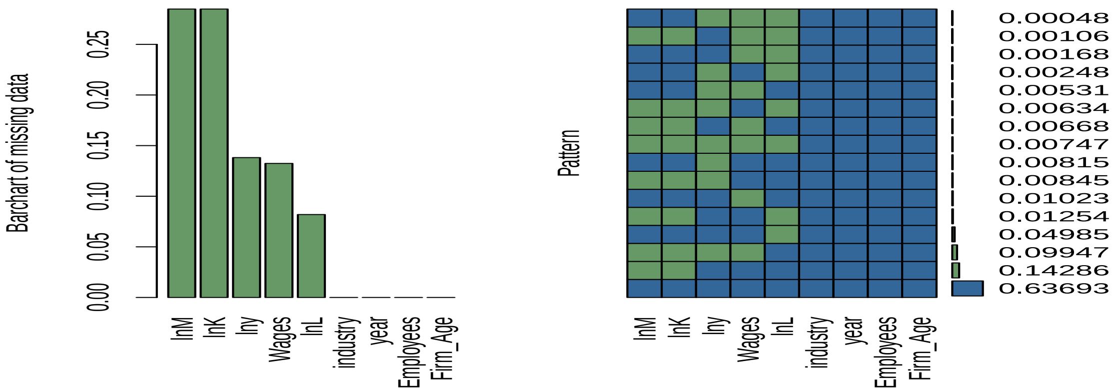
Table 1: Firm size and years in sample

	years in sample												
Firm size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
1 to 4	3.9	4.9	5.7	7.1	5.2	5.3	5.1	4.7	4.8	5.0	6.0	8.7	66.3
5 to 19	0.2	0.5	0.9	1.8	1.4	1.6	1.7	1.7	1.8	2.1	3.2	7.8	24.7
20 to 199	0.0	0.1	0.1	0.5	0.4	0.4	0.5	0.5	0.5	0.7	1.0	3.6	8.3
200 plus	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.7
Total	4.1	5.5	6.8	9.5	7.0	7.3	7.3	6.8	7.1	7.8	10.3	20.6	100.0

5.4 Missing data

We use an unbalanced panel of firms. Figure 1 shows the missing data pattern for the ABS experimental data used in the sample.

Figure 1: Missing data patterns in experimental datasets for all firms with ≥ 1 employee



Note. In each subfigure, the left panel is a bar chart showing the proportion of missing data for each variable. The right panel shows the missing data patterns and the proportion of each pattern; a green tile indicates missing data; a blue tile indicates non-missing data. The left panel is a bar chart showing the proportion of missing data for each variable. The right panel shows the missing data patterns in the data and the proportion of each pattern. These proportions are scaled to increase the readability of the plot (Templ et al., 2012). The variables ln$L$, ln$K$, ln$y$, ln$M$ and ln$Firm_Age$ are the logarithms of labour for firms, capital for firms, sales for firms, materials used for production and firm age, respectively. The number of employees is $Employees$ and the industry classification for firms is $Industry$.

Dropping firms where some variables are missing results in a dramatic reduction in sample size. Therefore we assume missing at random and impute missing variables for non-missing firms using a sequential regression approach in SAS, namely the `proc mi` procedure (see Appendix *B*). We create 10 imputed data sets upon which we base our estimation. We then reproduce this analysis 10 times and we select the results which maximise the likelihood function for the firm-level productivity model (3) below from the 10 imputations. Tables 2 and 3 compare the correlation coefficients for the variables of interest. These tables show that the correlation coefficients are consistent when we compare between complete cases and imputed datasets.

Table 2: Pearson Correlation Coefficients -- balanced dataset

	$\ln y_{jkt}^{(firm)}$	$\ln K_{jkt}$	$\ln M_{jkt}$	lnFirm_Age	$\ln \hat{z}_t^{(jk)}$	WAGES
$\ln y_{jkt}^{(firm)}$	1	0.5206	0.58389	0.13821	0.01096	0.72011
		<.0001	<.0001	<.0001	<.0001	<.0001
$\ln K_{jkt}$	0.5206	1	0.4829	0.10082	-0.12324	0.50686
	<.0001		<.0001	<.0001	<.0001	<.0001
$\ln M_{jkt}$	0.58389	0.4829	1	0.12111	0.02614	0.62887
	<.0001	<.0001		<.0001	<.0001	<.0001
lnFirm_Age	0.13821	0.10082	0.12111	1	-0.42234	0.15778
	<.0001	<.0001	<.0001		<.0001	<.0001
$\ln \hat{z}_t^{(jk)}$	0.01096	-0.12324	0.02614	-0.42234	1	-0.0151
	<.0001	<.0001	<.0001	<.0001		<.0001
WAGES	0.72011	0.50686	0.62887	0.15778	-0.01508	1
	<.0001	<.0001	<.0001	<.0001	<.0001	

Table 3: Pearson Correlation Coefficients -- imputed dataset

	$\ln y_{jkt}^{(firm)}$	$\ln K_{jkt}$	$\ln M_{jkt}$	lnFirm_Age	$\ln \hat{z}_t^{(jk)}$	WAGES
$\ln y_{jkt}^{(firm)}$	1	0.53644	0.55506	0.13707	0.0238	0.7391
		<.0001	<.0001	<.0001	<.0001	<.0001
$\ln K_{jkt}$	0.53644	1	0.4823	0.10351	-0.11546	0.5301
	<.0001		<.0001	<.0001	<.0001	<.0001
$\ln M_{jkt}$	0.55506	0.4823	1	0.07288	0.0215	0.5616
	<.0001	<.0001		<.0001	<.0001	<.0001
lnFirm_Age	0.13707	0.10351	0.07288	1	-0.42192	0.137
	<.0001	<.0001	<.0001		<.0001	<.0001
$\ln \hat{z}_t^{(jk)}$	0.0238	-0.11546	0.0215	-0.42192	1	0.0176
	<.0001	<.0001	<.0001	<.0001		<.0001
WAGES	0.73906	0.53006	0.56157	0.137	0.01758	1
	<.0001	<.0001	<.0001	<.0001	<.0001	

Results from our imputation approach match ABS results more closely than those where we drop all firms with any missing values. The analysis of the complete case data, which involves dropping 80 per cent of the data, produces a lot of volatility and inconsistency with ABS results therefore we prefer the imputation approach.

6 STATISTICAL MODELS

6.1 Worker equation

This analysis uses a modified wage equation adapted from Abowd et al. (2002). The worker panel is unbalanced, meaning that the available observations for each worker i, $i = 1, \dots, N$ can be different. Suppose that the observations for worker i are available at time $t = 1, \dots, T_i$. So $t = 1$ is the first time period and $t = T_i$ is the last time period for the available observations for worker i. Note that there can be gaps. A worker might appear in periods 1 and 3 but not in period 2, for example. We model y_{it}, the wages for worker i at time t, as

$$\ln(y_{it}) = \mathbf{x}_{it}^{\top}\boldsymbol{\alpha} + \theta_i + \mathbf{f}_{it}^{\top}\psi + \epsilon_{it}, \tag{1}$$

where $\mathbf{x}_{it}$ is a p-vector of characteristics of worker i at time t, $\boldsymbol{\alpha}$ is a p-vector of unknown coefficients of the worker characteristics, θ_i represents unobserved (time-invariant) worker effects, the components of the J-vector $\psi = (\psi_1, \cdots, \psi_J)^{\top}$ represent firm effects (e.g. specific factors such as pay structure that affect workers' wages), $\mathbf{f}_{it}^{\top} = (f_{i1t}, \cdots, f_{iJt})^{\top}$ is a firm indicator vector with components

$$f_{ijt} = \begin{cases} 1, & \text{if worker } i \text{ works for firm } j \text{ at time } t \\ 0, & \text{otherwise,} \end{cases}$$

and the random disturbances ϵ_{it} are assumed to satisfy $\epsilon_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

It is convenient to write the term $\mathbf{x}_{it}^{\top}$ which describes worker characteristics in Wilkinson and Rogers (1973) notation as $Sex + HighSkill + MediumSkill + WorkingSkill + Time + Poly(Age, 4) + Sex : Poly(Age, 4) + Sex : Time$. Here the indicator $Sex = 1$ if worker i is male and 0 otherwise. The indicator $HighSkill = 1$ if worker i has a tertiary qualification and 0 otherwise. The indicator $MediumSkill = 1$ if worker i has at most a diploma qualification and 0 otherwise. The indicator $WorkingSkill = 1$ if worker i has at most a certificate III qualification and 0 otherwise. Workers with qualifications lower than a certificate III qualification are treated as the baseline and included in the intercept. The variable $Time$ is represented by 11 time indicator variables, one for each year with $2001-02$ as baseline. The variable Age, the age of worker i at time t_{it}, is fitted by a quartic polynomial including linear, quadratic, cubic and quartic functions. We include a quartic function to better describe the data because fitting only quadratic and cubic terms does not describe the decline in workers' wage as they get older. We include the interaction terms $Sex : Poly(Age, 4)$ between Sex and Age and $Sex : Time$ between Sex and $Time$. This makes each $\mathbf{x}_{it}^{\top}\boldsymbol{\alpha}$ a sum of $p = 34$ terms.

Following Mare et al. (2017), we estimate (1), pooling across all workers at all time periods in all industries. We then derive an instrument for firm-specific labour inputs, which we use in (3) below, based upon the average fitted values for each firm j. Specifically, let $\widehat{\boldsymbol{\alpha}}$, $\widehat{\theta}_1, \dots \widehat{\theta}_N$ and

$\widehat{\psi}$ denote estimates of parameters in (1). We use a two stage least square approach to derive our instrumental variable (Wooldridge, 2006). The estimated person and firm effects from (1) are correlated with firm productivity because firms with higher quality workers or better management practices are likely to be more productive. The equation to derive the proposed instrumental variable is:

$$y_t^{(j)} = \mathbf{x}_t^{(j)}\widehat{\boldsymbol{\alpha}} + \varepsilon_{kt}^{(j)}, \text{ where} \tag{2}$$

where $\quad y_t^{(j)} = \sum_{i=1}^{N} f_{ijt}y_{it} \quad$ and $\quad \mathbf{x}_t^{(j)} = \sum_{i=1}^{N} f_{ijt}\mathbf{x}_{it},$

Note that the variables in (2) now have a firm superscript, j, to reflect the averaging of worker effects within each firm j. We use $\hat{z}_t^{(j)}$ to denote the instrumental variable derived as the predicted value from (2). When we want to emphasise below that firm j belongs to industry k, we also include the industry superscript k so that the estimated firm-average worker effect (the average effect of a worker in each firm) $\hat{z}_t^{(j)}$ from (2) becomes $\hat{z}_t^{(jk)}$.

6.2 Firm level productivity model

The firm volume outputs can be modelled as functions of the observed inputs such as capital, materials and labour in volume terms, and unobserved components in the production process (Fox and Smeets, 2011). We use a Cobb–Douglas production function, similar to Breunig and Wong (2008) and Mare et al. (2017), to model $y_{jkt}^{(firm)}$, the outputs, (i.e., sales adjusted for repurchase of stock) deflated by industry gross value added implicit price deflators by firm j in industry k at time t (ABS, 2018a), as:

$$\begin{aligned} \ln y_{jkt}^{(firm)} = \beta_k + \beta_{1k}\ln L_{jkt} + \beta_{2k}\ln K_{jkt} + \beta_{3k}\ln M_{jkt} + \\ \beta_{4k}\ln Firm\_Age_{jkt} + \tau_{kt} + \varepsilon_{jkt}, \end{aligned} \tag{3}$$

where $\ln L_{jkt}$ is the logarithm of labour inputs deflated by *Wage Price Index* and $\ln K_{jkt}$ is the logarithm of the cost of capital, which includes depreciation, capital rental expenses and capital work deductions, deflated by the industry consumption of fixed capital implicit price deflators (ABS, 2018a). The logarithm of material costs $\ln M_{jkt}$ is the inputs used in the production, deflated by *Producer Price Indexes: Intermediate Goods* (ABS, 2018b; also see the 'List of Symbols and Variables' for information). The logarithm of firm age is $\ln Firm\_Age_{jkt}$. We also include different intercepts β_k for each industry and time-fixed effects τ_{kt}. The multifactor productivity term ε_{jkt} is assumed to satisfy $\varepsilon_{jkt} \overset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$ to estimate unbiased coefficients for the Cobb–Douglas production function Zellner et al. (1966).

