Policy Research Working Paper 5346

Empirical Econometric Evaluation of Alternative Methods of Dealing with Missing Values in Investment Climate Surveys

Alvaro Escribano
Jorge Pena
J. Luis Guasch

The World Bank
Latin America and the Caribbean Region
Finance & Private Sector
June 2010

Policy Research Working Paper 5346

Abstract

Investment climate Surveys are valuable instruments that improve our understanding of the economic, social, political, and institutional factors determining economic growth, particularly in emerging and transition economies. However, at the same time, they have to overcome some difficult issues related to the quality of the information provided; measurement errors, outlier observations, and missing data that are frequently found in these datasets. This paper discusses the applicability of recent procedures to deal with missing observations in investment climate surveys. In particular, it presents a simple replacement mechanism—for application in models with a large number of explanatory variables—which in turn is a proxy of two methods: multiple imputations and an export-import algorithm. The performance of this method in the context of total factor productivity estimation in extended production functions is evaluated using investment climate surveys from four countries: India, South Africa, Tanzania, and Turkey. It is shown that the method is very robust and performs reasonably well even under different assumptions on the nature of the mechanism generating missing data.

This paper—a product of the Finance & Private Sector, Poverty Reduction and Economic Management, Latin America and the Caribbean Region—is part of a larger effort in the department to asses the determinants of productivity. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at jguasch@worldbank.org.

Empirical Econometric Evaluation of Alternative Methods of Dealing with Missing Values in Investment Climate Surveys*

by

Alvaro Escribano†, Jorge Pena‡ and J. Luis Guasch*

Keywords: Investment Climate surveys, missing observations, incomplete data, random sampling, sample selection, EM-algorithm and bootstrap.

JEL classification: C15, C24, C63, C81, C83.

* We have benefited from suggestions from Daniel Peña, Ariel Pakes, Rodolfo Stucchi and Eric Verhoogen.
† Telefonica-UC3M Chair of Economics of Telecommunications, Department of Economics, Universidad Carlos III de Madrid; alvaroe@eco.uc3m.es.
‡ Department of Economics, Universidad Carlos III de Madrid; jpizquie@eco.uc3m.es.
*Senior Adviser, World Bank, Head of the World Bank Global Experts Group on Private-Public-Partnerships, and Professor of Economics, Univeristy of California, San Diego; jguasch@worldbank.org

1. Introduction

The Investment Climate (IC) surveys (or Enterprise Surveys) have been created as part of a new strategy by the World Bank to put more emphasis on the intangible assets of developing countries such as knowledge, institutions and culture.[4] This new set of information that is becoming available to both scholars and policy makers is intended to be a valuable instrument to help improve our understanding of the economic, social, political and institutional factors determining economic growth, particularly in emerging and transition economies. However, at another level, IC surveys are also a source of trouble for researchers. In general, economic data are far from being perfect and when one is carrying out econometric or statistical analysis with a typical dataset too often we have to deal with the problem of missing values.[5] IC datasets are not an exception to this. Their imperfections make our job difficult and often even impossible (Griliches, 1986).

Incomplete data is an ubiquitous problem and standard econometric and statistical methods have nothing whatsoever to say about how to solve it. The simplest solution to this problem is to exclude from the analysis any cross-sectional observation with any missing value in it. This strategy is commonly known as *casewise deletion, listwise deletion* or *complete case analysis*. The advantage of this method lies obviously in its simplicity. The disadvantage is also rather evident to anyone who has used it: in many applications, *casewise deletion* excludes from the analysis a large fraction of the original sample. In the context of IC surveys, this is quite a high cost in terms of information lost, as well as the monetary cost arising from losing a large proportion of very expensive interviews.

The debate we wish to introduce is whether the researcher should apply some treatment on missing values when using investment climate surveys (ICSs) or rather whether it is preferable to operate with the complete case only. One of the main characteristics of the ICSs is the wide set of information they provide. Concretely, the surveys have been designed to perform a variety of economic and statistical analyses, among which especially interesting are those linking investment climate variables and several measures of firms' economic performance, such as productivity, labor demand, sales, exporting activity, FDI propensity, etc. This means having matrices of data with a remarkably large number of rows and therefore the possibility of using econometric models with a wide set of right hand side variables. Unfortunately, in many cases the problem of missing data is so serious that it prevents us from using those kinds of models. In some of these cases the missingness problem reduces the cross sectional observations available in the complete case to even 0% of the original sampling frame.[6] Should the researcher therefore

[4] Key determinants of the investment climate, which are included and properly measured in the Investment Climate (IC) series of surveys, include physical and institutional infrastructure, economic and political stability, rule of law, infrastructure, approaches to regulations and taxes, functioning of labor and finance markets, and broader features of governance, such as corruption. The World Bank group has long been a supporter of investment climate reform, recognizing the importance of shaping a business environment conducive to the successful start-up and operation of firms of all sizes in all sectors.

[5] Information is missing for various reasons. A sizeable fraction of the respondents refuse, forget or fail to answer some questions. In other cases, even well-trained interviewers may neglect to ask some questions. Sometimes respondents just say they do not have the information available to them or they do not know the answer to the question. Some questions are simply not applicable to some respondents (see Allison, 2001). All of these cases may be applicable to IC data.

[6] The number of observations available in the complete case decreases as we consider more and more investment climate variables. If we consider all the variables included in the survey, the complete case due to missing cells is

limit himself to using models with a reduced number of independent variables with the risk of introducing a more serious omitted variables problem? Or is it preferable to impute missing data in order to be able to use structural models with a wide set of explanatory variables? If we assume the latter as a reasonable solution, the question that arises then is: should we input missing cells in both LHS (independent) and RHS (dependent) variables, or, on the contrary, ,should we satisfy ourselves by replacing missing data in only those explanatory variables of the model?

During recent years statisticians have proposed many alternative methods to handle incomplete datasets that offer substantial improvements over *casewise deletion*. These approaches may be grouped into two families of methods: maximum likelihood and multiple imputation, see Allison (2001), Meng (2000) and Little and Rubin (1987) for a review. However, these methods depend on easily violated assumptions that, to make things worse, are difficult or even impossible to test. In this paper we discuss the applicability of these methods to four IC surveys with very different patterns of missing data among them: India, Turkey, South Africa and Tanzania. In particular, we propose a simple imputation mechanism (which we call the *ICA method*) that in part departs from the EM-algorithm, and that has been widely applied in various empirical works (Escribano et al, 2008a, b; Escribano, Guasch and Pena 2009 and Escribano et al. 2009). We compare the performance of this method with several alternative approaches to deal with incomplete data and we discuss the different assumptions we need to hold for the different imputation mechanisms to work well. We evaluate the validity of the different methods in the context of the extended production function of Escribano and Guasch (2005 and 2008).[7] The extended production function framework used here fits very well with the objective of the paper as the RHS of the equation is compounded by a broad set of explanatory variables.[8] On the other hand, although we concentrate on PF variables, the results of the analysis can be easily extended to any variable with missing information included in the ICSs.

We demonstrate that, besides the imputation method used, a detailed knowledge of the missingness mechanism in the context of ICSs is a requisite. The missing data problem is at the core of statistical and econometric analysis done with ICSs and therefore a proper treatment of the missing data mechanism is inevitable. We also show that the so-called *ICA method* proposed performs reasonably well, even under very different patterns of missing data. The differences of the *ICA method* with respect to other more sophisticated imputation mechanisms, such as EM algorithms, multiple imputation, bootstrap methods or Heckman models, are not remarkably significant, so we propose it as a benchmark, a homogeneous, simple and easy to implement method for models with large numbers of covariates in ICSs, and more importantly, for very complex and unbalanced patterns of missing data.

The structure of the paper is as follows. In section 2 we review the patterns of missing data observed in the four IC surveys considered. We compare the original sampling frame with the complete case and we see that in most cases the representativity of the original sample is

---

0% in most cases. However, if we construct models using only those investment climate variables with a response rate higher than 80%, the complete case increases from 20% to 30% of the original sampling frame.

[7] Although it is straightforward to apply this method to any kind of model, especially those involving a large number of RHS variables or structural system of equations.

[8] The underlying philosophy of the Escribano and Guasch (2005 and 2008) extended production function is to incorporate in a Cobb-Douglas (or Translog) function a large set of investment climate variables to correct for observable fixed effects.

modified and the total number of observations available for regression analysis is considerably reduced. We compare these numbers with the observations available after the replacement mechanism we propose. Section 3 presents the *ICA method* and other imputation mechanism used as comparators. We also comment on the different assumptions underlying the different methods proposed. We discuss to what extent the missing data mechanism (MDM) presented in the four surveys analyzed may be considered as missing completely at random (MCAR), missing at random (MAR), or non-ignorable. Section 4 shows the regression results for the extended production function under the different replacement methods. Finally, section 5 concludes. All tables and figures are included in an extensive appendix at the end of the paper.

2. Missing data and investment climate surveys

We introduce the problem at hand with Table 1.1 (see appendix tables and figures) which shows the total number of observations, the observations available in the complete case and the final number of observations we have after the replacement process we propose—which we discuss later on—in 43 different ICSs. All the surveys share similar characteristics in the sampling procedure applied and, more importantly, in the information provided. The number of observations lost varies among all the surveys considered. The replacement process considerably increases the sample size in all cases (the method is described in section 3).[9] The problem of incomplete data is common to all the IC surveys considered, although it is more persistent in countries like Thailand, Niger, Paraguay, Tanzania and Turkey, in which the percentage of observations available in the complete case is below 30%.[10] In Table 1.1 we only consider missing values in production function variables. When we consider all variables likely to be used in regression analysis (all investment climate variables), the complete case even reduces to 0% in some cases.[11]

[9] The sample with replacement fills missing values of all variables of the survey (both production function and IC variables).

[10] By means of simplification we understand by *complete case* the sample with replacement only in IC variables.