Endogeneity causes bias in estimating the production function (3). To mitigate the bias, many studies use predicted values from instrumental variables equations—that is, using lagged inputs as instruments for the current inputs (Gandhi et al., 2011). For example, Olley and Pakes (1996) and Breunig and Wong (2008) use lagged capital investment, Levinsohn and Petrin (2003) and Bakhtiari (2015) use lagged material inputs, and Fox and Smeets (2011) use lagged wage costs as instrumental variables. However, Reed (2015) cautions against the use of lagged instrumental variables to correct for the simultaneity bias. Our labour component comes from the estimated instrument, $\hat{z}_t^{(jk)}$, from (2). We remove components in (1) that can correlate with firm multifactor productivity ε_{jkt} in (3).

We fit separate models for each industry, we include k to emphasise the nesting of firm j in industry k. This specification restricts the same production technology within industries but varies across industries to allow each firm to have an individual productivity component. The model fitted to the data is:

$$\begin{aligned}\ln y_{jkt}^{(firm)} = \beta_k + \beta_{1k}\hat{z}_t^{(jk)} + \beta_{2k}\ln K_{jkt} + \beta_{3k}\ln M_{jkt} + \\ \beta_{4k}\ln Firm\_Age_{jkt} + \tau_{kt} + \varepsilon_{jkt}.\end{aligned} \tag{4}$$

The estimated parameters for (4) include the industry intercepts $\hat{\beta}_k$, labour inputs $\hat{\beta}_{1k}$, cost of capital $\hat{\beta}_{2k}$, materials costs $\hat{\beta}_{3k}$, firm age $\hat{\beta}_{4k}$, time-fixed effects $\hat{\tau}_{kt}$ and multifactor productivity:

$$\hat{\varepsilon}_{jkt} = \ln y_{jkt}^{(firm)} - (\hat{\beta}_k + \hat{\beta}_{1k}\hat{z}_t^{(jk)} + \hat{\beta}_{2k}\ln K_{jkt} + \hat{\beta}_{3k}\ln M_{jkt} + \hat{\beta}_{4k}\ln Firm\_Age_{jkt} + \hat{\tau}_{kt}).$$

Firm productivity is the ratio of output to measured inputs normalised relative to industry k mean. The firm productivity is defined as:

$$M\hat{f}p_{jkt} = \hat{\tau}_{kt} + \hat{\varepsilon}_{jkt}. \tag{5}$$

We specify the pooled production function regression to calculate industry weights and aggregate the contributions of firms to industry multi-factor productivity growth as

$$\begin{aligned}\ln y_{jkt}^{(firm)} = \beta_1\hat{z}_t^{(jk)} + \beta_2\ln K_{jkt} + \beta_3\ln M_{jkt} + \\ \beta_4\ln Firm\_Age_{jkt} + \lambda_k + \tau_t + \varepsilon_{jkt}^{(pooled)},\end{aligned} \tag{6}$$

where the industry and year effects (λ_k and τ_t) are estimated as fixed effects. We use (6) to define industry weights $\widehat{w}_{jkt}$ as:

$$\widehat{w}_{jkt} = \hat{\beta}_1\hat{z}_t^{(jk)} + \hat{\beta}_2\ln K_{jkt} + \hat{\beta}_3\ln M_{jkt} + \hat{\beta}_4\ln Firm\_Age_{jkt}. \tag{7}$$

6.3 Industry productivity

We follow Mare et al. (2017) and define the aggregate productivity index A_{kt} for an industry k at time t as:

$$A_{kt} = \widehat{w}^{*}_{jkt} M\hat{f}p_{jkt}, \tag{8}$$

and J_{kt} is the number of firms in industry k at time t. The multi-factor productivity term $M\hat{f}p_{jkt}$ and industry weights $\widehat{w}_{jkt}$ are defined in the previous section. Note that the weights $\widehat{w}^{*}_{jkt}$ satisfy $\sum_{j=1}^{J_{kt}} \widehat{w}^{*}_{jkt} = 1$ for each industry k and time t.

Next, aggregating to industry level, let $\widehat{w}_{kt} = \sum_{j=1}^{J_{kt}} \widehat{w}_{jkt}$ and $M\hat{f}p_{kt} = \sum_{j=1}^{J_{kt}} M\hat{f}p_{jkt}$. Then the aggregate productivity index A_t, for all industries at time t, is:

$$A_t = \widehat{w}^{**}_{kt} M\hat{f}p_{kt}, \tag{9}$$

where $\widehat{w}^{**}_{kt} = \dfrac{\widehat{w}_{kt}}{\sum_{k=1}^{K_t} \widehat{w}_{kt}}$,

and K_t is the number of industries at time t. Note the weights $\widehat{w}^{**}_{kt}$ satisfy $\sum_{k=1}^{K_t} \widehat{w}^{**}_{kt} = 1$ for each time t.

Griliches and Regev (1995) propose decomposing the changes in aggregate productivity from time $t-1$ to t into contributions from surviving (S), entering (EN) and exiting (EX) firms as:

$$\Delta A_{kt} = W_{kt} + B_{kt} + EN_{kt} + EX_{kt}, \tag{10}$$

where
$$\begin{aligned} W_{kt} &= \sum_{j \in S_{kt}} \overline{\widehat{w}}_{jk} \Delta M\hat{f}p_{jkt}, \\ B_{kt} &= \sum_{j \in S_{kt}} \Delta \widehat{w}_{jkt} (\overline{M\hat{f}p}_{jk} - \overline{A}_k), \\ EN_{kt} &= \sum_{j \in EN_{kt}} \widehat{w}_{jkt} (M\hat{f}p_{jk} - \overline{A}_k) \text{ and} \\ EX_{kt} &= \sum_{j \in EX_{kt}} \widehat{w}_{jkt-1} (M\hat{f}p_{jkt-1} - \overline{A}_k). \end{aligned}$$

The symbol Δ represents changes, so $\Delta A_{kt} = A_{kt} - A_{kt-1}$ is the change in aggregate productivity for industry k from time $t-1$ to time t. Bars represent averages between t and $t-1$, so $\overline{\widehat{w}}_{jk} = \frac{(\widehat{w}_{jkt} + \widehat{w}_{jkt-1})}{2}$ and $\overline{A}_k = \frac{(A_{kt} + A_{kt-1})}{2}$. The definitions of surviving $j \in S_{kt}$, entering $j \in EN_{kt}$ and exiting firms $j \in EX_{kt}$ are based on firm transitions on an annual basis over the observed sample period. Survivors are firms operating in t and $t-1$, exiting firms are firms that exist at time $t-1$ but not at time t and entering firms are firms that did not exist at time $t-1$ but did at time t. The contribution of the surviving firms is decomposed into two components: the within-industry reallocation W_{kt}, which measures the change in firm productivity weighted by the average of the weights at t and $t-1$ (i.e., $\overline{\widehat{w}}_{jk}$) and the between-industry reallocation B_{kt}, which measures deviations from the average productivity (i.e., $\overline{A}_k$) including the impact of firm entry and exit (Foster et al., 2001).

Fox and Smeets (2011) discuss the importance of using appropriate benchmarks to calculate the contributions of surviving, entering and exiting firms to aggregate productivity. This study uses an alternative method proposed by Melitz and Polanec (2015). Their dynamic Olley–Pakes decomposition incorporates a decomposition proposed by Olley and Pakes (1996), which captures the covariance of productivity changes and market share of an individual firm over time. Let J denote J_{kt} (i.e., the number of firms in industry k at time t) and the equation is:

$$\begin{aligned} A^*_{kt} &= \overline{M\hat{f}p}_{kt} + \frac{1}{(J-1)}\sum_{j=1}^{J}(\widehat{w}_{jkt} - \overline{\widehat{w}}_{kt})(M\hat{f}p_{jkt} - \overline{M\hat{f}p}_{kt}) \\ &= \overline{M\hat{f}p}_{kt} + Cov(\widehat{w}_{jkt}, M\hat{f}p_{jkt}), \end{aligned} \tag{11}$$

where $\overline{M\hat{f}p}_{kt} = \dfrac{\sum_j I_{jkt} M\hat{f}p_{jkt}}{\sum_{j=1}^{N} I_{jkt}}$ and $\overline{\widehat{w}}_{kt} = \dfrac{\sum_j I_{jkt}\widehat{w}_{jkt}}{\sum_{j=1}^{N} I_{jkt}}$, with

$$I_{jkt} = \begin{cases} 1, & \text{if firm } j \text{ operates in industry } k \text{ at time } t \\ 0, & \text{otherwise.} \end{cases}$$

The dynamic Olley–Pakes approach decomposes aggregate productivity into contributions from surviving, entering and exiting firms as:

$$\Delta A^*_{kt} = W^*_{kt} + B^*_{kt} + EN^*_{kt} + EX^*_{kt}, \tag{12}$$

where $W^*_{kt} = \Delta\overline{P}_{kt}$, $B^*_{kt} = \Delta Cov_{kt}$,

$$EN^*_{kt} = \sum_{j \in EN} \widehat{w}_{jkt}(A_{jkt\in EN} - A_{jkt\in S}) \text{ and}$$

$$EX^*_{kt} = \sum_{j \in EX} \widehat{w}_{jkt}(A_{jkt\in EX} - A_{jkt-1\in S}).$$

The dynamic Olley–Pakes decomposition approach uses more appropriate benchmarks for the entering and exiting firms (see the discussion in Section 8). For example, entering firms only generate positive growth when they have higher productivity than surviving firms at time t. Similarly, exiting firms can only generate a positive contribution if they have lower productivity than surviving firms at time $t-1$.

7 ESTIMATION METHODS

7.1 Data structure

The model in (1) can be written as a model for each worker i by stacking the observations over time. We obtain

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + 1_i\theta_i + \mathbf{F}_i\psi + \boldsymbol{\epsilon}_i, \tag{13}$$

where $\underset{T_i\times 1}{\mathbf{y}_i = \begin{bmatrix} y_{it_{i1}} \\ \vdots \\ y_{it_{iT_i}} \end{bmatrix}}$, $\mathbf{X}_i = \underset{T_i\times 34}{\begin{bmatrix} \mathbf{x}_{it_{i1}}^{\top} \\ \vdots \\ \mathbf{x}_{it_{iT_i}}^{\top} \end{bmatrix}}$, $1_i = \underset{T_i\times 1}{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}$ $\mathbf{F}_i = \underset{T_i\times J}{\begin{bmatrix} \mathbf{f}_{it_{i1}}^{\top} \\ \vdots \\ \mathbf{f}_{it_{iT_i}}^{\top} \end{bmatrix}}$, $\boldsymbol{\epsilon}_i = \underset{T_i\times 1}{\begin{bmatrix} \epsilon_{it_{i1}} \\ \vdots \\ \epsilon_{it_{iT_i}} \end{bmatrix}}$.

The model for the whole sample can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{P}\theta + \mathbf{F}\psi + \boldsymbol{\epsilon}, \tag{14}$$

where $\mathbf{y} = \underset{N^*\times 1}{\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}}$, $\mathbf{X} = \underset{N^*\times p}{\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}}$, $\mathbf{P} = \underset{N^*\times N}{\begin{bmatrix} 1_1 & & 0 \\ & \ddots & \\ 0 & & 1_N \end{bmatrix}}$, $\theta = \underset{N\times 1}{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}}$, $\mathbf{F} = \underset{N^*\times J}{\begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_N \end{bmatrix}}$,

$\boldsymbol{\epsilon} = \underset{N^*\times 1}{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}}$ and $N^* = \sum_{i=1}^{N} T_i$ is the total number of observations.