[11] As said, the problem of missing data is, to a lesser or greater extent, common to almost all the variables presented in the IC surveys. We here consider the missingness and its treatment in production function variables (sales,

Table 1.1: Observations available for regression analysis after and before imputing missing values and outliers in 43 ICSs

		Year of the survey	Obs. In the sampling frame	Complete case		After imputing missing cells	
				#Obs.	% with respect to sampling frame	#Obs.	% with respect to sampling frame
Latin America	Argentina	2006	746	372	*49.9*	664	*89.0*
	Bolivia	2006	409	209	*51.1*	336	*82.2*
	Colombia	2006	649	525	*80.9*	618	*95.2*
	Mexico	2006	1,161	778	*67.0*	1,093	*94.1*
	Panama	2006	243	97	*39.9*	223	*91.8*
	Peru	2006	361	230	*63.7*	337	*93.4*
	Paraguay	2006	440	111	*25.2*	315	*71.6*
	Uruguay	2006	396	155	*39.1*	304	*76.8*
	Chile	2006	697	382	*54.8*	629	*90.2*
	Costa Rica	2005	1029	643	*62.5*	970	*94.3*
	Ecuador	2006	394	235	*59.6*	346	*87.8*
	El Salvador	2006	467	296	*63.4*	439	*94.0*
	Honduras	2006	263	189	*71.9*	243	*92.4*
	Guatemala	2006	328	262	*79.9*	316	*96.3*
	Nicaragua	2006	365	230	*63.0*	341	*93.4*
Africa	Algeria	2002	1,904	1,114	*58.5*	1,412	*74.2*
	Benin	2004	591	364	*61.6*	475	*80.4*
	Botswana	2006	114	109	*95.6*	113	*99.1*
	Cameroon	2006	119	117	*98.3*	118	*99.2*
	Egypt	2004	2,931	1,317	*44.9*	2,629	*89.7*
	Eritrea	2002	237	61	*25.7*	179	*75.5*
	Ethiopia	2002	1,281	1,048	*81.8*	1,142	*89.1*
	Kenya	2003	852	360	*42.3*	585	*68.7*
	Madagascar	2005	870	383	*44.0*	623	*71.6*
	Malawi	2005	320	208	*65.0*	288	*90.0*
	Mali	2003	462	242	*52.4*	309	*66.9*
	Mauritius	2005	636	271	*42.6*	417	*65.6*
	Morocco	2003	2,550	2,352	*92.2*	2,422	*95.0*
	Namibia	2006	106	100	*94.3*	104	*98.1*
	Senegal	2003	783	253	*32.3*	535	*68.3*
	South Africa*	2003	1,737	1,229	*70.8*	1,492	*85.9*
	Tanzania*	2003	828	325	*39.3*	561	*67.8*
	Uganda	2003	900	368	*40.9*	695	*77.2*
	Zambia	2002	564	391	*69.3*	417	*73.9*
Asia	Indonesia	2003	1,214	486	*40.0*	1,041	*85.7*
	Malaysia	2001	1,732	605	*34.9*	1,317	*76.0*
	Philippines	2003	1,432	1,092	*76.3*	1,272	*88.8*
	Thailand	2004	2,766	646	*23.4*	1,502	*54.3*
	Pakistan	2007	2358	990	*42.0*	2,144	*90.9*
	Bangladesh	2006	4804	2,533	*52.7*	3,946	*82.1*
	India*	2005	6849	4448	*64.9*	5750	*84.0*
Europe	Croatia	2007	419	219	*52.3*	372	*88.8*
	Turkey*	2005	2646	771	*29.1*	1,619	*61.2*

Complete case includes those observations without missing values and or outliers in sales, materials, capital, labor cost and labor
Source: Authors' calculations with IC data.

We focus the analysis on the investment climate surveys of India, Turkey, South Africa and Tanzania because they represent almost all the situations regarding the structure of missing data we may find.[12] For India, in the complete case we lose 35% of the original sampling frame, while after replacing we only lose 16%. Turkey and Tanzania lose a similar percentage of observations, 70.9% and 60.7% respectively. South Africa only loses 29.2%.

Table 1.2 looks in depth at the description of the missingness problem of the four countries selected. In this case, for the computation of the observations available in the complete case, we use all those IC variables included in the survey likely to be used in a regression analysis framework. This means using more than 115 variables in India, 90 in Turkey, 168 in South Africa and 162 in Tanzania. For each country we consider two benchmark cases: the first one includes both PF and IC variables in the computation of the complete case, while the second only considers the IC variables. In the extreme case, when we consider all those IC variables, the complete case reduces to 0% of the complete case in all the countries; it doesn't matter whether we include PF or not. Note that the observations available in the complete case increase as we exclude from the computation of the complete case those IC variables with the largest proportion of empty cells reported. In order to have a large enough number of observations we would need to exclude from the analysis those IC variables with a response rate lower than 95%. Even in this case, and also considering the PF variables, we should be forced to exclude 41.1% of the interviews in India, 76.9% in Turkey, 60.2% in South Africa and 66.2% in Tanzania. The evidence concerning the size of the problem of missing information we have to deal with is overwhelming.

materials, capital and employment), although all we say about imputing missing information in production function variables can be easily extended to any other IC variable.

[12] These datasets have in turn been analyzed in the following works: Escribano, Guasch and de Orte (2009) for India, Escribano, Guasch, de Orte and Pena (2008b and c) for the case of Turkey and Escribano, Guasch and Pena (2009) for South Africa and Tanzania.

Table 1.2: Missing values in IC variables and their incidence on complete case

A. India

IC variables included	# variables	[1]		[2]	
		# obs. Available	% over total	# obs. Available	% over total
All IC variables [(a)]	115	0	0.0	0	0.0
those IC vars. with response rate >70% [(b)]	80	500	7.3	588	8.6
those IC vars. with response rate >80% [(c)]	71	942	13.8	1188	17.3
those IC vars. with response rate >90% [(d)]	63	1663	24.3	2202	32.2
those IC vars. with response rate >95% [(e)]	40	2109	30.8	2817	41.1

B. Turkey

IC variables included	# variables	[1]		[2]	
		# obs. Available	% over total	# obs. Available	% over total
All IC variables [(a)]	90	1	0.0	4	0.2
those IC vars. with response rate >70% [(b)]	78	426	16.1	740	28.0
those IC vars. with response rate >80% [(c)]	77	472	17.8	1226	46.3
those IC vars. with response rate >90% [(d)]	75	523	19.8	1394	52.7
those IC vars. with response rate >95% [(e)]	65	697	26.3	2034	76.9

C. South Africa

IC variables included	# variables	[1]		[2]	
		# obs. Available	% over total	# obs. Available	% over total
All IC variables [(a)]	168	0	0.0	0	0.0
those IC vars. with response rate >70% [(b)]	112	93	5.1	114	6.3
those IC vars. with response rate >80% [(c)]	108	391	21.6	451	24.9
those IC vars. with response rate >90% [(d)]	92	620	34.3	769	42.5
those IC vars. with response rate >95% [(e)]	81	828	45.8	1089	60.2

D. Tanzania

IC variables included	# variables	[1]		[2]	
		# obs. Available	% over total	# obs. Available	% over total
All IC variables [(a)]	162	0	0.0	0	0.0
those IC vars. with response rate >70% [(b)]	98	6	0.7	9	1.1
those IC vars. with response rate >80% [(c)]	89	32	3.9	69	8.3
those IC vars. with response rate >90% [(d)]	71	118	14.3	251	30.3
those IC vars. with response rate >95% [(e)]	40	227	27.4	548	66.2

[1] PF variables are also included In the computation of the final number of observations available in the complete case.
[2] PF variables are not included In the computation of the final number of observations available in the complete case.
[(a)] All IC variables are included in the computation of the number of observations available in the complete case.
[(b)] Only those IC variables with a response rate higher than 70% are included in the computation of the number of observations available in the complete case.
[(c)] Only those IC variables with a response rate higher than 80% are included in the computation of the number of observations available in the complete case.
[(d)] Only those IC variables with a response rate higher than 90% are included in the computation of the number of observations available in the complete case.
[(e)] Only those IC variables with a response rate higher than 80% are included in the computation of the number of observations available in the complete case.
Source: Authors' estimation with ICSs.

In the remaining part of this section we first present the pattern of missing values observed in the four surveys considered. We also evaluate the representativity of the sample with replacement and the complete case with respect to the sampling frame.

2.1 Sampling and characteristics of the ICSs

The sampling of the ICSs is based on a World Bank template used in a large number of countries and customized in collaboration with regional statistical agencies to reflect country-specific issues and policy areas of interest. In order to ensure proper representation of the sectors of interest,[13] respondents are carefully selected. The sampling process is normally based on national industry databases and census of firms or establishments,[14] which provide the necessary information on the particular population of establishments. To ensure proper representation of firms, stratification is usually done based on three standards: size, sector and location.[15]

The information contained in the ICSs is composed of a wide set of around 400 variables. Eventually, the number of variables likely to be used in regression analysis is reduced to around 120-200.[16] The Investment Climate Surveys provide information regarding firms' experience in a range of areas related to economic performance: financing, governance, corruption, crime, regulation, tax policy, labor relations, conflict resolution, infrastructures, supplies and marketing, quality, technology, and training among others. The ICSs also provide information on the productivity (or production function) variables, sales output (sales are used as measure of output), employment, intermediate materials, capital stock and labor cost. The resulting panel information is short in the time dimension, since it includes only 2 or 3 years of productivity data (in our case 2 years for Turkey and 3 for India, South Africa and Tanzania), and has 1 year of information for the investment climate variables. Finally, it is important to note that all information is based on recall data and not on book values or accounting.

2.2 The missing information problem at first glance

Figures 1.1 to 1.4 show the complex and unbalanced patterns of missing values observed in the PF variables in the four countries considered. The most common case is finding observations with information for all the PF but one. In India, the percentage of establishments reporting information for all the PF variables except capital is 16.3%. In the rest of the countries, this percentage is slightly lower but significantly high too. It is less common to observe data on all the PF variables except sales, materials or employment, although in Tanzania the percentage of firms reporting all the figures except sales is relatively important, 9.8%. The cases for which data is collected for only two PF variables represent, in all the countries, less than 1% of total data. Finally, it is very common to have data collected only for labor; this percentage represents 13.3% in India, 27.9% in Turkey, 5.5% in South Africa and 15.7% in Tanzania.

[13] Here we focus only on the manufacturing sector. By classifying the establishments by their ISIC code we generally end up with establishments from the following eight sectors: a) Food and beverages; b) Textiles and apparel; c) Chemicals; d) Non-metallic mineral products; e) Metallic products; f) Machinery and equipment; g) Electrical machinery; h) Transport equipment.

[14] The unit of reference in the ICSs is the establishment, although in this paper we refer indistinctively to both establishments and firms.

[15] Concretely, the establishments are selected according to a random sampling by industry and region. Taking into account this issue we use standard errors allowing for clustering by industry and region (apart from the conventional correction for heteroskedasticity *a la* White). In some surveys there is also oversampling of large firms.

[16] We understand by *"likely to be used in regression analysis"* all those variables describing the investment climate in which firms operate and likely to be related to firms' economic performance.

Figure 1.1: INDIA, Patterns of missing values in PF variables

Sales	Materials	Capital	Labor	# of m.v	# of obs.	% of obs.
				0	4631	67.6
				1	1113	16.3
				3	913	13.3
				2	89	1.3
				2	47	0.7
				2	28	0.4
				1	18	0.3
				1	10	0.1

Figure 1.2: TURKEY, Patterns of missing values in PF variables

Sales	Materials	Capital	Labor	# of m.v	# of obs.	% of obs.
				0	818	30.9
				3	737	27.9
				1	345	13.0
				2	189	7.1
				1	185	7.0
				2	133	5.0
				4	96	3.6
				1	87	3.3
				2	35	1.3
				3	6	0.2
				2	5	0.2
				3	5	0.2
				1	3	0.1
				2	2	0.1

Figure 1.3: SOUTH AFRICA, Patterns of missing values in PF variables

Sales	Materials	Capital	Labor	# of m.v	# of obs.	% of obs.
				0	1265	69.9
				1	220	12.2
				4	123	6.8
				3	99	5.5
				1	47	2.6
				2	24	1.3
				1	17	0.9
				2	7	0.4
				2	4	0.2
				1	1	0.1
				2	1	0.1
				3	1	0.1

Figure 1.4: TANZANIA, Patterns of missing values in PF variables

Sales	Materials	Capital	Labor	# of m.v	# of obs.	% of obs.
				0	313	37.8
				3	130	15.7
				1	81	9.8
				1	74	8.9
				1	51	6.2
				1	38	4.6
				4	37	4.5
				2	30	3.6
				2	26	3.1
				2	25	3.0
				3	9	1.1
				2	5	0.6
				2	3	0.4
				3	3	0.4
				2	2	0.2
				3	1	0.1

Notes:
Yellow means information available on the corresponding variable. White means information is missing.
Source: Authors' calculations with ICS data.

Tables from 2.1 to 2.4 of the appendix show the distribution of the number of observations available in the original sampling frame, in the complete case and in the sample with replacement, along with the percentage of observations lost with respect to the original sampling frame. From Table 2.1 the percentage of observations lost in India in the complete case varies when we move industry by industry and size by size. Flagrant cases of loss of observations are small firms operating in the non-metallic products sector (61.9%) or the medium-sized firms of the food sector (55.37%). The replacement process allows retrieving for the analysis a considerable percentage of observations. After the replacement we only lost 28.6% and 22.6% in the two cells mentioned previously. In Turkey the percentages of observations lost by size and industry (see Table 2.2) range from 40% (medium-sized firms in the transport equipment sector) to 87.3% (small firms in textiles and apparels industry). South Africa lost 50% of small firms in textiles and apparel and chemical, rubber and plastics sectors (see Table 2.3). Lastly, Tanzania lost more than 70% of small firms in paper, edition and publishing and machinery and equipment and 73% of large firms in textiles and apparels (Table 2.4).

Table 2.1: INDIA, Percentage of observations lost due to missing values by industry and size

	Size	Small		Medium		Large		Total	
Industry		#Obs	%Lost[(d)]	#Obs	%Lost	#Obs	%Lost	#Obs	%Lost
Food	Sampling frame[(a)]	333		177		87		597	
	Complete case[(b)]	177	46.9	79	55.4	51	41.4	307	48.6
	With replacement[(c)]	248	25.5	137	22.6	69	20.7	454	24
Textiles & Leather	Sampling frame	426		255		207		888	
	Complete case	251	41.1	210	17.7	139	32.9	600	32.4
	With replacement	325	23.7	235	7.8	178	14	738	16.9
Apparel	Sampling frame	360		315		150		825	
	Complete case	247	31.4	267	15.2	120	20	634	23.2
	With replacement	287	20.3	290	7.9	138	8	715	13.3
Chemicals & Chemical prds	Sampling frame	426		333		171		930	
	Complete case	262	38.5	218	34.5	130	24	610	34.4
	With replacement	337	20.9	282	15.3	150	12.3	769	17.3
Plastics & Rubbers	Sampling frame	279		189		12		480	
	Complete case	193	30.8	112	40.7	11	8.3	316	34.2
	With replacement	243	12.9	157	16.9	11	8.3	411	14.4
Non-metallic products	Sampling frame	105		63		48		216	
	Complete case	40	61.9	38	39.7	32	33.3	110	49.1
	With replacement	75	28.6	50	20.6	39	18.8	164	24.1
Structural metal & metal prds	Sampling frame	618		252		39		909	
	Complete case	328	46.9	131	48	21	46.2	480	47.2
	With replacement	526	14.9	214	15.1	31	20.5	771	15.2
Machinery & Equipment	Sampling frame	1074		687		243		2004	
	Complete case	749	30.3	482	29.8	160	34.2	1,391	30.6
	With replacement	912	15.1	603	12.2	213	12.4	1728	13.8
Total	Sampling frame	3621		2271		957		6849	
	Complete case	2,247	38	1,537	32.3	664	30.6	4,448	35.1
	With replacement	2953	18.5	1968	13.3	829	13.4	5750	16.1

Notes:
[(a)] "Sampling frame" refers to the total number of observations (firms surveyed multiplied by the number of years of information).
[(b)] "Complete case" refers to the complete case of production function variables (sales, materials, capital and labor), missing values in other IC variables—other than production function— are not considered.
[(c)] "With replacement" refers to the sample after imputing IC variables according to the ICA Method; missing values in other IC variables—other than production function—are not considered. Notice that only observations with information available in at least one of sales, labor, labor cost, materials or capital, are imputed
[(d)] "Perc. lost" refer to the percentage of observations lost with respect to the sampling frame.
Source: Authors calculations with IC data.

Table 2.2: TURKEY, Percentage of observations lost due to missing values by industry and size

	Size	Small		Medium		Large		Total	
Industry		#Obs	%Lost[(d)]	#Obs	%Lost	#Obs	%Lost	#Obs	%Lost
Food and Beverages	Sampling frame[(a)]	192		170		202		564	
	Complete case[(b)]	56	70.8	57	66.5	82	59.4	195	65.4
	With replacement[(c)]	134	30.2	116	31.8	150	25.7	400	29.1
Textiles and Apparel	Sampling frame	110		230		398		738	
	Complete case	14	87.3	47	79.6	115	71.1	176	76.2
	With replacement	48	56.4	130	43.5	257	35.4	435	41.1
Chemicals	Sampling frame	118		98		136		352	
	Complete case	24	79.7	29	70.4	51	62.5	104	70.5
	With replacement	60	49.2	67	31.6	87	36.0	214	39.2
Non-metallic mineral products	Sampling frame	54		66		46		166	
	Complete case	15	72.2	20	69.7	19	58.7	54	67.5
	With replacement	46	14.8	51	22.7	30	34.8	127	23.5
Metal products (ex. M&E)	Sampling frame	94		98		92		284	
	Complete case	30	68.1	43	56.1	34	63.0	107	62.3
	With replacement	68	27.7	82	16.3	59	35.9	209	26.4
Machinery and Equipment	Sampling frame	98		78		80		256	
	Complete case	37	62.2	31	60.3	38	52.5	106	58.6
	With replacement	79	19.4	52	33.3	63	21.3	194	24.2
Electrical machinery	Sampling frame	58		40		36		134	
	Complete case	19	67.2	19	52.5	15	58.3	53	60.4
	With replacement	42	27.6	34	15.0	24	33.3	100	25.4
Transport equipment	Sampling frame	64		30		58		152	
	Complete case	31	51.6	18	40.0	15	74.1	64	57.9
	With replacement	54	15.6	25	16.7	46	20.7	125	17.8
Total	Sampling frame	788		810		1048		2646	
	Complete case	226	71.3	264	67.4	369	64.8	859	67.5
	With replacement	531	32.6	557	31.2	716	31.7	1804	31.8

Notes:
[(a)] "Sampling frame" refers to the total number of observations (firms surveyed multiplied by the number of years of information).
[(b)] "Complete case" refers to the complete case of production function variables (sales, materials, capital and labor), missing values in other IC variables—other than production function— are not considered.
[(c)] "With replacement" refers to the sample after imputing IC variables according to the ICA Method; missing values in other IC variables—other than production function—are not considered. Notice that only observations with information available on at least one of sales, labor, labor cost, materials or capital, are imputed
[(d)] "Perc. lost" refers to the percentage of observations lost with respect to the sampling frame.
Source: Authors calculations with IC data.

Table 2.3: SOUTH AFRICA, Percentage of observations lost due to missing values by industry and size

	Size	Small		Medium		Large		Total	
Industry		#Obs	%Lost[(d)]	#Obs	%Lost	#Obs	%Lost	#Obs	%Lost
Food & beverages	Sampling frame[(a)]	22		80		87		189	
	Complete case[(b)]	13	40.9	49	38.8	69	20.7	131	30.7
	With replacement[(c)]	14	36.4	66	17.5	82	5.7	162	14.3
Textiles & apparel	Sampling frame	12		43		120		175	
	Complete case	6	50	32	25.6	69	42.5	107	38.9
	With replacement	10	16.7	33	23.3	101	15.8	144	17.7
Chemicals, rubber & plastics	Sampling frame	42		119		118		279	
	Complete case	21	50	79	33.6	87	26.3	187	33
	With replacement	29	31	111	6.7	101	14.4	241	13.6
Paper, edition & publishing	Sampling frame	13		89		54		156	
	Complete case	10	23.1	65	27	45	16.7	120	23.1
	With replacement	10	23.1	78	12.4	49	9.3	137	12.2
Machinery & equipment	Sampling frame	47		252		256		555	
	Complete case	25	46.8	198	21.4	212	17.2	435	21.6
	With replacement	35	25.5	222	11.9	241	5.9	498	10.3
Wood & furniture	Sampling frame	13		74		58		145	
	Complete case	7	46.2	55	25.7	39	32.8	101	30.3
	With replacement	11	15.4	69	6.8	50	13.8	130	10.3
Non-metallic products	Sampling frame	13		23		30		66	
	Complete case	3	76.9	18	21.7	22	26.7	43	34.8
	With replacement	6	53.8	18	21.7	26	13.3	50	24.2
Other	Sampling frame	27		63		57		147	
	Complete case	19	29.6	38	39.7	47	17.5	104	29.3
	With replacement	25	7.4	50	20.6	51	10.5	126	14.3
Total	Sampling frame	189		743		780		1712	
	Complete case	104	45	534	28.1	590	24.4	1228	28.3
	With replacement	140	25.9	647	12.9	701	10.1	1488	13.1

Notes:

[(a)] "Sampling frame" refers to the total number of observations (firms surveyed multiplied by the number of years of information).