7.2 Preconditioned conjugate gradient algorithm

Abowd et al. (1999) highlighted the challenges of fitting model (1) due to the large number of workers and firms. The US study contains $N > 1$ million workers and $J > 50,000$ firms. The Australian prototype dataset contains $N > 10$ million workers and $J > 1.5$ million firms for around 130 million observations over eleven years for the worker equation (1). This study uses the direct estimation methodology proposed by Abowd et al. (2002) which involves first solving a large sparse linear system with a preconditioned conjugate gradient algorithm, and then imposing constraints on the parameters to identify unique worker and firm effects. The conjugate gradient algorithm solves the sparse linear system $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{A}$ is a symmetric positive definite matrix, $\boldsymbol{\beta}$ is an unknown vector and $\mathbf{c}$ is a known vector. For ordinary least square estimation of parameters in (1), the system is defined with

$$\mathbf{A} = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{P} & \mathbf{X}^\top\mathbf{F} \\ \mathbf{P}^\top\mathbf{X} & \mathbf{P}^\top\mathbf{P} & \mathbf{P}^\top\mathbf{F} \\ \mathbf{F}^\top\mathbf{X} & \mathbf{F}^\top\mathbf{P} & \mathbf{F}^\top\mathbf{F} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \theta \\ \psi \end{bmatrix} \text{ and } \mathbf{c} = \begin{bmatrix} \mathbf{X}^\top\mathbf{y} \\ \mathbf{P}^\top\mathbf{y} \\ \mathbf{F}^\top\mathbf{y} \end{bmatrix}. \tag{15}$$

Since $\mathbf{A}$ is a large, sparse matrix, iterative methods like the conjugate gradient algorithm perform better if we transform $\mathbf{A}$ to improve its condition number (Shewchuk, 1994). There are many options for creating a preconditioning matrix, including incomplete Cholesky factorisation or diagonal preconditioning which uses a diagonal matrix whose diagonal entries are identical to the diagional elements of $\mathbf{A}$ (see Song (2013) for a review). The preconditioning matrix used in the algorithm is a variant of incomplete Cholesky factorisation. Let

$$\mathbf{U} = \begin{bmatrix} \mathbf{Z} & 0 & 0 \\ 0 & \mathbf{P}^{1/2} & 0 \\ 0 & 0 & \mathbf{F}^{1/2} \end{bmatrix},$$

where $\mathbf{Z}$ is the upper triangular matrix obtained from the Cholesky decomposition of $\mathbf{X}^\top\mathbf{X}$, $\mathbf{P}^{1/2}$ is the diagonal matrix with the square roots of the diagonal terms of $\mathbf{P}^\top\mathbf{P}$ on the diagonal and $\mathbf{F}^{1/2}$ is the diagonal matrix with the square roots of the diagonal terms of $\mathbf{F}^\top\mathbf{F}$ on the diagonal. Following Fasshauer (2007), rewrite the system as

$$\widetilde{\mathbf{A}}\widetilde{\boldsymbol{\beta}} = \tilde{\mathbf{c}},$$

$$\text{where} \quad \widetilde{\mathbf{A}} = \mathbf{U}^{-\top}\mathbf{A}\mathbf{U}^\top = \begin{bmatrix} I & \mathbf{Z}^{-\top}\mathbf{X}^\top\mathbf{P}\mathbf{P}^{1/2} & \mathbf{Z}^{-\top}\mathbf{X}^\top\mathbf{F}\mathbf{F}^{1/2} \\ \mathbf{P}^{-1/2}\mathbf{P}^\top\mathbf{X}\mathbf{Z}^\top & I & \mathbf{P}^{-1/2}\mathbf{P}^\top\mathbf{F}\mathbf{F}^{1/2} \\ \mathbf{F}^{-1/2}\mathbf{F}^\top\mathbf{X}\mathbf{Z}^\top & \mathbf{F}^{-1/2}\mathbf{F}^\top\mathbf{P}\mathbf{P}^{1/2} & I \end{bmatrix},$$

$$\widetilde{\boldsymbol{\beta}} = \mathbf{U}^{-\top}\boldsymbol{\beta} \quad \text{and} \quad \tilde{\mathbf{c}} = \mathbf{U}^{-\top}\mathbf{c}.$$

The preconditioned conjugate gradient algorithm used in this study was developed by Dongarra (1991) and implemented in Fortran (see Algorithm 1). Let (k) denote the current and $(k+1)$ the next iteration. The CG method computes $\widetilde{\boldsymbol{\beta}}^{(k+1)}$ by iterating

$$\widetilde{\boldsymbol{\beta}}^{(k+1)} = \widetilde{\boldsymbol{\beta}}^{(k)} + \tilde{\alpha}^{(k)}\widetilde{\mathbf{d}}^{(k)},$$

where $\tilde{\alpha}^{(k)}$ is a scalar given by

$$\tilde{\alpha}^{(k)} = \frac{\tilde{\mathbf{r}}^{(k)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k)}}{\widetilde{\mathbf{d}}^{(k)\top}\widetilde{\mathbf{A}}\widetilde{\mathbf{d}}^{(k)}}, \text{ with } \tilde{\mathbf{r}} = \tilde{\mathbf{c}} - \widetilde{\mathbf{A}}\widetilde{\boldsymbol{\beta}}, \text{ and}$$
$$\widetilde{\mathbf{d}}^{(k+1)} = \tilde{\mathbf{r}}^{(k+1)} + \tilde{\delta}^{(k+1)}\widetilde{\mathbf{d}}^{(k)}, \text{ with } \quad \tilde{\delta}^{(k+1)} = \frac{\tilde{\mathbf{r}}^{(k+1)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k+1)}}{\tilde{\mathbf{r}}^{(k)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k)}}.$$

The basic pseudo code is

Algorithm 1 preconditioned conjugate gradient algorithm

procedure
compute the preconditioning matrix $\mathbf{U}$
compute $\widetilde{\mathbf{A}}$ and $\tilde{\mathbf{c}}$
initial $\mathbf{r}^{(0)} = \tilde{\mathbf{c}}$ and let $\mathbf{d}^{(0)} = \mathbf{U}^{-1}\mathbf{r}^{(0)}$
for $k = 1, 2, 3, \cdots$ do
$\tilde{\alpha}^k = \frac{\tilde{\mathbf{r}}^{(k)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k)}}{\widetilde{\mathbf{d}}^{(k)\top}\widetilde{\mathbf{A}}\widetilde{\mathbf{d}}^{(k)}}$
$\widetilde{\boldsymbol{\beta}}^{(k+1)} = \widetilde{\boldsymbol{\beta}}^{(k)} + \tilde{\alpha}^{(k)}\widetilde{\mathbf{d}}^{(k)}$
$\tilde{\mathbf{r}}^{(k+1)} = \mathbf{r}^{(k)} - \tilde{\alpha}^k\widetilde{\mathbf{A}}\widetilde{\mathbf{d}}^{(k)}$
$\tilde{\delta}^{(k)} = \frac{\tilde{\mathbf{r}}^{(k+1)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k+1)}}{\tilde{\mathbf{r}}^{(k)\top}\mathbf{U}^{-1}\tilde{\mathbf{r}}^{(k)}}$
$\widetilde{\mathbf{d}}^{(k+1)} = \tilde{\mathbf{r}}^{(k+1)} + \tilde{\delta}^{(k+1)}\widetilde{\mathbf{d}}^{(k)}$
until the difference between $\widetilde{\boldsymbol{\beta}}^{(k)}$ and $\widetilde{\boldsymbol{\beta}}^{(k+1)}$ is less than 10^{-7}
end procedure

The convergence criterion of $\frac{|\tilde{\mathbf{r}}|}{|\tilde{\mathbf{c}}|} < 10^{-7}$ that we use is similar to that used by others (e.g., Abowd et al., 2002, Hallez et al., 2007).

7.3 Identification using grouping algorithm

The preconditioned CG algorithm does not provide a unique solution for the firm and worker effects. The solutions depend on the initial values, preconditioning matrices and convergence criteria and the implicit constraints used in the algorithm are not necessarily conveniently interpretable. The implicit constraints require the state equations to be satisfied at each iteration. Koopmans (1949), Koopmans et al. (1950) and Fisher (1966) discussed the need to impose model constraints to identify the underlying economic relationship in the observed data. This is because it is possible for two parametric equations to have the same likelihood function unless some restrictions are imposed to uniquely identify parameters. There are an infinite number of possible constraints and solutions. Fujikoshi (1993) summarises several possible approaches for two-way cross classified unbalanced data.

7.3.1 Issues in Identification

We use a simplified version of model (1) in this subsection to illustrate the issues faced in imposing appropriate model restrictions on the model for workers' wages. For simplicity, we consider a single fixed t and replace the observable worker characteristics terms $\mathbf{x}_{it}^{\top}\boldsymbol{\alpha}$ by the fixed unknown constant μ. With these simplifications, the model (1) has expectation

$$E\{\ln(y_{it})\} = \mu + \theta_i + \mathbf{f}_{it}^{\top}\psi = \mu + \theta_i + \psi_j, \tag{16}$$

when worker i works for firm j at time t. With t fixed, it is convenient to make the dependence on j more explicit and, just for this subsection, replace y_{it} by y_{ij}. We consider a two-way table of 5 workers labelled θ_i for $i = 1, \cdots, 5$ and 4 firms labelled ψ_j for $j = 1, \cdots, 4$. If we only have one observation in every cell, we can represent the table as shown in figure 3. In practice, we often do not have one observation in every cell. A simple example is shown in figure 4.

We describe the data in figure 3 as balanced and in figure 4 as unbalanced. The saturated model, the main effect without interaction model for the *balanced* data, is given by (16). The model matrix $(\mathbf{P}, \mathbf{F})$ is given in figure 2(a). The relationships between the columns in the model matrix in figure 2(a) are

$$\beta_0 = \sum_{i=1}^{5} \theta_i \tag{17a}$$

$$\beta_0 = \sum_{j=1}^{4} \psi_j, \tag{17b}$$

where the sums are interpreted as the sums of the vectors in the columns labelled by μ, the θ_i and the ψ_j. These relationships show that the model is over-parameterised with ten parameters when only eight are needed so is rank deficient. This means that there are an infinite number of solutions that satisfy the ordinary least squares normal equation (1). The simplest way to identify unique solutions is by using the corner point constraint to set redundant parameters to zero, i.e. $\theta_5 = \psi_4 = 0$ (Holmes et al., 1997). This is shown in figure 2(b). After imposing the corner point constraint, the model is of full rank so the normal equations have a unique solution.

Figure 2: Model matrices for balanced two-way table

(a) No constraints

β_0	θ_1	θ_2	θ_3	θ_4	θ_5	ψ_1	ψ_2	ψ_3	ψ_4
1	1	0	0	0	0	1	0	0	0
1	1	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	0	1
1	0	1	0	0	0	1	0	0	0
1	0	1	0	0	0	0	1	0	0
1	0	1	0	0	0	0	0	1	0
1	0	1	0	0	0	0	0	0	1
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	0	1	0	0
1	0	0	1	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	1
1	0	0	0	1	0	1	0	0	0
1	0	0	0	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	1	0	0	0
1	0	0	0	0	1	0	1	0	0
1	0	0	0	0	1	0	0	1	0
1	0	0	0	0	1	0	0	0	1

full rank →

(b) With corner point constraints

β_0	θ_1	θ_2	θ_3	θ_4	ψ_1	ψ_2	ψ_3
1	1	0	0	0	1	0	0
1	1	0	0	0	0	1	0
1	1	0	0	0	0	0	1
1	1	0	0	0	0	0	0
1	0	1	0	0	1	0	0
1	0	1	0	0	0	1	0
1	0	1	0	0	0	0	1
1	0	1	0	0	0	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	0	1	0
1	0	0	1	0	0	0	1
1	0	0	1	0	0	0	0
1	0	0	0	1	1	0	0
1	0	0	0	1	0	1	0
1	0	0	0	1	0	0	1
1	0	0	0	1	0	0	0
1	0	0	0	0	1	0	0
1	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0

Figure 3: Balanced two-way table

	ψ_1	ψ_2	ψ_3	ψ_4
θ_1	A	A	A	A
θ_2	A	A	A	A
θ_3	A	A	A	A
θ_4	A	A	A	A
θ_5	A	A	A	A

Figure 4: Unbalanced two-way table [1]

	ψ_1	ψ_2	ψ_3	ψ_4
θ_1	A	A	NA	NA
θ_2	A	A	NA	NA
θ_3	A	A	NA	NA
θ_4	NA	NA	A	A
θ_5	NA	NA	A	A

In comparison, the model matrix for the unbalanced data is shown in figure $5(a)$. As can be seen from figure 4, the observation pattern forms two groups. This model is also rank deficient. If we apply corner point constraints by setting redundant parameters to zero, i.e. $\theta_5 = \psi_4 = 0$, the model matrix is shown in figure $5(b)$.

[1] A = available NA = unavailable

Figure 5: Model matrices for unbalanced two-way table

(a) No constraints

β_0	θ_1	θ_2	θ_3	θ_4	θ_5	ψ_1	ψ_2	ψ_3	ψ_4
1	1	0	0	0	0	1	0	0	0
1	1	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	1	0	0	0
1	0	0	1	0	0	0	1	0	0
1	0	0	0	1	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	1	0
1	0	0	0	0	1	0	0	0	1

not full rank

(b) With corner point constraints

β_0	θ_1	θ_2	θ_3	θ_4	ψ_1	ψ_2	ψ_3
1	1	0	0	0	0	1	0
1	0	1	0	0	1	0	0
1	0	1	0	0	0	1	0
1	0	0	1	0	1	0	0
1	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1
1	0	0	0	1	0	0	0
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0

full rank

full rank

β_0	θ_1	θ_2	θ_4	ψ_1	ψ_3
1	1	0	0	1	0
1	1	0	0	0	0
1	0	1	0	1	0
1	0	1	0	0	0
1	0	0	0	1	0
1	0	0	0	0	0
1	0	0	1	0	1
1	0	0	1	0	0
1	0	0	0	0	1
1	0	0	0	0	0

(c) Corner point constraints with group structure

However, the model matrix in figure 5(b) is still singular because $\theta_1 + \theta_2 + \theta_3 = \psi_1 + \psi_2$.

$$\overset{\theta_1+\theta_2+\theta_3}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \equiv \overset{\psi_1+\psi_2}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

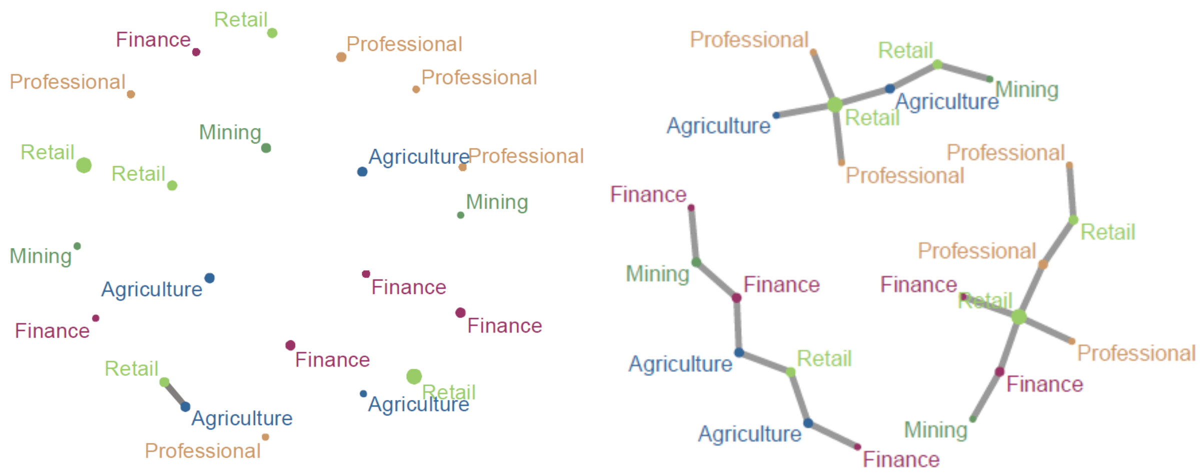
figure 4 shows that the unbalanced data separates into two groups called connected groups (Searle, 1987). We need to take the grouping structure into account to identify unique firm and worker effects. There are an infinite number of possible constraints to make the model matrix of full rank; the particular choice from these is arbitrary. An example is to impose $\psi_2 = 0$. The model matrix for the resulting full rank model is shown in figure 5(c).

Abowd et al. (2002) recognised the need to find connected groups of workers and firms to set model constraints to analyse linked employer and employee data. Firms and workers can be connected by a worker changing jobs or by multiple job holders who work for different firms. These connected groups are formed in such a way that no one worker or firm can be included in more than one group. figure $6(a)$ and figure $6(b)$ show how the algorithm connects firms and workers into mutually exclusive groups. The size of the circle represents the size of the firms to show that connections can occur between firms of different sizes. An edge connects two firms through a worker changing jobs from one firm to the other or holding jobs in both firms. These connected groups are mutually exclusive because there are no additional worker movements. See Algorithm 2 for details.

Figure 6: Connected groups

(a) First connection

(b) Groups formed



Abowd et al. (2002) proposed a grouping algorithm to create groups of connected workers and firms in the data for $g = 1, \cdots, G$ groups (see Algorithm 2).

Algorithm 2 grouping algorithm

procedure
  Order by firm id and then worker id.
  for $group = 1$: assign first firm j to group $g = 1$,
    partitioning step
    repeat
      add all workers employed by a firm j in group $g = 1$ to group $g = 1$.
      add all firms that have employed a worker i in group $g = 1$ to group $g = 1$.
    until no more firms or workers can be added to group $g = 1$.
    end partitioning step
  for $group = 2$: $\forall$ worker $i \notin g = 1$ and $\forall$ firm $j \notin g = 1$ assign first firm j to $g = 2$,
    repeat *partitioning* step and add all workers and firms in group $g = 2$ to group $g = 2$.
  for $group = 3$: $\forall$ worker $i \notin g = 1, 2$ and $\forall$ firm $j \notin g = 1, 2$ assign first firm j to $g = 3$,
    repeat *partitioning* step and add all workers and all firms in group $g = 3$ to group $g = 3$.
    ⋮
  for $group = G$: $\forall$ worker $i \notin g = 1, 2, \cdots, G-1$ and $\forall$ firm $j \notin g = 1, 2, \cdots, G-1$ assign first firm j to $g = G$,
    repeat *partitioning* step and add all workers and all firms in group $g = G$ to group $g = G$.
  until all firms are assigned.
  end procedure

The algorithm divides connected workers and firms into mutually exclusive groups. A group is defined as all workers and firms that are connected through some migration of workers between firms in that group, and such that there is no migration of a worker within the group to any firm outside the group. The main result is that the ensuing model matrix is of full rank so the solutions to the ordinary least squares normal equations are unique.

8 EMPIRICAL RESULTS

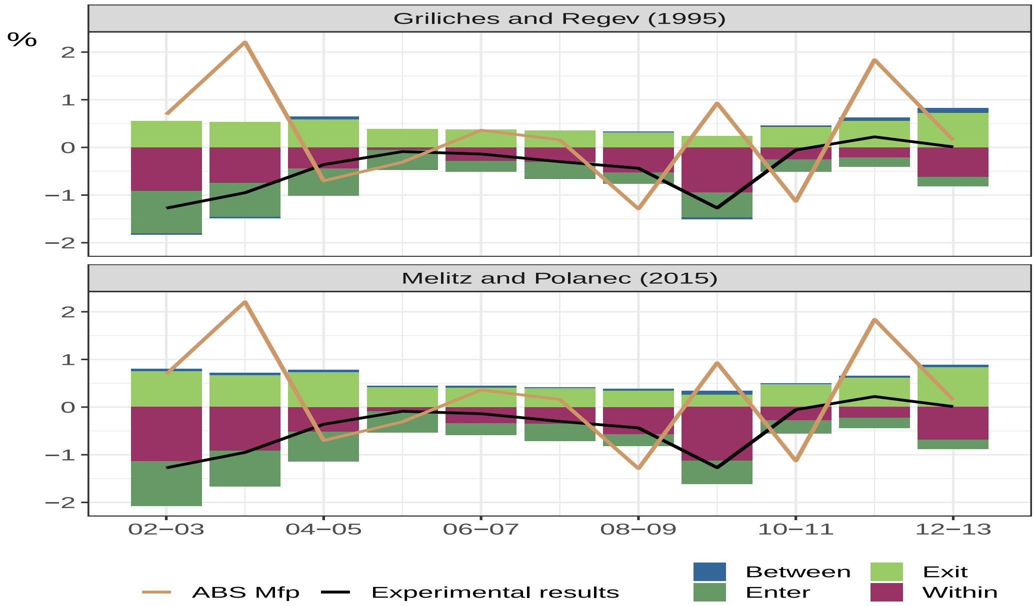
8.1 Firm dynamics and aggregate productivity

This study shows the usefulness of firm-level analysis for comparing the contribution that entering, exiting and surviving firms make to aggregate productivity. These contributions are quite different at the industry level. The analytical results can be extended to explore the link between the contribution of younger firms and overall growth to inform policies and encourage economic growth (see Andrews et al., 2015).

Figure 7 shows the estimated contributions from surviving, entering and exiting firms to aggregate productivity using the methods of Griliches and Regev (1995) and Melitz and Polanec (2015). Nguyen and Hansell (2014) and Melitz and Polanec (2015) note the importance of taking into account the appropriate counterfactual to derive the contributions from surviving, entering and exiting firms. We concur, particularly for the results from smaller industries (see Appendix C). The results show that the differences between the methods of Griliches and Regev (1995) and Melitz and Polanec (2015) are greater for entering and exiting firms in smaller industries. We have also explored the aggregation method proposed by Foster et al. (2001); the results are similar to those for the approach of Griliches and Regev (1995).

Figure 7 shows that our results are broadly consistent with published ABS annual productivity measures at the aggregate level. Our analysis provides useful insights into the variability of firms' contributions to aggregate productivity growth; this information is not available in ABS publications. We find similar productivity growth movements over time except in 2004–05, 2009–10 and 2010–11 when we compare published ABS and our experimental results. The differences may be due to the fact that we use different prices to derive the volume measures. This introduces differences in relative prices when estimating firm productivity. These differences result in different substitution effects between labour and capital and between goods and services, which can lead to different results. (see Dumagan and Balk, 2016, and Duarte and Restuccia, 2017, on the role of relative prices in estimating productivity.) In addition, we use firm-level capital cost instead of firm-level capital stock measures for our analysis. This is because there is no information on firm-level asset prices. This information is required to derive capital stock measures using the perpetual inventory method (Walters and Dippelsman, 1986).

Figure 7: All industry decomposition



Note. Between and Within are the contributions from surviving firms and Enter and Exit are the contributions from entering and exiting firms to the aggregate productivity indicated by Experimental results. The derivation of these measures can be found in (10) and (12) Griliches and Regev (1995), Melitz and Polanec (2015). ABS Mfp is the published ABS Estimates of Industry Multifactor Productivity (ABS, 2013).

Like Nguyen and Hansell (2014), this study has found that the net contribution from entering and exiting firms is smaller in manufacturing than in services industries in general. The within-industry contribution component generally has a smaller contribution in services industries. This may imply that entering and exiting firms are the main source of productivity changes. At the industry level, our experimental results show similar patterns with the ABS results, particularly for the Agriculture, Forestry and Fishing (A), Construction (E), Financial and Insurance Services (K) and Administrative Service industries. The industries with notable differences are Mining (B) and Electricity, Gas and Water (D), especially in 2012–13. As discussed, the difference may be caused by different price deflators and the different methods used to derive capital measures.

8.2 Firm level model results

This study confirms the importance of correcting for endogeneity in estimating the firm-level production function. The estimated labour coefficients for the firm models' wages (WAGES) are higher than $\ln L$ for all industries. Table 4 shows the estimated coefficients for the firm-level model results for All industries (using (6)) and Agriculture, Forestry and Fishing industry (using (4)). Appendix *D* contains results for all other industries).

We show both estimated coefficients using *Balanced* and *Imputed* datasets. The first column under the *Balanced* and *Imputed* subheadings are the results from the instrumental variable and the second column contains the results from the ordinary least square. For All industries, the estimated coefficient of WAGES is 0.723 in balanced and 0.706 in imputed datasets (in ALL.OLS columns), stronger than the estimated instrumental variable, $\ln\hat{z}_t^{(jk)}$, which is 0.691 in balanced and 0.648 in imputed datasets (in ALL.2SLS columns). Similarly, we observe similar industry results. Agriculture, Forestry and Fishing industry (A), the estimated coefficient of

WAGES is 0.212 in balanced and 0.248 in imputed datasets (in A.OLS columns), stronger than the estimated instrumental variable, $\ln\hat{z}_t^{(jk)}$, which is −0.036 in balanced and −0.016 in imputed datasets (in A.2SLS columns). The lower estimations of labour component are consistent with similar studies using instrumental variables to correct for endogeneity (see Breunig and Wong, 2008, and Levinsohn and Petrin, 2003). This correction is important to avoid bias in the aggregate industry decomposition results.