[(b)] "Complete case" refers to the complete case of production function variables (sales, materials, capital and labor), missing values in other IC variables—other than production function— are not considered.

[(c)] "With replacement" refers to the sample after imputing IC variables according to the ICA Method; missing values in other IC variables—other than production function—are not considered. Notice that only observations with information available in at least one of sales, labor, labor cost, materials or capital, are imputed

[(d)] "Perc. lost" refers to the percentage of observations lost with respect to the sampling frame.

Source: Authors calculations with IC data.

Table 2.4: TANZANIA, Percentage of observations lost due to missing values by industry and size

	Size	Small		Medium		Large		Total	
Industry		#Obs	%Lost[d]	#Obs	%Lost	#Obs	%Lost	#Obs	%Lost
Food & beverages	Sampling frame[a]	105		87		51		243	
	Complete case[b]	47	55.2	44	49.4	17	66.7	108	55.6
	With replacement[c]	82	21.9	57	34.5	31	39.2	170	30
Textiles & apparel	Sampling frame	33		41		19		93	
	Complete case	10	69.7	14	65.9	5	73.7	29	68.8
	With replacement	26	21.2	24	41.5	8	57.9	58	37.6
Chemicals, rubber & plastics	Sampling frame	23		55		24		102	
	Complete case	10	56.5	18	67.3	14	41.7	42	58.8
	With replacement	13	43.5	40	27.3	16	33.3	69	32.4
Paper, edition & publishing	Sampling frame	27		39		9		75	
	Complete case	8	70.4	19	51.3	6	33.3	33	56
	With replacement	16	40.7	30	23.1	9	0	55	26.7
Machinery & equipment	Sampling frame	49		29		9		87	
	Complete case	14	71.4	6	79.3	6	33.3	26	70.1
	With replacement	36	26.5	21	27.6	8	11.1	65	25.3
Wood & furniture	Sampling frame	133		53		9		195	
	Complete case	52	60.9	13	75.5	3	66.7	68	65.1
	With replacement	89	33.1	23	56.6	5	44.4	117	40
Non-metallic products	Sampling frame	11		16		6		33	
	Complete case	3	72.7	11	31.3	5	16.7	19	42.4
	With replacement	9	18.2	12	25	6	0	27	18.2
Total	Sampling frame	381		320		127		828	
	Complete case	144	62.2	125	60.9	56	55.9	325	60.7
	With replacement	271	28.9	207	35.3	83	34.6	561	32.2

Notes:

[a] "Sampling frame" refers to the total number of observations (firms surveyed multiplied by the number of years of information).

[b] "Complete case" refers to the complete case of production function variables (sales, materials, capital and labor), missing values in other IC variables—other than production function— are not considered.

[c] "With replacement" refers to the sample after imputing IC variables according to the ICA Method; missing values in other IC variables—other than production function—are not considered. Notice that only observations with information available in at least one of sales, labor, labor cost, materials or capital, are imputed

[d] "Perc. lost" refers to the percentage of observations lost with respect to the sampling frame.

Source: Authors calculations with IC data.

Tables 3.1, 3.2, 3.3 and 3.4 attempt to illustrate how the representativity of the sampling frame changes with respect to the complete case and the sample with replacement.[17] In all cases, the percentages vary slightly in the complete case with respect to the sampling frame. The percentages of the sample with replacement are more similar to the sampling frame. For instance, in India from Table 3.1, panel a), the percentage of '*food*' firms falls from 8.7% to 6.9%, while after the replacement it is 7.9%. Symmetrically, the percentage of '*apparel*' firms jumps from 12% to 14.3% in the complete case and to 12.4% in the sample with replacement. Similar patterns can be observed in the remaining countries. Finally, from these tables response rates do differ across countries, but within countries they are remarkably uniform across regions and industries.

Table 3.1: INDIA, Representativity of sampling frame, complete case and sample with replacement

	Sampling frame[(a)]		Complete case[(b)]		With replacement[(c)]	
	# Obs	Perc over total	# Obs	Perc over total	# Obs	Perc over total
a) by Industry						
Food	597	8.7	307	6.9	454	7.9
Textiles & Leather	888	13	600	13.5	738	12.8
Apparel	825	12	634	14.3	715	12.4
Chemicals & Chemical prds	930	13.6	610	13.7	769	13.4
Plastics & Rubbers	480	7	316	7.1	411	7.1
Non-metallic products	216	3.2	110	2.5	164	2.9
Structural metal & metal prds	909	13.3	480	10.8	771	13.4
Machinery & Equipment	2,004	29.3	1,391	31.3	1,728	30.1
Total	6,849	100	4,448	100	5,750	100
b) by size						
Small	3,621	52.9	2,247	50.5	2,953	51.4
Medium	2,271	33.2	1,537	34.6	1,968	34.2
Large	957	14	664	14.9	829	14.4
Total	6,849	100	4,448	100	5,750	100

Notes:
[(a)] "Sampling frame" refers to the total number of observations (firms surveyed multiplied by the number of years of information).
[(b)] "Complete case" refers to the complete case of production function variables (sales, materials, capital and labor), missing values in other IC variables—other than production function— are not considered.
[(c)] "With replacement" refers to the sample after imputing IC variables according to the ICA Method; missing values in other IC variables—other than production function—are not considered. Notice that only observations with information available in at least one of sales, labor, labor cost, materials or capital, are imputed
Source: Authors calculations with ICSs data.

[17] In order to evaluate how representativity changes from the sampling frame to the complete case, we would need to have information on the weight of each category over the reference population. Unfortunately, this information is not available. As second best, we can still demonstrate how representativity changes from the data we have. Let us suppose population is split into two strata, and that the original sample selects a given number of observations for strata 1 and 2, and as a result X and Y are the percentages that represent the weight of each strata in the population. In the complete case, we introduce the missing data problem so instead of X and Y we have X′, Y′. If we suppose that the sampling frame is representative of the population then the complete case is said to be representative if, and only if, the weights in the complete case are proportional to the weights in the sampling frame; that is $X \approx X'$ and $Y \approx Y'$.

Table 3.2: TURKEY, Representativity of sampling frame, complete case and sample with replacement

	Sampling frame[a]		Complete case[b]		With replacement[c]	
	# Obs	Perc over total	# Obs	Perc over total	# Obs	Perc over total
a) by Industry						
Food and Bev.	564	21.3	195	22.7	400	22.2
Textiles and Apparel	738	27.9	176	20.5	435	24.1
Chemicals	352	13.3	104	12.1	214	11.9
Non-metallic mineral products	166	6.3	54	6.3	127	7.0
Metal products (ex. M&E)	284	10.7	107	12.5	209	11.6
Machinery and Equipment	256	9.7	106	12.3	194	10.8
Electrical machinery	134	5.1	53	6.2	100	5.5
Transport equipment	152	5.7	64	7.5	125	6.9
Total	2,646	100	859	100.0	1,804	100.0
b) by size						
Small	788	29.8	226	26.3	531	29.4
Medium	810	30.6	264	30.7	557	30.9
Large	1048	39.6	369	43.0	716	39.7
Total	2,646	100.0	859	100.0	1,804	100.0

Notes:
Same as Table 3.1.

Table 3.3: SOUTH AFRICA, Representativity of sampling frame, complete case and sample with replacement

	Sampling frame[a]		Complete case[b]		With replacement[c]	
	# Obs	Perc over total	# Obs	Perc over total	# Obs	Perc over total
a) by Industry						
Food & beverages	189	10.9	131	10.7	159	10.7
Texts & apparel	180	10.4	107	8.7	143	9.6
Chemicals rubber & plastics	285	16.4	187	15.2	241	16.2
Paper, edition & publishing	159	9.2	120	9.8	137	9.2
Machinery & equipment	561	32.3	435	35.4	497	33.4
Wood & furniture	147	8.5	102	8.3	131	8.8
Non-metallic products	66	3.8	43	3.5	49	3.3
Other	150	8.6	104	8.5	129	8.7
Total	1,737	100	1,229	100	1,486	100
b) by size						
Small	189	11	104	8.5	139	9.4
Medium	743	43.4	534	43.5	647	43.7
Large	780	45.6	590	48	696	47
Total	1,712	100	1,228	100	1,482	100

Notes:
Same as Table 3.1.

Table 3.4: TANZANIA, Representativity of sampling frame, complete case and sample with replacement

	Sampling frame[(a)]		Complete case[(b)]		With replacement[(c)]	
	# Obs	Perc over total	# Obs	Perc over total	# Obs	Perc over total
a) by Industry						
Food & beverages	243	29.3	108	33.2	170	30.3
Textiles & apparel	93	11.2	29	8.9	58	10.3
Chemicals, rubber & plastics	102	12.3	42	12.9	69	12.3
Paper, edition & publishing	75	9.1	33	10.2	55	9.8
Machinery & equipment/Metallic products	87	10.5	26	8	65	11.6
Wood & furniture	195	23.6	68	20.9	117	20.9
Non-metallic products	33	4	19	5.8	27	4.8
Total	828	100	325	100	561	100
b) by size						
Small	381	46	144	44.3	271	48.3
Medium	320	38.6	125	38.5	207	36.9
Large	127	15.3	56	17.2	83	14.8
Total	828	100	325	100	561	100

Notes:
Same as Table 3.1.

3. Imputation of missing values: The *ICA method*

Rubin (1976) rigorously defined the assumptions that might plausibly be made about missing data mechanisms (MDM).[18] When the MDM is ignorable, the objective of the replacement methods is not to augment the sample size, but to preserve the sample representativity, to gain efficiency in the estimation and to retrieve for the analysis a large number of very expensive interviews. The alternative to these methods is the *listwise deletion*, which is not a panacea even when the MDM is ignorable. Operating with the complete case is only acceptable if incomplete cases attributable to missing data comprise a small percentage, say 5% or less, of the number of total cases (Schafer, 1997), and when the complete case preserves the representativeness of the original sampling frame. In addition, in models with a large number of regressors, missing data problems may encourage analysts to leave out of the regression some explanatory variables with a high proportion of missing values. As Cameron and Trivedi (2005) point out, this practice may be misleading as it leads to an omitted variables problem, which is more serious than the missing data problem *per se*.

To see how the various mechanisms applied to deal with missing data perform, it is useful to depart from a population model of interest. A repeated task that applied researchers carry out in the context of IC data is the estimation of production functions to perform a variety of productivity analyses. Concretely, let us suppose the extended production function as in Escribano and Guasch (2005 and 2008). The population model is given by

$$\log Y_{it} = \alpha_0 + \alpha_L \log L_{it} + \alpha_M \log M_{it} + \alpha_K \log K_{it} + \alpha'_{IC} IC_i + \alpha'_D D_{it} + u_{it}, \quad (1)$$

[18] Data on Y variable is said to be missing completely at random (*MCAR*) if *P(Y missing | Y, X)= P(Y missing)*, where X is a matrix of other variables on data. Data is missing at random (*MAR*) if *P(Y missing | Y, X)= P(Y missing | X)*. Missing data is nonignorable if *P(Y missing | Y, X)= P(Y missing | X, Y)*.

where *logY, logL, logM* and *logK* represents output, labor, materials and capital all in logs, *IC* is the time-invariant vector of investment climate and other control variables and *D* is a vector of industry/region/size/time dummies. Since the usual time, industry, region and size fixed effects are included in the vector D, and the usual fixed effects are assumed to be observable and included in IC vector, u is assumed to be a usual *i.i.d* error.[19]

Equation (1) is of special interest for the purpose of this paper as it implies using a large proportion of the variables included in the ICSs. Furthermore, it is especially useful to illustrate the trade-off between plausible biases inherent in measurement errors that could arise after replacing missing data and the omitted variables bias associated with the complete case. Concretely, in the four cases considered, the final vector of significant IC variables is intended to include 27 variables in India, 18 in Turkey, 31 in South Africa and 25 in Tanzania.[20] The definition of the variables used, classified into five broad groups (infrastructures, red tape, finance, quality, and other), is in the appendix on definition of variables.

For identification in (1) if we observe all data and under regularity conditions, it is clear that, following Wooldridge (2007), we need $E(u_{it} \mid \log L_{it}, \log M_{it}, \log K_{it}, IC_i, D_{it}) = 0$. Now let the pattern of missing values for each observation *i* at moment *t* be given by s_{it}, where s_{it}=0 if missing value and 1 otherwise. So what we observe is

$$s_{it} \log Y_{it} = s_{it}(\alpha_0 + \alpha_L \log L_{it} + \alpha_M \log M_{it} + \alpha_K \log K_{it} + \alpha'_{IC} IC_i + \alpha'_D D_{it}) + s_{it} u_{it}. \quad (2)$$

If the pattern of missing values is *M.A.R* or *M.C.A.R* then the necessary conditions for equation (4) to be identified are $E(s_{it} u_{it}) = 0$, $E[(s_{it} J)(s_{it} u_{it})] = E[(s_{it} J u_{it})] = 0$ with $J = \log L_{it}, \log M_{it}, \log K_{it}, IC_i, D_{it}$. In the additional case of *exogenous sample selection*, when the pattern of missing values is determined only by the explanatory variables of (1),—for instance the missing values have some patterns on time, size, industries, regions or even between exporters/non-Exporters firms, domestic/foreign, etc—we also need that

$$E(s_{it} u_{it} \mid s_{it} \log L_{it}, s_{it} \log M_{it}, s_{it} \log K_{it}, s_{it} IC_i, s_{it} D_{it}) = s_{it} E(u_i \mid s_{it} \log L_{it}, s_{it} \log M_{it}, s_{it} \log K_{it}, s_{it} IC_i, s_{it} D_{it}) = 0.$$

That is, for the identification condition in this case to hold, we need to control for any exogenous variable affecting the pattern of missing values, and this is the way we proceed in the estimation of the productivity equations. Note that once we have controlled for all these variables, we can estimate (2) in the *complete case* consistently, although at the cost of losing efficiency and in some cases the representativity of the original sampling frame.

[19] Concretely, equation (1) is based on the methodology proposed in Escribano and Guasch (2005 and 2008) with further developments in Escribano et al (2008a and b). The selection of variables is detailed in these papers, and it is based on a general to particular procedure. Although for the purpose of this paper, we are not interested in the properties of the model, but wish to test the sensitivity of the results to the imputation method used, it is interesting to clarify that the underlying philosophy of this methodology is to use the time-invariant vectors of IC variables to correct for observable fixed effects.

[20] Although the initial set of IC vectors comprises more than 150 variables, a reduction process from the general to the specific was applied in order to find the final sets of significant variables. The final set of variables is required to be robust to 12 different TFP measures. More details are in Escribano and Guasch (2005 and 2008).

When the pattern of missing values *s* is correlated with the dependent variable of (1) we are in the presence of a self-selection case.[21] In this case the missing values are not ignorable and we cannot get rid of incomplete observations. In this case, equation (2) must be estimated by other sample selection corrections, such as the Heckman selection model.

In what follows we discuss the first imputation mechanism proposed to deal with the problem of incomplete data; *the ICA method*.

3.1 Imputation of missing values: The ICA method

Our method of imputing missing data, which we call the *ICA method*, shares the expectation step of the Expectation-Maximization (EM) algorithm proposed in the seminal paper of Dempster, Laird and Rubin (1977), a method that, within the maximum likelihood approaches, has been widely applied in several scientific fields (see McLachlan and Krishnan (1997) for a review). In particular, the replacement strategy used departs from the expectation of the production function variables conditional on the industry, region and size the corresponding observation belongs to ('*expectation step*'). Or equivalently, we replace the missing value by the expectation of the distribution of the variable conditional on the information on sector, region and size according to next equation

$$E(J_{it} \mid D_{R,it}, D_{I,it}, D_{S,it}) = \rho_0 + \rho'_{R,J} D_{R,it} + \rho'_{I,J} D_{I,it} + \rho'_{S,J} D_{S,it} \quad J = Y, L, M, K \tag{3}$$

where Y, L, M and K represents output, labor, materials and capital and D_R, D_I and D_S are vectors of region, industry and size dummies respectively. Notice that we choose (3) such that it represents the special features of the IC datasets—in IC surveys industry, region and size are the variables used to stratify the sample.

After excluding from the replacement process those observations with all the production function variables missing,[22] estimated values to replace incomplete data are given by

$$\tilde{J}_{it} = \hat{\rho}_0 + \hat{\rho}_{R,J} D_{R,it} + \hat{\rho}_{T,J} D_{I,it} + \hat{\rho}_{T,J} D_{S,it} \quad J = Y, L, M, K \tag{4}$$

Unlike the EM algorithm,[23] the ICA method has the advantage of separating the imputation of missing data from the estimation of the parameters of the population model. More precisely, separating the imputation mechanism of a population model is the main characteristic of the *multiple imputation* approaches, which allows using them with virtually any kind of data and any kind of model. The *ICA method* is, in fact, a general multiple imputation mechanism in which we assume that each imputed variable can be represented as a linear function of the

[21] Notice that as equation (1) is equivalent to: $\log Y_{it} - \alpha_L \log L_{it} - \alpha_M \log M_{it} - \alpha_K \log K_{it} = \alpha_0 + \alpha'_{IC} IC_i + \alpha'_D D_{it} + u_{it}$, where on the right hand side we have the productivity index. We are clearly concerned with the possible correlation of the MDM with productivity or TFP as it may induce biases in the estimators of the vector β.

[22] The ICA method is conservative in the sense that we do not replace missing cells for those observations with all but one PF variables unobserved. We force the industry-region-size cells to have at least 18 values to estimate consistently the sample average. Moreover, in order to avoid biases caused by outlier observations, we use the within-group median instead of the within-group mean.

[23] The EM algorithm imputes missing data conditional on a given population model, and therefore chooses the candidates' values to replace the missing cells that maximize the likelihood function conditional on a vector of parameters of that model.

variables used to stratify the sample (dummies of industry, region and size), and therefore the fitted values can be used to replace missing data.

Hence, the first assumption we need is that the imputed variable can be represented as a multiple linear function of other variables. The second condition that needs to be met for multiple imputation to work well is that all the variables, including those replaced and those used to replace, have normal distributions (see Allison, 2001).[24]

According to equation (3) and (4), equation (2) represents the '*maximization step*', which is now given by

$$s_{it}^{*} \log \tilde{Y}_{it} = s_{it}^{*}(\alpha_0 + \alpha_L \log \tilde{L}_{it} + \alpha_M \log \tilde{M}_{it} + \alpha_K \log \tilde{K}_{it} + \alpha'_{IC} IC_i + \gamma' D_{it}) + s_{it}^{*} \tilde{u}_{it} \quad (5)$$

where *y, l, m* and *k* with a tilde on top represent the imputed variables and *s** is the new pattern of missing values after the replacement process.[25] With identification conditions in the MAR case given by $E(s_{it}^{*}\tilde{u}_{it}) = 0$, $E[(s_{it}^{*}J)(s_{it}^{*}\tilde{u}_{it})] = E[(s_{it}^{*}J\tilde{u}_{it})] = 0$ with $J = \tilde{l}_{it}, \tilde{m}_{it}, \tilde{k}_{it}, IC_i, D_{it}$, while in the case of *exogenous sample selection* we need that

$$E(s_{it}^{*}\tilde{u}_{it} \mid s_{it}^{*} \log \tilde{L}_{it}, s_{it}^{*} \log \tilde{M}_{it}, s_{it}^{*} \log \tilde{K}_{it}, s_{it}^{*} IC_i, s_{it}^{*} D_{it}) = s_{it}^{*} E(\tilde{u}_{it} \mid s_{it}^{*} \log \tilde{L}_{it}, s_{it}^{*} \log \tilde{M}_{it}, s_{it}^{*} \log \tilde{K}_{it}, s_{it}^{*} IC_i, s_{it}^{*} D_{it}) = 0.$$

That is, we need to control for any explanatory variable correlated with *s** to get consistency either in the inputs or IC variables.

When the two assumptions mentioned above (normality and linearity of imputed variables on dummies of industry, region and size) do not hold, the replacement strategy is no longer consistent. Very little can be said about the asymptotic distributions of the estimators obtained under these circumstances because they have not yet been derived. In a general fashion, in these cases we can understand our replaced variables as the *classic* problem of *variables measured with error*. In order to illustrate this, let our model be given by $y_i = x_i\beta + u_i$, where y_i represents sales and x_i is a vector of inputs. Suppose that in the population we have that $E(u_i \mid x_i) = 0$, and that x_i is missing when $i \in S$. When we predict x_i $i \in S$ such that $\hat{x}_i = x_i + v_i$ where $\hat{x}_i$ is our predicted value, then the model becomes $y_i = \tilde{x}_i\beta + \tilde{v}_i\beta + u_i$. Where when $i \notin S$ $\tilde{x}_i = x_i$ and $v_i=0$, while if $i \in S$ $\tilde{x}_i = \hat{x}_i$ and $\tilde{v}_i = x_i - \hat{x}_i$. Therefore, consistency of estimates of β depends on whether $E(\tilde{v}_i \mid \tilde{x}_i) = 0$. Consistency follows if the linear regression of the inputs on industry, region and size variables gives us a noisy measure of the true level of the variables. Otherwise we will have a v_i and the parameters obtained from regression analysis would be consequently downward biased, and the magnitude of the bias will depend on the standard deviation of the error term relative to the standard deviation of the variable and the proportion of replaced values.[26]

[24] Although these are strong assumptions, the imputation method seems to works well even when the variables have distributions that are manifestly not normal, see Schafer (1997).