Table 4: All industries (ALL) and Agriculture, Forestry and Fishing (A) industry results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	ALL.2SLS	ALL.OLS	ALL.2SLS	ALL.OLS	A.2SLS	A.OLS	A.2SLS	A.OLS
$\ln\hat{z}_t^{(jk)}$	0.691***		0.648***		−0.036**		−0.016***	
	(0.001)		(0.0003)		(0.018)		(0.004)	
WAGES		0.723***		0.706***		0.212***		0.248***
		(0.001)		(0.0003)		(0.002)		(0.001)
LnK	0.223***	0.251***	0.246***	0.239***	0.503***	0.447***	0.453***	0.402***
	(0.001)	(0.001)	(0.0002)	(0.0002)	(0.002)	(0.002)	(0.001)	(0.001)
LnM	0.242***	0.183***	0.238***	0.170***	0.114***	0.069***	0.129***	0.076***
	(0.0004)	(0.0005)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)
Firm_Age	0.120***	0.054***	0.159***	0.060***	0.008	−0.061***	0.022***	−0.033***
	(0.001)	(0.001)	(0.0005)	(0.0005)	(0.005)	(0.005)	(0.003)	(0.002)
Year2003	0.170***	0.328***	0.260***	0.445***	−0.148***	−0.094***	−0.075***	−0.038***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.014)	(0.012)	(0.007)	(0.006)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Year2013	1.914***	0.107***	1.700***	0.266***	−0.275***	−0.016	−0.099***	0.103***
	(0.004)	(0.003)	(0.002)	(0.002)	(0.063)	(0.016)	(0.015)	(0.008)
BAS_divB	−0.416***	−1.118***	−0.131***	−0.747***				
	(0.014)	(0.014)	(0.006)	(0.006)				
⋮	⋮	⋮	⋮	⋮				
BAS_divS	−0.472***	−1.022***	−0.439***	−0.700***				
	(0.003)	(0.003)	(0.002)	(0.002)				
Observations	2,296,984	2,296,984	10,039,638	10,039,638	162,766	162,766	662,553	662,553
Adjusted R^2	0.992	0.992	0.990	0.990	0.363	0.399	0.357	0.405
Note:								*p<0.1; **p<0.05; ***p<0.01

8.3 Worker level model results

It is essential to include firms connected by workers to uniquely identify worker and firm effects (Abowd et al., 1999). Table 5 shows the pattern of workers who have different employers in the sample. The columns indicate the number of years that a worker stays in the sample and the rows correspond with the number of employers workers have over the 11 years of data. It is more likely for workers to work for more employers when they stay in the sample for longer. There are significant worker movements between firms in the sample. Only 23.27% of workers have one employer over the 11-year period

Table 5: Number of job changes and number of years in the sample

number of employers	number of years in sample 1	2	3	4	5	6	7	8	9	10	>10	Total
1	3.54	3	2.12	5.63	0.83	0.74	0.66	0.62	0.72	0.87	4.53	23.27
2	1	2.17	1.79	3.69	1.07	0.97	0.86	0.8	0.88	1.03	4.57	18.82
3	0.33	1	1.16	2.22	1.04	0.99	0.91	0.85	0.92	1.02	4.13	14.56
4	0.12	0.46	0.61	1.26	0.81	0.87	0.84	0.81	0.88	0.95	3.5	11.11
5	0.05	0.21	0.32	0.7	0.54	0.66	0.7	0.72	0.78	0.83	2.87	8.37
6	0.02	0.1	0.16	0.39	0.34	0.46	0.54	0.59	0.66	0.69	2.29	6.24
7	0.01	0.05	0.09	0.22	0.2	0.31	0.39	0.45	0.53	0.57	1.8	4.63
8	·	0.02	0.05	0.12	0.12	0.2	0.28	0.34	0.41	0.44	1.4	3.4
9	·	0.01	0.02	0.07	0.07	0.13	0.19	0.25	0.32	0.35	1.07	2.48
10	·	0.01	0.01	0.04	0.04	0.08	0.13	0.18	0.24	0.26	0.82	1.81
>10	·	0.01	0.02	0.07	0.07	0.15	0.28	0.44	0.66	0.81	2.81	5.31
Total	5.07	7.04	6.36	14.39	5.14	5.57	5.78	6.03	6.99	7.83	29.81	100

Note. Number of employers measures how many unique ABN a worker i has over the sample period and number of years in sample measures how many unique year counts a worker i has in the sample.

Table 6 shows the correlation structure of the estimated components in the worker model. This study finds a positive correlation between worker and firm effects. This is in line with the finding of Iranzo et al. (2008) but different from Abowd et al. (2002). Andrews et al. (2008) suggest that the negative correlation in previous studies may arise from a lack of worker mobility, which is not the case in this Australian sample.

Table 6: Pearson correlation coefficients of estimated components

	$logL$	θ	ψ	$\mathbf{X}\boldsymbol{\alpha}$	ϵ
$logL$	—	0.3063***	0.5490***	–0.2115***	0.5923***
θ	0.3063***	—	0.1058***	–0.9793***	–0.0085***
ψ	0.5490***	0.1058***	—	–0.0966***	–0.0021***
$\mathbf{X}\boldsymbol{\alpha}$	–0.2115***	–0.9793***	–0.0966***	—	–0.0267***
ϵ	0.5923***	–0.0085***	–0.0021***	–0.0267***	—

Note. $Prob > |r|$ under $N = 130,281,096$.
*p<0.1; **p<0.05; ***p<0.01.

9 CONCLUSIONS AND FUTURE DIRECTIONS

This study shows the value of using microdata to better understand the components of industry-level productivity growth. It explores methods for fitting a model for workers by solving a large sparse linear system of equations and uses the estimated results to correct for endogeneity in the firm's decisions about how much labour to employ. The paper also calculates the contribution of entering, exiting and surviving firms to aggregate productivity at the industry level.

Our results show the importance of correcting for endogeneity in estimating the production function. The productivity contributions from surviving, entering and exiting firms are quite different across different industries. Understanding these differences may be useful to inform policy.

Across all industries, we generally find that firm exit is the most important contributor to productivity growth. Firm entry generally has a negative impact on industry-level productivity growth. This is similar to what was found by Breunig and Wong (2008) for the 1990s in Australia. It is not surprising, as many new firms end up not surviving. They may lack access to industry-specific knowledge and skills.

Within-firm productivity increases are generally a positive contributor to industry-level productivity, but are very small in about half of the industry groups we examine. Re-allocation effects for continuing firms are virtually non-existent. Almost all of the reallocation is happening through entry and exit.

This would suggest that policies which facilitate firm entry and exit are likely to help in achieving increased productivity gains. Policies which provide large advantages to incumbent firms (such as cumbersome regulation which is difficult to comply with for new entrants) are likely to detract from productivity growth.

Our analysis could be extended in several ways. First, with a better proxy for worker skill such as education, we could better account for the effects of workers. Capturing workers' skill dispersion across and between firms would be useful. Secondly, it would be interesting to explore other estimation approaches like Constant Elasticity of Substitution production functions that allow the elasticity of substitution between capital and labour inputs to better understand the relative prices effects (McFadden, 1963) and (Steenkamp, 2017) . Increased data access and better measures of key variables are both required for such analyses.

10 REFERENCES

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002), "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Report, US Census Bureau, `ftp://ftp2.census.gov/ces/tp/tp-2002-06.pdf`[Accessed: 12022016].

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999), "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333, `http://www.jstor.org/stable/2999586`[Accessed: 03082017].

ABS (2009), "ANZSCO - Australian and New Zealand Standard Classification of Occupations," 1220.0, `http://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/4AF138F6DB4FFD4BCA2571E200096BAD?opendocument`[Accessed: 01082017].

— (2013), "Estimates of Industry Multifactor Productivity, cat. no. 5260.0.55.002," `https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/5260.0.55.002Main+Features12012-13?OpenDocument`[Accessed: 01022018].

— (2015), "Information Paper: Construction of Experimental Statistics on Employee Earnings and Jobs from Administrative Data, 2011-12," `https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/6311.0main+features12011-12`[Accessed: 01032018].

— (2018a), "6427.0 - Producer Price Indexes, Australia, Sep 2018," `http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6427.0Sep%202018?OpenDocument`[Accessed: 01102018].

— (2018b), "Administrative Data Research for the 2021 Census," `http://www.abs.gov.au/websitedbs/d3310114.nsf/home/2021+census+administrative+data+research`[Accessed: 01022019].

ABS and DIIS (2017), "Business Longitudinal Analysis Data Environment," `https://industry.gov.au/Office-of-the-Chief-Economist/Data/Pages/Business-Longitudinal-Analysis-Data-Environment.aspx`[Accessed: 06062017].

Agresti, A. (2007), *An introduction to categorical data analysis*, New Jersey: John Wiley & Sons, Inc.

Andrews, D., Bartelsman, E., and Criscuolo, C. (2015), "Firm dynamics and productivity growth in Europe," Report, mimeo.

Andrews, M. J., Gill, L., Schank, T., and Upward, R. (2008), "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 673–697, `http://dx.doi.org/10.1111/j.1467-985X.2007.00533.x`[Accessed: 01062017].

Bakhtiari, S. (2015), "Productivity, outsourcing and exit: the case of Australian manufacturing," *Small Business Economics*, 44, 425–447.

Bartelsman, E. J. and Dhrymes, P. J. (1998), "Productivity Dynamics: U.S. Manufacturing Plants, 1972–1986," *Journal of Productivity Analysis*, 9, 5–34.

Bartelsman, E. J. and Doms, M. (2000), "Understanding Productivity Lessons from Longitudinal Microdata," *Journal of Economic Literature*, 38, 569–594.

Breunig, R. and Wong, M.-H. (2008), "A Richer Understanding of Australia's Productivity Performance in the 1990s: Improved Estimates Based Upon Firm-Level Panel Data," *Economic Record*, 84, 157–176.

Chien, C.-H. and Mayer, A. (2015a), "A New Analytical Platform to Explore Linked Data," Report, Australian Bureau of Statistics, `http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.151`[Accessed: 05082016].

— (2015b), "Use of a prototype Linked Employer-Employee Database to describe characteristics of productive firms," Report, Australian Bureau of Statistics, `https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.055`[Accessed: 05082016].

Chien, C.-H., Welsh, A. H., and Moore, J. D. (2018), "Synthetic Microdata - A Possible Dissemination Tool," Report, Australian Bureau of Statistics, `http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.163`[Accessed: 16112018].

Czepiel, S. A. (2002), "Maximum likelihood estimation of logistic regression models: theory and implementation," `http://ww.saedsayad.com/docs/mlelr.pdf`[Accessed: 01032019].

Dongarra, J. J. (1991), *Solving linear systems on vector and shared memory computers*, Philadelphia: Society for Industrial and Applied Mathematics.

Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control Theory and Implementation*, Lecture Notes in Statistics, New York: Springer.

Duarte, M. and Restuccia, D. (2017), "Relative Prices and Sectoral Productivity," Report, National Bureau of Economic Research.

Dumagan, J. C. and Balk, B. M. (2016), "Dissecting aggregate output and labour productivity change: a postscript on the role of relative prices," *Journal of Productivity Analysis*, 45, 117–119.

Fasshauer, G. (2007), "477 577 Numerical Linear Algebra Computational Mathematics I," `http://www.math.iit.edu/~fass/477_577.html`[Accessed: 20012016].

Fisher, F. M. (1966), *The identification problem in econometrics*, McGraw-Hill.

Foster, L., Haltiwanger, J. C., and Krizan, C. J. (2001), *Aggregate productivity growth: lessons from microeconomic evidence*, University of Chicago Press, pp. 303–372.

Fox, J. T. and Smeets, V. (2011), "DOES INPUT QUALITY DRIVE MEASURED DIFFERENCES IN FIRM PRODUCTIVITY?" *International Economic Review*, 52, 961–989, `http://www.jstor.org/stable/41349184`[Accessed: 20012016].

Fujikoshi, Y. (1993), "Two-way ANOVA models with unbalanced data," *Discrete Mathematics*, 116, 315–334, `http://www.sciencedirect.com/science/article/pii/0012365X9390410U`[Accessed: 20012016].

Gandhi, A., Navarro, S., and Rivers, D. A. (2011), "On the identification of production functions: How heterogeneous is productivity?" http://www.ssc.wisc.edu/~agandhi/homepage/Amit_Gandhi_files/production_9_25_13_FULL.pdf[Accessed: 20012016].

Griliches, Z. and Regev, H. (1995), "Firm productivity in Israeli industry 1979–1988," *Journal of Econometrics*, 65, 175–203, http://www.sciencedirect.com/science/article/pii/030440769401601U[Accessed: 20012017].

Hallez, H., Vanrumste, B., Grech, R., Muscat, J., De Clercq, W., Vergult, A., D'Asseler, Y., Camilleri, K. P., Fabri, S. G., and Van Huffel, S. (2007), "Review on solving the forward problem in EEG source analysis," *Journal of neuroengineering and rehabilitation*, 4, 46.

Hansell, D. and Rafi, B. (2018), "Firm-Level Analysis Using the ABS' Business Longitudinal Analysis Data Environment (BLADE)," *Australian Economic Review*, 51, 132–138, https://doi.org/10.1111/1467-8462.12253[Accessed: 05032019].

Holmes, A., Poline, J., and Friston, K. (1997), "Characterizing brain images with the general linear model," in *Human Brain Function*, eds. Frackowiak, R., Friston, K., Frith, C., Dolan, R., and Mazziotta, J., Academic Press USA, pp. 59–84.

Iranzo, S., Schivardi, F., and Tosetti, E. (2008), "Skill Dispersion and Firm Productivity: An Analysis with Matched Employer-Employee Data," *Journal of Labor Economics*, 26, 247–285.