[25] Variables included in the IC and C vectors are imputed by using the same procedure. However, by means of illustration and simplification here we only discuss the identification condition as if only PF variables were imputed.

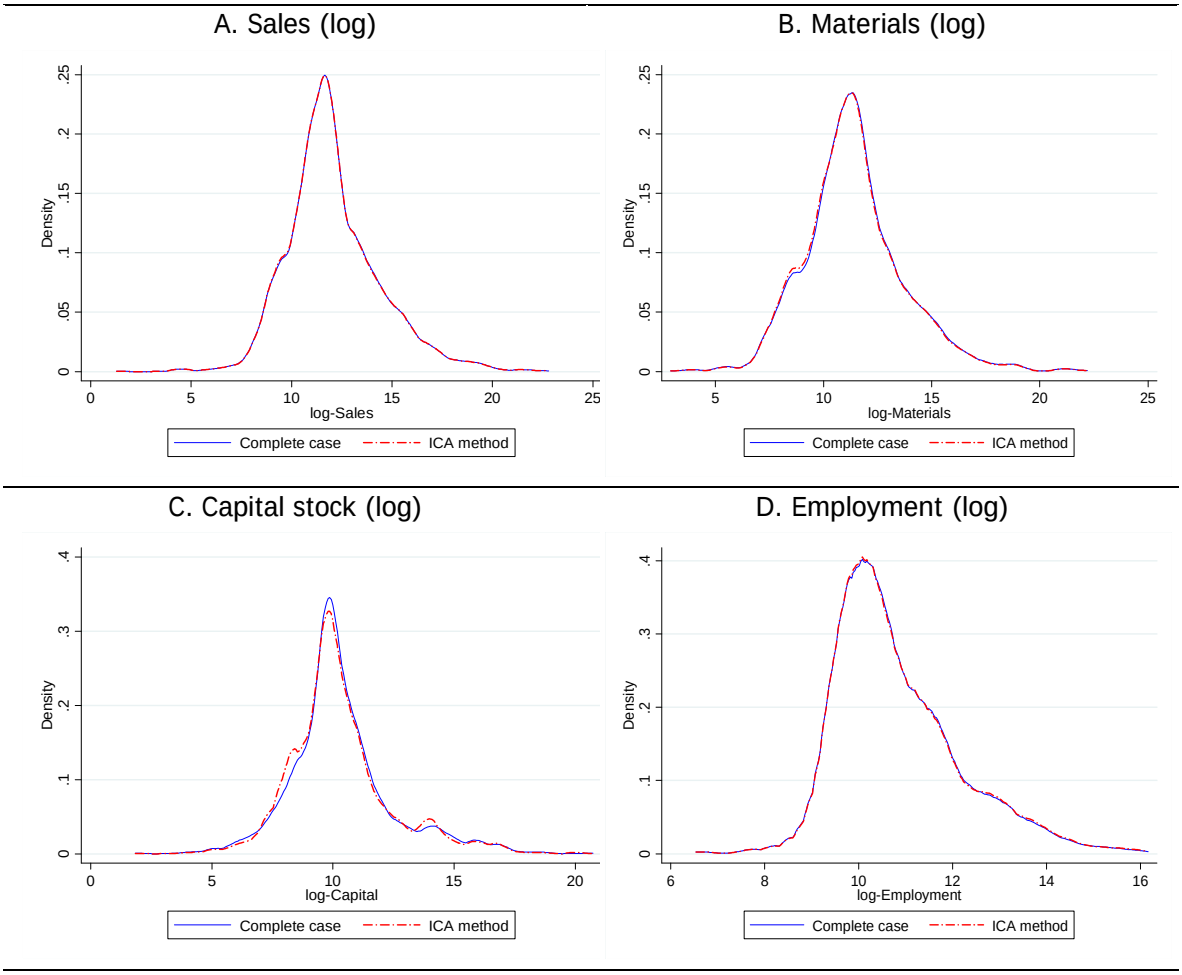
[26] We thank Ariel Pakes for useful suggestions at this point.

3.2 Performance of the ICA method

The performance of the ICA method is illustrated by plotting the Kernel densities of the PF variables in the complete case and after imputing missing data. Those are in figures 2.1 to 2.4 in the appendix at the end of the paper. Overall, from these figures the distributions of the ICA method and the complete case tend to be similar when the proportion of missing values is not too high. Divergences appear as the proportion of unobserved sample becomes larger.

Figure 2.1: INDIA, evaluation of performance of the ICA method

I. Kernel[1] estimates of output and input densities in the complete case and in the sample after imputing missing values by the ICA method



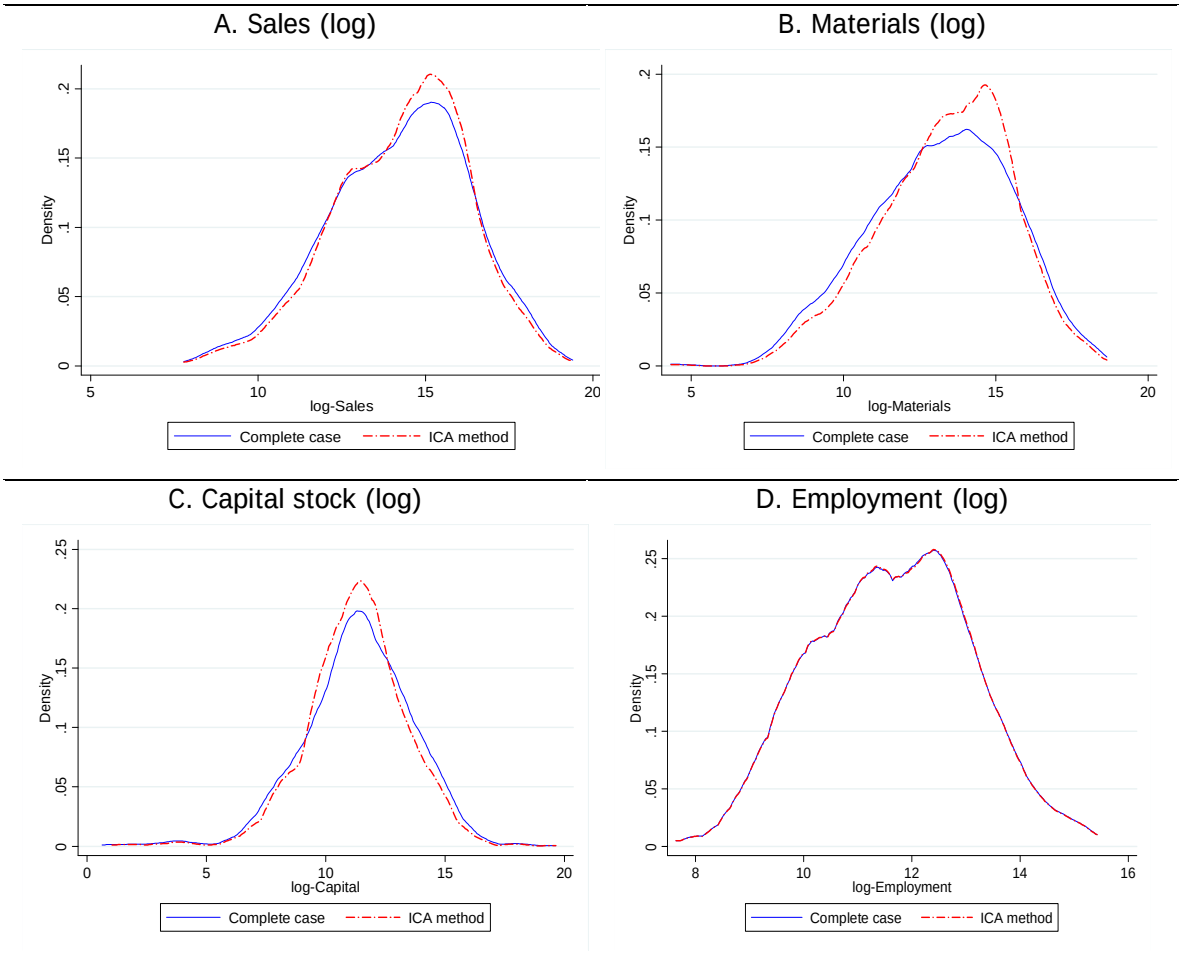
II. Table of descriptive statistics and tests of equality of distributions of output and inputs in the complete case and in the sample with imputation by the ICA method

		# Obs. (# imputed)	Mean	Std. Dev.	Min.	Max.	One-sample K-S Test (p-value)
Sales (log)	Complete case	5841	12.08	2.30	1.30	22.79	0.000
	ICA meth.	5935 (94)	12.07	2.29	1.30	22.79	0.000
Materials (log)	Complete case	5597	11.44	2.30	2.94	22.20	0.000
	ICA meth.	5933 (336)	11.40	2.28	2.94	22.20	0.000
Capital (log)	Complete case	4555	10.31	2.11	1.85	20.73	0.000
	ICA meth.	5918 (1363)	10.28	2.10	1.85	20.73	0.000
Empl (log)	Complete case	6164	10.82	1.33	6.54	16.16	0.000
	ICA meth.	6321 (157)	10.82	1.34	6.54	16.16	0.000

Notes:
1 Epanechnikov kernel. Each point estimated within a range of 300 values.
The null hypothesis of the one-sample Kolmogorov-Smirnov Test is that the cumulative distribution differs from the hypothesized theoretical normal distribution.
Source: Authors' estimations with ICSs data.