Koopmans, T. C. (1949), "Identification problems in economic model construction," *Econometrica, Journal of the Econometric Society*, 125–144.

Koopmans, T. C., Rubin, H., and Leipnik, R. B. (1950), *Measuring the equation systems of dynamic economics*, New York: John Wiley and Sons, vol. Monograph No. 10, pp. 53–237.

Lafrance, A. and Baldwin, J. R. (2011), "Firm Turnover and Productivity Growth in Selected Canadian Services Industries, 2000 to 2007," Tech. rep., Statistics Canada, Analytical Studies Branch.

Lentz, R. and Mortensen, D. T. (2010), "Labor market models of worker and firm heterogeneity," *Annu. Rev. Econ.*, 2, 577–602.

Levinsohn, J. and Petrin, A. (2003), "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341, http://www.jstor.org/stable/3648636[Accessed: 25012017].

Mare, D. C., Hyslop, D. R., and Fabling, R. (2017), "Firm productivity growth and skill," *New Zealand Economic Papers*, 302–326.

McFadden, D. (1963), "Constant elasticity of substitution production functions," *The Review of Economic Studies*, 30, 73–83.

Melitz, M. J. and Polanec, S. (2015), "Dynamic Olley-Pakes productivity decomposition with entry and exit," *The Rand journal of economics*, 46, 362–375.

Nguyen, T. and Hansell, D. (2014), "Firm dynamics and productivity growth in Australian manufacturing and business services Oct 2014," Report, ABS, https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.052[Accessed: 20052017].

Olley, G. S. and Pakes, A. (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, 27, 85–96.

Reed, W. R. (2015), "On the Practice of Lagging Variables to Avoid Simultaneity," *Oxford Bulletin of Economics and Statistics*, 77, 897–905.

Reiter, J. P. (2005), "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 185–205.

Schomaker, M. and Heumann, C. (2014), "Model selection and model averaging after multiple imputation," *Computational Statistics and Data Analysis*, 71, 758–770.

Searle, S. R. (1987), *Linear models for unbalanced data*, New York: Wiley.

Shewchuk, J. R. (1994), "An introduction to the conjugate gradient method without the agonizing pain," http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf[Accessed: 03082017].

Song, H. (2013), "ECS231 Large-Scale Scientific Computation Course College of Engineering," http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.592.1575&rep=rep1&type=pdf[Accessed: 03082017].

Steenkamp, D. (2017), "Productivity and relative prices," Thesis, https://researchspace.auckland.ac.nz/handle/2292/33749[Accessed: 03082017].

Syverson, C. (2011), "What Determines Productivity?" *Journal of Economic Literature*, 49, 326–365.

Templ, M., Alfons, A., and Filzmoser, P. (2012), "Exploring incomplete data using visualization techniques," *Advances in Data Analysis and Classification*, 6, 29–47.

Walters, R. and Dippelsman, R. J. (1986), *Estimates of depreciation and capital stock, Australia*, Australian Bureau of Statistics, https://nla.gov.au/nla.cat-vn446271[Accessed: 03082017].

Wilkinson, G. N. and Rogers, C. E. (1973), "Symbolic Description of Factorial Models for Analysis of Variance," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22, 392–399.

Wooldridge, J. M. (2006), *Introductory econometrics: a modern approach*, Mason, OH: Thomson/South-Western, 3rd ed.

Zellner, A., Kmenta, J., Dr, xe, and ze, J. (1966), "Specification and Estimation of Cobb-Douglas Production Function Models," *Econometrica*, 34, 784–795.

A IMPUTATION METHODS FOR CATEGORICAL DATA

We use the available information from the experimental dataset to allocate firm j belonging to an unknown industry U into different industries. The font—$\boldsymbol{\mathcal{X}}$—represents observed dataset in the notation. The formula to allocate firms into different industries is:

$$
\begin{aligned}
Pr(j = k \mid \boldsymbol{\mathcal{X}}_{jkt}) &= \frac{exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}{1 + \sum_{k=1}^{K-1} exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}, k = 1, \cdots, K-1 \\
\vdots \quad &= \quad \vdots \\
Pr(j = K \mid \boldsymbol{\mathcal{X}}_{jkt}) &= \frac{1}{1 + \sum_{k=1}^{K-1} exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}.
\end{aligned} \tag{18}
$$

The 1 terms in the denominator and in the numerator of $Pr(j = K \mid \boldsymbol{\mathcal{X}}_{jkt})$ ensure that probabilities over the response categories equal 1 (Czepiel, 2002, Agresti, 2007). It is convenient to write the term $\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k$ in Wilkinson and Rogers's (1973) notation. The term $\boldsymbol{\mathcal{X}}$ contains $Firm\_Age + Employees + \tau$ where $Firm\_Age$ is the age of firms and $Employees$ is the number of employees that firm j has. The variable τ is represented by 10 time-indicator variables, one for each year with 2001–02 as the baseline. This makes each $\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k$ a sum of 12 terms. The formula is applied to the complete cases to obtain the industry coefficients $\mathbf{a}_k$ with $k = 1, \cdots, 17$ industries. We combine these estimated coefficients with firm characteristics data $\boldsymbol{\mathcal{X}}_{jkt}$ for firms with the missing industry. We allocate firm j to an industry with the highest predictive probability.

B IMPUTATION METHODS FOR CONTINUOUS DATA

Next, we assume MAR and impute missing values in the combined ABS and IPGOD datasets by imputed industry. We use sequential regression in SAS `proc mi` procedure for the imputation. We adapt a similar notation to Reiter (2005). The experimental dataset consists of $[\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{y}$ is an $N \times 1$ vector that includes the dependent variable, and $\boldsymbol{\mathcal{X}}$ is an $N \times 15$ matrix that includes all the independent variables from (3). This gives 15 unknown regression parameters in (3). We impute missing variables ln$y$, ln$K$ and ln$M$. The observed dataset consists of two $N \times 16$ matrices, $\boldsymbol{\mathcal{D}} = [\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{\mathcal{X}}$ includes all the independent variables from (3); and the response indicator matrix $\boldsymbol{\mathcal{R}}$ which we use to partition $\boldsymbol{\mathcal{D}}$ into the observed $\boldsymbol{\mathcal{D}}^{obs}$ and the missing $\boldsymbol{\mathcal{D}}^{mis}$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the design matrix for imputing missing data in ln$y$, ln$K$ and ln$M$, respectively.

We impute the missing values in ln$y$, ln$K$ and ln$M$ separately using sequential regression (SR). The SR method uses appropriate regression models for different variable types. For example, continuous variables are imputed using a normal model and binary variables using a logit model. The SR method generates a continuous vector $\boldsymbol{y}^{seq}$ from the parameters directly estimated from the fitted regression following Raghunathan et al. (2001). The SR formula for generating missing data for $\boldsymbol{y}$ is:

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}. \tag{19}$$

We apply (19) three times, with $\boldsymbol{y}$ denoting each of the three variables ln$y$, ln$K$ and ln$M$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the design matrix for creating missing data in ln$y$, ln$K$ and ln$M$, respectively. If the missing data variable is ln$y$, then $\boldsymbol{\mathcal{X}}$ includes all the independent variables from (3). In comparison, if the missing data variable is ln$K$, then $\boldsymbol{\mathcal{X}}^{(K)}$ includes all the independent variables and ln$y$ but excludes ln$K$. Similarly, if the missing data variable is ln$M$, then $\boldsymbol{\mathcal{X}}^{(M)}$ includes all the independent variables and ln$y$ but excludes ln$M$. Algorithm 3 describes the basic concept of the algorithm (Drechsler, 2011).

Algorithm 3 Sequential regression algorithm

```
procedure
    Step 1: draw a new value θ = (σ², β) from Pr(θ | y_obs)
        draw variance from σ²|X_obs ~ (y_obs − X_obs β̂)′(y_obs − X_obs β̂) χ⁻²_{n−k}, where n is the total number
        of observations and k is the number of parameters
        draw coefficients from β | σ², X_obs ~ N(β̂, (X′_obs X_obs)⁻¹ σ²)
    Step 2: draw an imputed value y^seq from Pr(y^seq | y_obs, θ)
        draw from fitted regression y^seq | β, σ², X_obs ~ N(X_obs β, σ²)
    repeat Step 1 and Step 2 to impute each variable sequentially
```

We create 10 imputed datasets in each imputed industry and we select the best imputed dataset which maximises the likelihood for equation (3) from the 10 datasets in each industry (Schomaker and Heumann, 2014, Chien et al., 2018).

Figure 8: Industry Decomposition

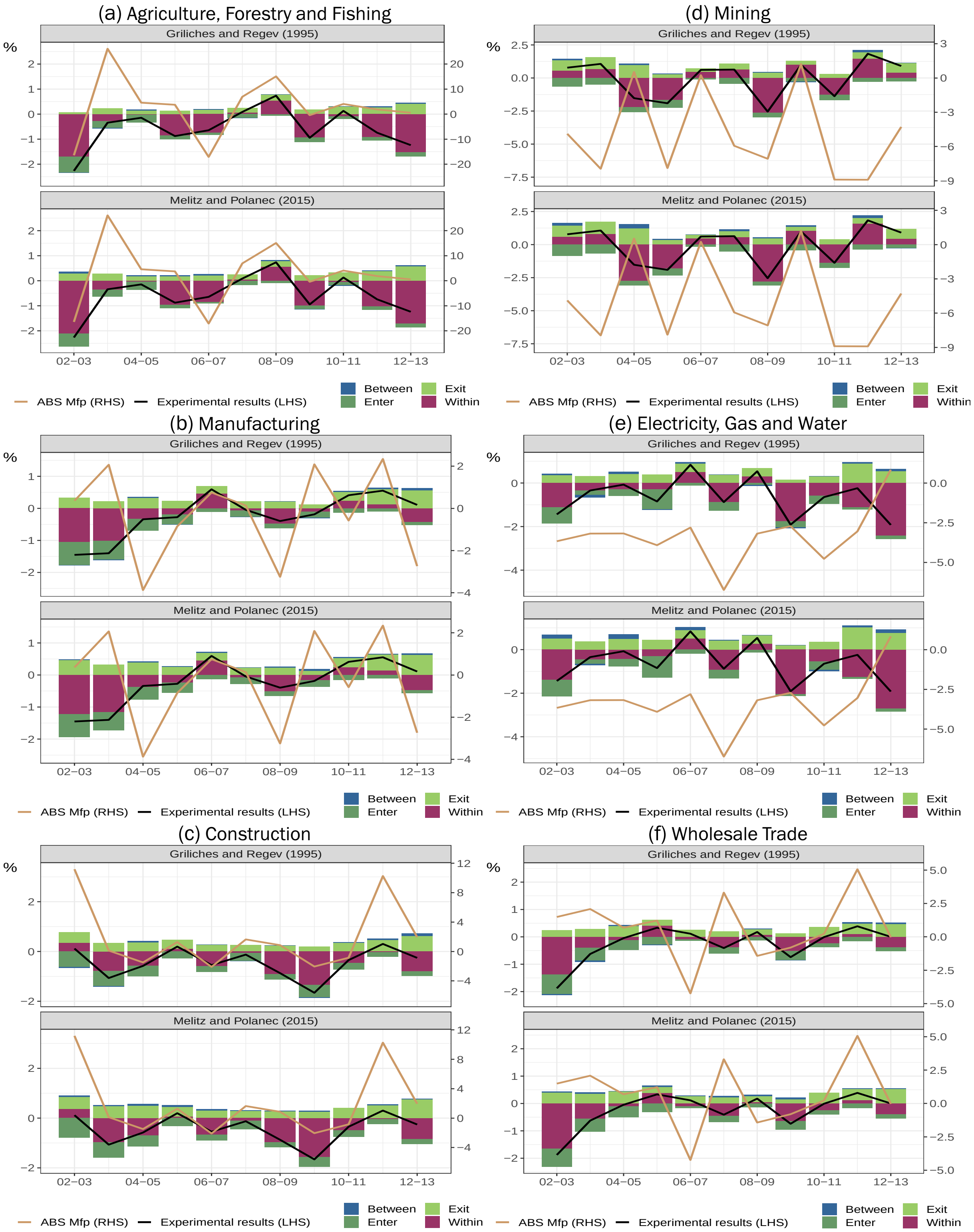


Figure 9: Industry Decomposition

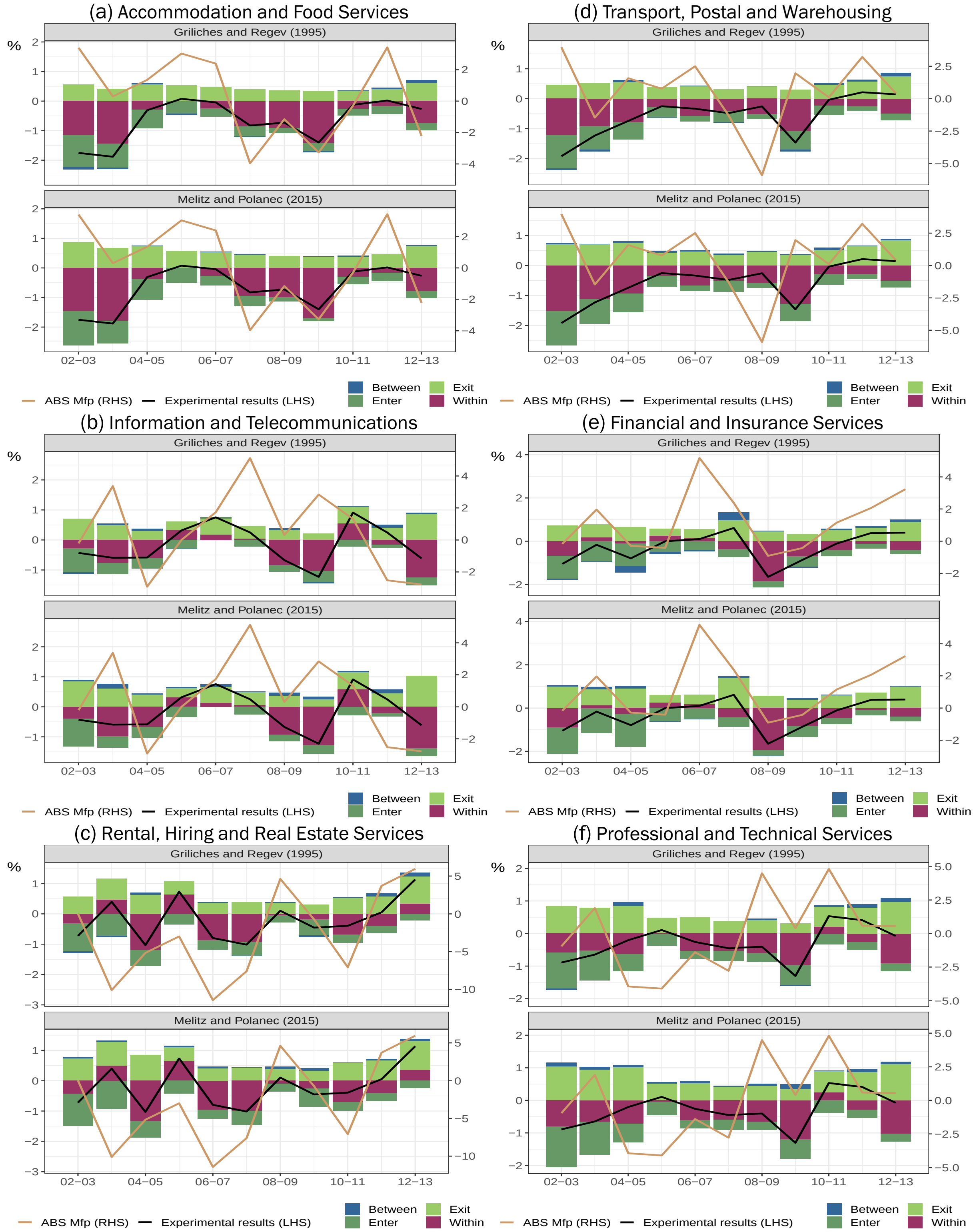
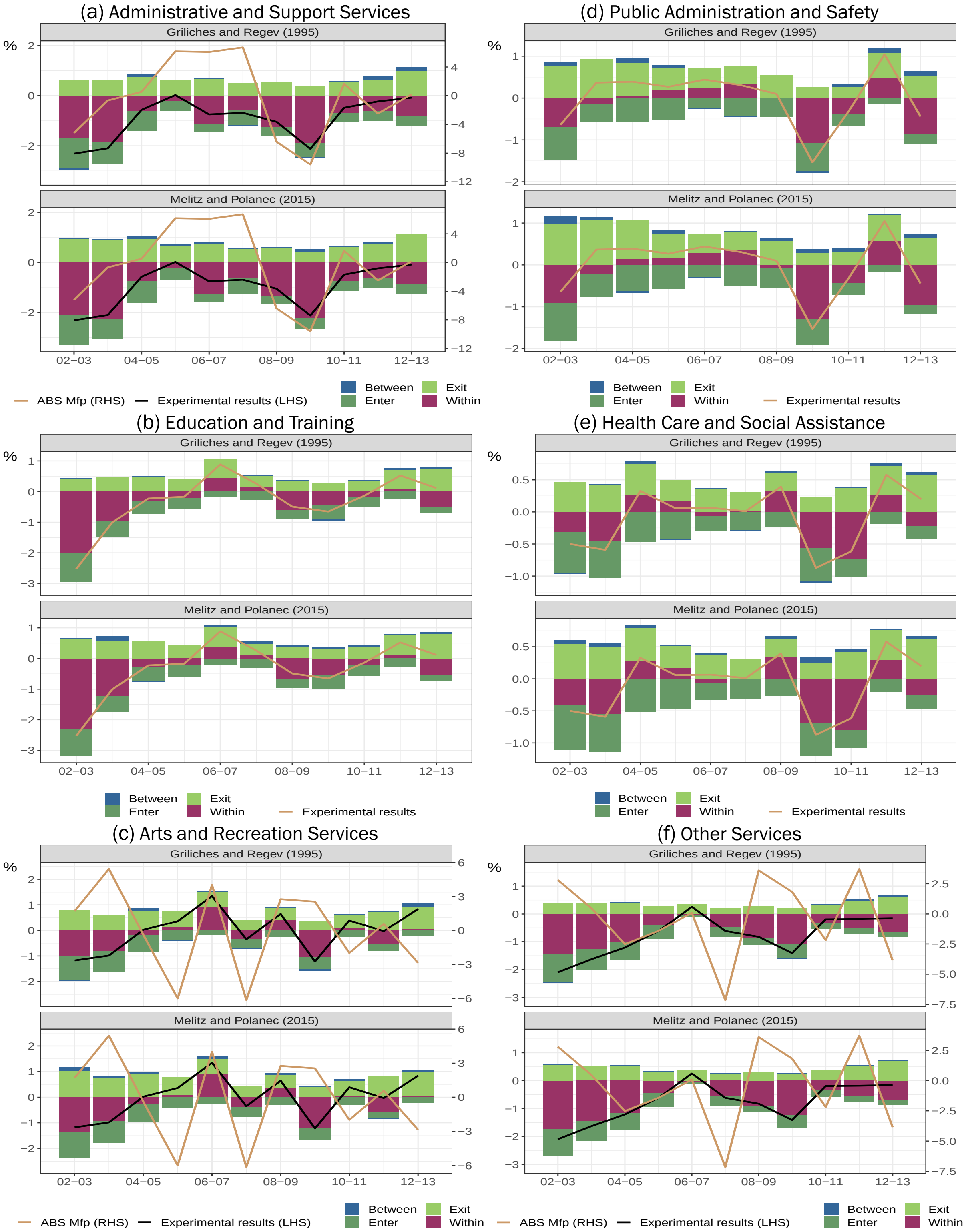


Figure 10: Industry Decomposition



D FIRM MODEL RESULTS

Table 7: Mining (B) and Manufacturing (C) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	B.2SLS	B.OLS	B.2SLS	B.OLS	C.2SLS	C.OLS	C.2SLS	C.OLS
$\ln\hat{z}_t^{(jk)}$	0.584***		0.235***		0.565***		0.090***	
	(0.115)		(0.014)		(0.012)		(0.003)	
WAGES		0.563***		0.546***		0.527***		0.469***
		(0.017)		(0.005)		(0.002)		(0.001)
LnK	0.313***	0.218***	0.226***	0.162***	0.200***	0.127***	0.179***	0.126***
	(0.011)	(0.011)	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
LnM	0.159***	0.090***	0.203***	0.108***	0.317***	0.189***	0.352***	0.216***
	(0.007)	(0.007)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
lnFirm_Age	−0.012	0.016	0.002	−0.008	0.111***	0.041***	0.135***	0.049***
	(0.022)	(0.019)	(0.007)	(0.006)	(0.002)	(0.002)	(0.002)	(0.001)
Observations	4,902	4,902	36,559	36,559	288,335	288,335	645,869	645,869
R^2	0.289	0.413	0.261	0.431	0.303	0.414	0.320	0.430
Adjusted R^2	0.287	0.411	0.261	0.431	0.303	0.414	0.320	0.430

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Electricity, Gas, Water and Waste Services (D) and Construction (E) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	D.2SLS	D.OLS	D.2SLS	D.OLS	E.2SLS	E.OLS	E.2SLS	E.OLS
$\ln\hat{z}_t^{(jk)}$	0.409***		0.114***		0.123***		0.023***	
	(0.101)		(0.015)		(0.010)		(0.002)	
WAGES		0.541***		0.444***		0.403***		0.355***
		(0.016)		(0.006)		(0.002)		(0.001)
LnK	0.294***	0.180***	0.313***	0.232***	0.170***	0.129***	0.164***	0.134***
	(0.011)	(0.010)	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
LnM	0.134***	0.069***	0.145***	0.084***	0.229***	0.169***	0.245***	0.183***
	(0.007)	(0.007)	(0.003)	(0.003)	(0.001)	(0.001)	(0.0005)	(0.0005)
lnFirm_Age	0.079***	0.016	0.122***	0.060***	0.030***	−0.012***	0.063***	0.018***
	(0.019)	(0.016)	(0.008)	(0.007)	(0.002)	(0.002)	(0.001)	(0.001)
Observations	5,022	5,022	28,837	28,837	373,859	373,859	1,477,460	1,477,460
Adjusted R^2	0.234	0.382	0.269	0.387	0.222	0.313	0.248	0.336