Figure 2.2: TURKEY, evaluation of performance of the ICA method

I. Kernel[1] estimates of output and input densities in the complete case and in the sample after imputing missing values by the ICA method



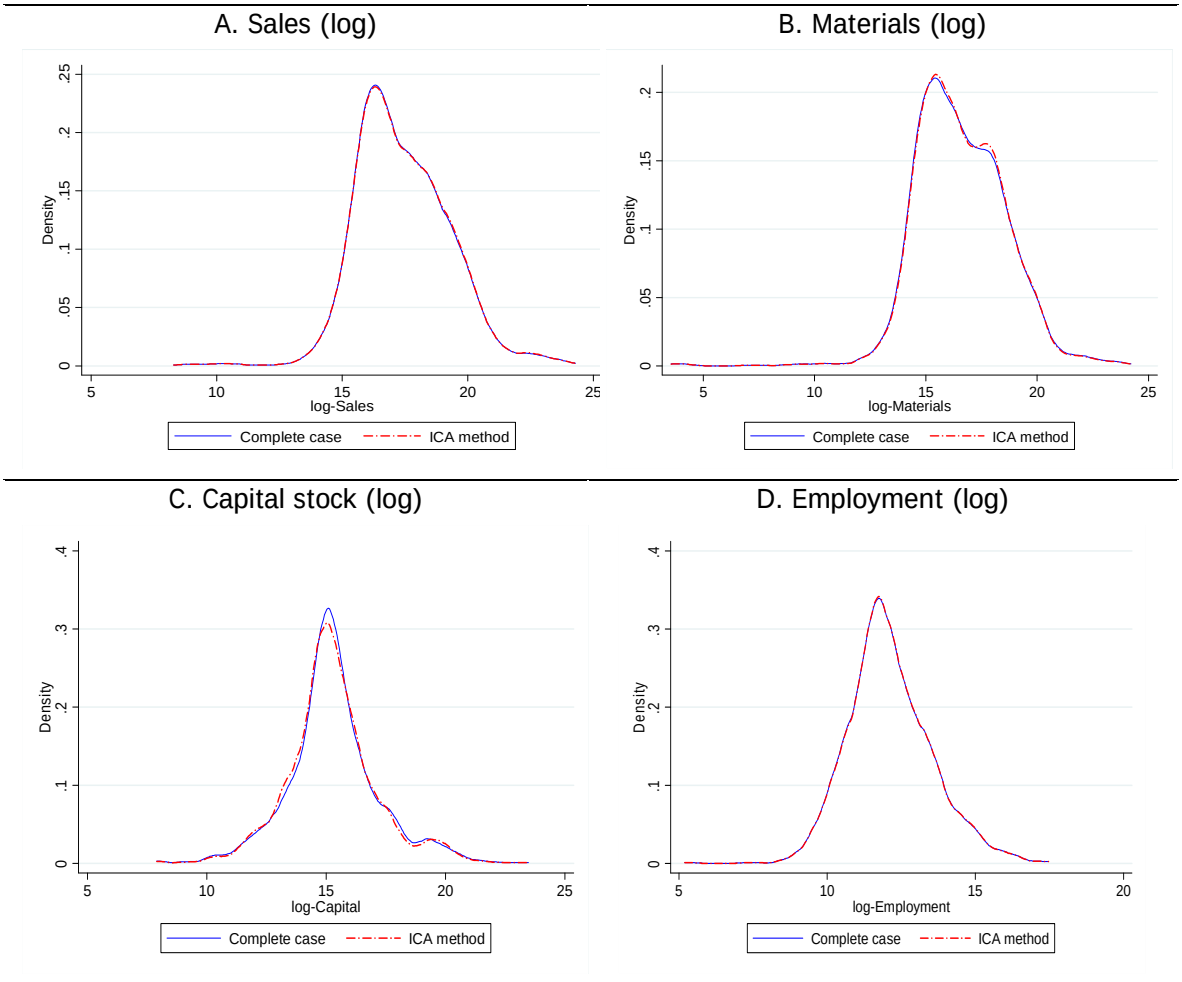
II. Table of descriptive statistics and tests of equality of distributions of output and inputs in the complete case and in the sample with imputation by the ICA method

		# Obs. (# imputed)	Mean	Std. Dev.	Min.	Max.	One-sample K-S Test (p-value)
Sales (log)	Complete case	1497	14.24	2.10	7.78	19.40	0.004
	ICA meth.	1821 (324)	14.30	1.99	7.78	19.40	0.000
Materials (log)	Complete case	1293	13.19	2.31	4.33	18.65	0.020
	ICA meth.	1822 (529)	13.37	2.13	4.34	18.65	0.000
Capital (log)	Complete case	1289	11.39	2.26	0.63	19.65	0.015
	ICA meth.	1816 (527)	11.32	2.05	1.05	19.65	0.004
Empl (log)	Complete case	2529	11.63	1.45	7.64	15.42	0.001
	ICA meth.	2548 (19)	11.63	1.45	7.64	15.42	0.001

Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
The null hypothesis of the one-sample Kolmogorov-Smirnov Test is that the cumulative distribution differs from the hypothesized theoretical normal distribution.
Source: Authors' estimations with ICSs data.

Figure 2.3: SOUTH AFRICA, evaluation of performance of the ICA method

I. Kernel[1] estimates of output and input densities in the complete case and in the sample after imputing missing values by the ICA method



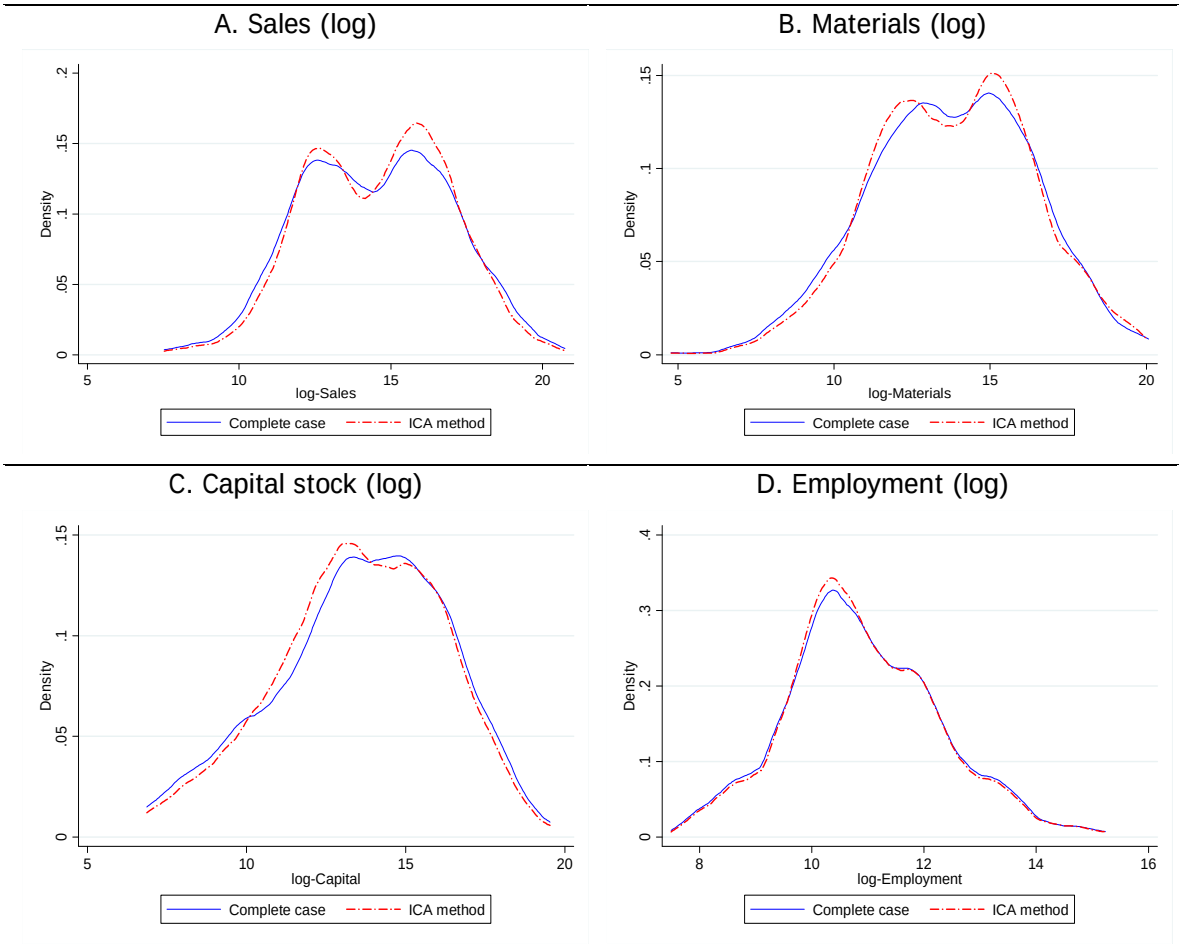
II. Table of descriptive statistics and tests of equality of distributions of output and inputs in the complete case and in the sample with imputation by the ICA method

		# Obs. (# imputed)	Mean	Std. Dev.	Min.	Max.	One-sample K-S Test (p-value)
Sales (log)	Complete case	1497	14.24	2.10	7.78	19.40	0.000
	ICA meth.	1821 (324)	14.30	1.99	7.78	19.40	0.000
Materials (log)	Complete case	1293	13.19	2.31	4.33	18.65	0.000
	ICA meth.	1822 (529)	13.37	2.13	4.34	18.65	0.000
Capital (log)	Complete case	1289	11.39	2.26	0.63	19.65	0.000
	ICA meth.	1816 (527)	11.32	2.05	1.05	19.65	0.000
Empl (log)	Complete case	2529	11.63	1.45	7.64	15.42	0.000
	ICA meth.	2548 (19)	11.63	1.45	7.64	15.42	0.000

Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
The null hypothesis of the one-sample Kolmogorov-Smirnov Test is that the cumulative distribution differs from the hypothesized theoretical normal distribution.
Source: Authors' estimations with ICSs data.

Figure 2.4: TANZANIA, evaluation of performance of the ICA method

I. Kernel[1] estimates of output and input densities in the complete case and in the sample after imputing missing values by the ICA method



II. Table of descriptive statistics and tests of equality of distributions of output and inputs in the complete case and in the sample with imputation by the ICA method

		# Obs. (# imputed)	Mean	Std. Dev.	Min.	Max.	One-sample K-S Test (p-value)
Sales (log)	Complete case	1497	14.24	2.10	7.78	19.40	0.012
	ICA meth.	1821 (324)	14.30	1.99	7.78	19.40	0.001
Materials (log)	Complete case	1293	13.19	2.31	4.33	18.65	0.169
	ICA meth.	1822 (529)	13.37	2.13	4.34	18.65	0.093
Capital (log)	Complete case	1289	11.39	2.26	0.63	19.65	0.053
	ICA meth.	1816 (527)	11.32	2.05	1.05	19.65	0.027
Empl (log)	Complete case	2529	11.63	1.45	7.64	15.42	0.006
	ICA meth.	2548 (19)	11.63	1.45	7.64	15.42	0.002

Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
The null hypothesis of the one-sample Kolmogorov-Smirnov Test is that the cumulative distribution differs from the hypothesized theoretical normal distribution.
Source: Authors' estimations with ICSs data.