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: Wholesale Trade (F) and Retail Trade (G) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	F.2SLS	F.OLS	F.2SLS	F.OLS	G.2SLS	G.OLS	G.2SLS	G.OLS
$\ln\hat{z}_t^{(jk)}$	0.570***		0.110***		0.622***		0.162***	
	(0.017)		(0.004)		(0.010)		(0.003)	
WAGES		0.493***		0.442***		0.450***		0.412***
		(0.003)		(0.002)		(0.002)		(0.001)
LnK	0.167***	0.085***	0.160***	0.096***	0.139***	0.088***	0.132***	0.090***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
LnM	0.336***	0.235***	0.363***	0.252***	0.362***	0.246***	0.395***	0.271***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
lnFirm_Age	0.112***	0.052***	0.138***	0.059***	0.187***	0.111***	0.221***	0.126***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Observations	213,389	213,389	480,515	480,515	434,058	434,058	1,072,727	1,072,727
Adjusted R^2	0.254	0.330	0.292	0.370	0.241	0.310	0.273	0.345
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 10: Accommodation and Food Services (H) and Transport, Postal and Warehousing (I) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	H.2SLS	H.OLS	H.2SLS	H.OLS	I.2SLS	I.OLS	I.2SLS	I.OLS
$\ln\hat{z}_t^{(jk)}$	0.721***		0.217***		0.604***		0.086***	
	(0.013)		(0.003)		(0.032)		(0.003)	
WAGES		0.447***		0.420***		0.417***		0.424***
		(0.002)		(0.001)		(0.005)		(0.001)
LnK	0.189***	0.132***	0.174***	0.122***	0.291***	0.208***	0.278***	0.213***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.004)	(0.003)	(0.001)	(0.001)
LnM	0.389***	0.268***	0.452***	0.315***	0.132***	0.091***	0.125***	0.080***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
lnFirm_Age	0.279***	0.213***	0.299***	0.224***	0.103***	0.029***	0.170***	0.097***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.007)	(0.006)	(0.002)	(0.002)
Observations	258,373	258,373	721,244	721,244	45,677	45,677	463,843	463,843
Adjusted R^2	0.360	0.433	0.383	0.461	0.229	0.327	0.257	0.380
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 11: Telecommunications (J) and Financial and Insurance Services (K) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	J.2SLS	J.OLS	J.2SLS	J.OLS	K.2SLS	K.OLS	K.2SLS	K.OLS
$\ln\hat{z}_t^{(jk)}$	1.202***		0.095***		1.366***		0.081***	
	(0.061)		(0.009)		(0.053)		(0.003)	
WAGES		0.593***		0.566***		0.592***		0.529***
		(0.009)		(0.003)		(0.008)		(0.001)
LnK	0.277***	0.138***	0.239***	0.142***	0.224***	0.092***	0.248***	0.139***
	(0.007)	(0.006)	(0.003)	(0.002)	(0.006)	(0.006)	(0.001)	(0.001)
LnM	0.207***	0.116***	0.246***	0.138***	0.202***	0.137***	0.188***	0.132***
	(0.005)	(0.005)	(0.002)	(0.002)	(0.004)	(0.004)	(0.001)	(0.001)
lnFirm_Age	0.056***	0.047***	0.144***	0.050***	0.090***	0.092***	0.124***	0.072***
	(0.013)	(0.011)	(0.005)	(0.004)	(0.012)	(0.011)	(0.002)	(0.002)
Observations	13,619	13,619	89,794	89,794	21,013	21,013	471,502	471,502
Adjusted R^2	0.267	0.430	0.266	0.461	0.237	0.384	0.242	0.435
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 12: Rental, Hiring and Real Estate Services (L) and Professional Services (M) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	L.2SLS	L.OLS	L.2SLS	L.OLS	M.2SLS	M.OLS	M.2SLS	M.OLS
$\ln\hat{z}_t^{(jk)}$	1.663***		0.225***		0.757***		0.156***	
	(0.036)		(0.004)		(0.019)		(0.002)	
WAGES		0.546***		0.404***		0.607***		0.577***
		(0.006)		(0.002)		(0.003)		(0.001)
LnK	0.278***	0.176***	0.262***	0.209***	0.215***	0.104***	0.152***	0.098***
	(0.004)	(0.004)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
LnM	0.200***	0.123***	0.204***	0.141***	0.144***	0.085***	0.165***	0.089***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.002)	(0.001)	(0.0005)	(0.0004)
lnFirm_Age	0.106***	0.076***	0.181***	0.117***	0.027***	0.008**	0.064***	0.009***
	(0.008)	(0.007)	(0.002)	(0.002)	(0.004)	(0.003)	(0.001)	(0.001)
Observations	40,665	40,665	386,405	386,405	124,096	124,096	1,298,560	1,298,560
R^2	0.288	0.391	0.260	0.356	0.178	0.400	0.161	0.427
Adjusted R^2	0.288	0.391	0.260	0.355	0.178	0.400	0.161	0.427
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 13: Administrative and Support Services (N) and Public Administration and Safety (O) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	N.2SLS	N.OLS	N.2SLS	N.OLS	O.2SLS	O.OLS	O.2SLS	O.OLS
$\ln\hat{z}_t^{(jk)}$	0.876***		0.174***		0.241***		0.118***	
	(0.036)		(0.004)		(0.079)		(0.012)	
WAGES		0.576***		0.549***		0.581***		0.563***
		(0.005)		(0.001)		(0.013)		(0.004)
LnK	0.246***	0.141***	0.222***	0.136***	0.227***	0.132***	0.224***	0.145***
	(0.004)	(0.003)	(0.001)	(0.001)	(0.009)	(0.008)	(0.003)	(0.003)
LnM	0.130***	0.069***	0.143***	0.073***	0.179***	0.105***	0.255***	0.148***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.007)	(0.006)	(0.003)	(0.002)
lnFirm_Age	0.044***	0.003	0.116***	0.043***	0.043***	−0.018	0.093***	0.025***
	(0.007)	(0.006)	(0.002)	(0.002)	(0.015)	(0.013)	(0.006)	(0.005)
Observations	43,106	43,106	441,659	441,659	6,653	6,653	56,044	56,044
Adjusted R^2	0.246	0.410	0.261	0.451	0.276	0.448	0.375	0.550
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 14: Education and Training (P) and Public Administration and Safety (Q) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	P.2SLS	P.OLS	P.2SLS	P.OLS	Q.2SLS	Q.OLS	Q.2SLS	Q.OLS
$\ln\hat{z}_t^{(jk)}$	1.227***		0.184***		0.756***		0.228***	
	(0.067)		(0.008)		(0.038)		(0.004)	
WAGES		0.592***		0.555***		0.593***		0.583***
		(0.010)		(0.002)		(0.006)		(0.001)
LnK	0.218***	0.096***	0.195***	0.110***	0.237***	0.127***	0.120***	0.069***
	(0.008)	(0.007)	(0.002)	(0.002)	(0.005)	(0.004)	(0.001)	(0.001)
LnM	0.203***	0.108***	0.265***	0.155***	0.125***	0.078***	0.186***	0.111***
	(0.005)	(0.005)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
lnFirm_Age	0.087***	0.037***	0.168***	0.041***	0.114***	0.025***	0.202***	0.012***
	(0.014)	(0.012)	(0.004)	(0.004)	(0.007)	(0.006)	(0.002)	(0.002)
Observations	10,478	10,478	181,655	181,655	35,078	35,078	629,550	629,550
Adjusted R^2	0.311	0.481	0.297	0.485	0.181	0.351	0.153	0.358
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 15: Arts and Recreation Services (R) and Other Services (S) industries results

	Balanced		*Imputed*		*Balanced*		*Imputed*	
	R.2SLS	R.OLS	R.2SLS	R.OLS	S.2SLS	S.OLS	S.2SLS	S.OLS
$\ln\hat{z}_t^{(jk)}$	1.093***		0.154***		0.247***		0.049***	
	(0.052)		(0.008)		(0.013)		(0.003)	
WAGES		0.589***		0.558***		0.498***		0.459***
		(0.008)		(0.002)		(0.003)		(0.001)
LnK	0.208***	0.113***	0.148***	0.065***	0.158***	0.101***	0.144***	0.106***
	(0.005)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
LnM	0.234***	0.152***	0.288***	0.192***	0.279***	0.179***	0.348***	0.229***
	(0.004)	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
lnFirm_Age	0.139***	0.073***	0.182***	0.062***	0.100***	0.035***	0.132***	0.054***
	(0.010)	(0.009)	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)
Observations	19,502	19,502	146,098	146,098	196,393	196,393	748,764	748,764
Adjusted R^2	0.316	0.462	0.290	0.478	0.267	0.383	0.315	0.427

Note: *p<0.1; **p<0.05; ***p<0.01

E SUMMARY STATISTICS

Table 16: Summary statistics: firm-level productivity model data

Statistic	N	$P_{1^{st}}$	$P_{50^{th}}$	$P_{99^{th}}$	St. Dev.
Balanced data					
$\ln y_{jkt}^{(firm)}$	2,296,984	7.28	10.57	13.34	1.14
$\ln \hat{z}_t^{(jk)}$	2,296,984	6.06	7.49	9.70	1.06
$\ln K_{jkt}$	2,296,984	5.01	8.88	11.60	1.27
$\ln M_{jkt}$	2,296,984	5.40	10.59	14.11	1.74
$\ln Firm\_Age$	2,296,984	0.00	1.79	2.94	0.79
WAGES	2,296,984	6.53	9.82	11.90	0.99
Imputed data					
$\ln y_{jkt}^{(firm)}$	10,039,638	7.03	10.55	13.33	1.24
$\ln \hat{z}_t^{(jk)}$	10,039,638	6.03	7.38	9.77	1.08
$\ln K_{jkt}$	10,039,638	4.41	8.69	11.76	1.48
$\ln M_{jkt}$	10,039,638	5.06	10.18	14.41	1.94
$\ln Firm\_Age$	10,039,638	0.00	1.79	2.94	0.83
WAGES	10,039,638	6.30	9.75	12.03	1.13

$\ln y_{jkt}^{(firm)}$ is logarithm of output (i.e., sales adjusted for repurchase of stock) deflated by industry gross value added implicit price deflators.

$\ln \hat{z}_t^{(jk)}$ the logarithm of estimated labour inputs.

$\ln K_{jkt}$ is the logarithm of capital that includes depreciation, capital rental expenses and capital work deductions deflated by the industry consumption of fixed capital implicit price deflators.

$\ln M_{jkt}$ is the logarithm of material costs deflated by *Producer Price Indexes: Intermediate Goods* (ABS, 2018b).

$\ln Firm\_Age$ is the logarithm of firm age. Firm age is derived as the current year minus the year of incorporation.

$\ln WAGES$ is the logarithm of wage costs (reported in Business Activities Statements) deflated by *Wage Price Index: All Industries.*

Table 17: Summary statistics: worker equation

Statistic	N	Mean	St. Dev.	Min	Max
$SKILLH$	130,281,096	0.31	0	0	1
$SKILLHM$	130,281,096	0.11	0	0	1
$SKILLM$	130,281,096	0.12	0	0	1
2003	130,281,096	0.07	0	0	1
2004	130,281,096	0.07	0	0	1
2005	130,281,096	0.07	0	0	1
2006	130,281,096	0.07	0	0	1
2007	130,281,096	0.08	0	0	1
2008	130,281,096	0.08	0	0	1
2009	130,281,096	0.08	0	0	1
2010	130,281,096	0.12	0	0	1
2011	130,281,096	0.11	0	0	1
2012	130,281,096	0.10	0	0	1
2013	130,281,096	0.09	0	0	1
AGE	130,281,096	37	37	17	64
AGE^2	130,281,096	1549	1369	289	4096
AGE^3	130,281,096	70029	50653	4913	262144
AGE^4	130,281,096	3370501	1874161	83521	16777216
$SEX : AGE$	130,281,096	19	18	0	64
$SEX : AGE^2$	130,281,096	792	324	0	4096
$SEX : AGE^3$	130,281,096	35825	5832	0	262144
$SEX : AGE^4$	130,281,096	1726528	104976	0	16777216
$SEX : 2003$	130,281,096	0.03	0	0	1
$SEX : 2004$	130,281,096	0.04	0	0	1
$SEX : 2005$	130,281,096	0.04	0	0	1
$SEX : 2006$	130,281,096	0.04	0	0	1
$SEX : 2007$	130,281,096	0.04	0	0	1
$SEX : 2008$	130,281,096	0.04	0	0	1
$SEX : 2009$	130,281,096	0.04	0	0	1
$SEX : 2010$	130,281,096	0.06	0	0	1
$SEX : 2011$	130,281,096	0.06	0	0	1
$SEX : 2012$	130,281,096	0.05	0	0	1
$SEX : 2013$	130,281,096	0.05	0	0	1

The indicator variable High Skill ($SKILLH$) equals 1 if a worker has at least a tertiary qualification and 0 otherwise.
The indicator variable Medium Skill ($SKILLHM$) equals 1 if a worker has at most a diploma qualification and 0 otherwise.
The indicator variable Working Skill ($SKILLM$) equals 1 if a worker has at most a Certificate III qualification and 0 otherwise.
There are 11 time indicator variables from 2003 to 2013, one for each year. Note that 2003 represents financial year 2002–03.
AGE is the logarithm of worker age. Worker age is derived as the current year minus the year of birth. AGE^2, AGE^3 and AGE^4 are worker age in quadratic, cubic and quartic.
$SEX : AGE$, $SEX : AGE^2$, $SEX : AGE^3$ and $SEX : AGE^4$ are the interaction terms between worker sex (SEX) and polynomial AGE.
$SEX : 2003$, ···, $SEX : 2013$ are the interaction terms between SEX and time indicator variables.

F FIRM ENTRY AND EXIT RATES

We follow Nguyen and Hansell (2014) and define firm entry rate as the number of new firms divided by the total number of incumbent and entering firms in a given year. Exit rate is defined as the number of firms exiting the market given a year and divided by the incumbents in the previous year.

Figure 11: Firm entry and exit rates

Note. ALL** represents all industries.

11 ACKNOWLEDGEMENTS

Authors would like to express our gratitude to the following ABS colleagues - Lisette Aaron, Michael Beahan, Diane Braskic, Stephen Cross, Dr Philip Gould, Rowan Hatley, Dr Sarah Hinde, Jennifer Humphrys, Jacqui Jones, Grace Kim, Katherine Keenan, Lachlan McColl, Janelle McFarlane, Sybille McKeown, Katrina Richardson, Paul Roberts, Helen Robson, Dr Siu-Ming Tam, David Taylor, Liza Tiy, Mark Viney, Hui Wei and Carter Wong for their helpful comments and support for this research, Dr David Maré and Cornell University Labor Dynamics Institute for making Fortran code available, and those who provided comments at the Business Analytic and Econometrics session of the Australian Statistical Conference 2016 that helped improve our paper. We remain solely responsible for the views expressed in this paper.

ABS Catalogue No. 1351.0.55.164

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

Reissue

Research Paper

Approaches To Analysing Micro-Drivers Of Aggregate Productivity

Methodology Transformation Branch

Methodology Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 30 OCT 2020

FOR MORE INFORMATION . . .

www.abs.gov.au the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au