From a more detailed analysis of Figure 2.1, which illustrates the case of India, it is clear that there are not significant differences in the distributions of any of the PF variables in the complete case and in the sample with replacement by the ICA method, which is supported by the Kolmogorov-Smirnov tests. Furthermore, both the sample mean and the standard deviation do not change significantly before and after the imputation process (especially important is the fact that the standard

deviation does not decline after the imputation). These observations hold for all the PF variables, even for the case of the capital stock, for which the proportion of imputed values is much higher than in the remaining variables. The case of South Africa case represented in Figure 2.3 reaches the same conclusions as the India sample.

On the other hand, the performance of the ICA method in the cases of Turkey and Tanzania shows significantly different behavior from the previous cases. Thus, in Turkey where the response rate of PF variables is below 40%, the kernel estimates suggest slight differences in the shape of the distributions, and, although the sample means are rather similar, the standard deviation estimated after imputing missing values decreases as the proportion of missing values increases. The same holds for the case of Tanzania, although in this case the problem becomes more acute as the sample distributions are far from normal, rejecting the null hypothesis of the Kolmogorov-Smirnov tests.

The extent to which the ICA method gives us a good approximation of the population distribution of the variables and therefore leads to a consistent estimation of equation (1) depends on the determinants of the MDM. Studying and analyzing the characteristics of the MDM is precisely the aim of sections 4 and 5, where we investigate the links between the patterns of missing values and productivity, sales and other key characteristics at the firm level such as accountability, informality, corruption, crime, innovative activity, etc. This analysis will be significantly important in the remaining sections, when we compare the ICA method with extensions and other different imputation mechanisms, which rely on different assumptions about the nature of the missingness mechanism.

4. The nature of the missing data mechanism

The following section aims to present a careful descriptive analysis of the characteristics of those firms having missing values, in order to judge whether the missing data mechanism may be treated as missing at random or not.

4.1 Why do some establishments refuse to provide or avoid providing certain information?

At this point, one question of great concern is the nature of the generating data process: missing completely at random, missing at random or non-ignorable missing data. Different assumptions can be made about the nature of the mechanism generating missing values. In general, missing values may be considered a consequence of some of the following causes: a) firms refuse to answer some questions (they do not have the information at hand, they simply do not know the information, they do not want to report it, they forget to answer some questions, etc); b) the interviewer neglects to ask some questions; and c) the question does not apply to some firms.

Since missing data arising from an oversight of the interviewer or because the question simply does not apply represents a small share of the total number of missing values and may be assumed as random, we are clearly concerned with the cases in which firms avoid, refuse or simply do not answer some questions. Here one can make some assumptions as to why firms do not report certain figures to the interviewer. Maybe firms do not report data on production function variables because of lack of accountability. It could also be a matter of informality. Those firms that do not report all sales to IRS authorities may have an incentive to avoid reporting these figures to the data collector as well, even though data is confidential. In this vein, one may also consider that missing

values could be correlated with the level of corruption within the environment in which firms operate.

Productivity or level of sales could also explain missing values: the higher the level of sales (or productivity) the lower the number of missing values. The explanation could simply be that weaker/less profitable firms do not keep proper accountability, or maybe the managers of weaker firms are less likely to know the PF figures (it is important to point out here that PF variables come from *recall* data). At this point, the question is whether the pattern of missing values is directly correlated with sales or TFP or if itis correlated indirectly through other variables such as share of exports, imports, access to infrastructures, capacity, innovation, R&D, quality, use of IC technologies, informality, corruption, accountability, etc, which are known to be strongly associated with sales and TFP.[27]

If the pattern of missing values is directly correlated with the dependent variable of our model—sales or TFP in our case—then the MAR or MCAR assumptions no longer hold. In this case, the missing value mechanism is said to be non-ignorable and the missing data mechanism needs to be modeled together with the structural model we are trying to estimate. On the other hand, when the missing data mechanism is related with sales or TFP indirectly through other—independent or exogenous variables in the dataset, the missing data mechanism is considered to be missing at random, which under regularity conditions is equivalent to saying that missing data is ignorable.[28] In this case we can get rid of missing data and operate only with the complete case once we have controlled for the variables correlated with the missingness mechanism. However, some caveats need to be made regarding *casewise deletion* as we will see in later sections.

The descriptive analysis we propose in this section allows us to obtain deeper and more thorough knowledge of the MDM. This is especially useful when the MDM is non-ignorable (not MAR and therefore not MCAR). As Meng (2000) signals, ignorability is untestable from the observed data, so caution is required when drawing conclusions from models with imputed data. Furthermore, sensitivity analysis and subjective knowledge of the nature of the MDM play a critical role here, as Molerberghs et al. (1999) illustrate. In fact, modeling the MDM is a very active line of research with a number of unresolved problems (see e.g. Heitjan, 1994 and 1999; Ibrahim, et al., 1999). From now on, the aim is, therefore, to describe the characteristics of those firms reporting missing values. The types of questions we are aiming to address are: has the missingness mechanism some relevant information for the parameters we are attempting to estimate? Or, in other words, are the parameters of the MDM related to the parameters of our model? And, as a consequence, is the MDM ignorable?

4.2 Is a missing value more likely to be found within small firms?

[27] Notice that we are concerned with the correlation of the MDM with either sales or TFP. We use the extended production function of equation (1) where a wide set of IC and C variables is plugged into a general PF in order to control for observable fixed effects. The correlation of MDM with sales may introduce bias in the input-output elasticities estimates, whereas the correlation with TFP could imply biased IC parameters estimates.

[28] A separate question is whether MAR is equivalent to ignorable missing data. Even when the missing data mechanism is assumed to be MAR, an additional assumption is needed to ensure that empty cells can be ignored: the parameters of the missing data process need to be unrelated with the parameters of the model we are willing to estimate. However, MAR and ignorability are almost always considered as equivalent assumptions in the literature, since the assumption that the parameters defining the missingness model are unrelated to the structural model is easily satisfied (see Allison, 2001 and Heitjan and Basu, 1996 for illustrations).

Firstly, we are concerned with the possibility of systematic bias in the response rates to questions on sales and inputs. Table 4 shows the number of missing values in sales and inputs according to size, which are known to correlate strongly with productivity (and also with sales).[29] The pattern in response rates is that small firms (those with fewer than twenty employees) tend to respond less often in India and South Africa. The pattern is somewhat different in Turkey and Tanzania where missing values in the inputs are uniformly distributed across categories of firms' sizes, with the exception of capital stock which has a higher proportion of missing values within small firms. At this point, these results could suggest the presence of some degree of systematic bias of the response rates in India and South Africa. Nonetheless, further investigation is needed to give additional insight into this question. The fact that small firms report less information also suggests that response rates to detailed sales and costs questions could have more to do with accounting and capacity—less affordable for small firms.

[29] Categories of size are: small, fewer than 20 employees; medium, between 20 and 100 employees; large, more than 100 employees.

Table 4: Number of missing values in production function variables by size

		Small	Medium	Large
	a) INDIA			
Totals by size		3,621	2,271	957
Sales	Number of missing [(a)]	646	257	95
	Perc over totals by size [(b)]	17.8	11.3	9.9
Labor	Number of missing	0	0	0
	Perc over totals by size	0	0	0
Materials	Number of missing	688	278	101
	Perc over totals by size	19	12.2	10.6
Capital	Number of missing	1258	640	245
	Perc over totals by size	34.7	28.2	25.6
	b) TURKEY			
Totals by size		788	810	1048
Sales	Number of missing	335	365	449
	Perc over totals by size	42.5	45.1	42.8
Labor	Number of missing	34	37	46
	Perc over totals by size	4.3	4.6	4.4
Materials	Number of missing	346	396	521
	Perc over totals by size	43.9	48.9	49.7
Capital	Number of missing	462	388	507
	Perc over totals by size	58.6	47.9	48.4
	c) SOUTH AFRICA			
Totals by size		197	783	804
Sales	Number of missing	40	95	76
	Perc over totals by size	20.3	12.1	9.5
Labor	Number of missing	23	54	43
	Perc over totals by size	11.7	6.9	5.3
Materials	Number of missing	53	111	97
	Perc over totals by size	26.9	14.2	12.1
Capital	Number of missing	69	204	154
	Perc over totals by size	35	26.1	19.2
	d) TANZANIA			
Totals by size		361	302	127
Sales	Number of missing	129	121	40
	Perc over totals by size	35.7	40.1	31.5
Labor	Number of missing	28	21	11
	Perc over totals by size	7.8	7	8.7
Materials	Number of missing	114	87	38
	Perc over totals by size	31.6	28.8	29.9
Capital	Number of missing	53	111	97
	Perc over totals by size	14.7	36.8	76.4

Small: less than 20 employees; medium: between 20 and 100 employees; large: more than 100 employees.
[(a)] Number of missing includes both missing values and outliers in the corresponding variables.
[(b)] Percentage over the total number of observations in each category of firms' size.
Source: Authors calculations with IC data.

4.3 Are missing values distributed uniformly across different categories of firms?

Tables 5.1 to 5.4 offer further empirical underpinning on whether the MDM is related to a firm's' weakness, or rather are other firms' attributes what determine the probability of observing a missing value. Table 5.1 focuses on the case of India. It compares the share of firms reporting at least one missing value on PF variables in the whole sample, with the share of firms reporting missing values by categories of key IC variables. In the case of India, 32.8% of firms report at least one missing value in PF variables. This percentage varies when we take into account categories of IC variables. Thus, those firms that do not use e-mail or experience power outages tend to respond less often to PF questions, respectively 39.0% and 37.8% of firms with missing information within these two categories. It is indicative of the nature of the MDM that those firms hiding some share of sales and/or workforce from IRS tax authorities have more missing values in PF variables on average (see the rows corresponding to Informality (I) and Informality (II)). With regard to corruption, those firms that operate in a more corrupt environment report fewer missing values. Similar conclusions can be obtained from crime; those firms having suffered criminal attempts also tend to avoid reporting PF figures.

Symptomatic of the nature of the MDM in India is the fact that firms with access to a credit line and with the annual statements reviewed by a external auditor, report a lower proportion of missing values (PF information is lost for 40.4% of firms without access to credit and 50.2% of firms with the annual statements not audited externally, report at least one missing value). This indicates that a plausible explanation for the missing values is the lack of proper accountability or even informality.

Continuing with Table 5.1, other indicative variables of the pattern of missing values are the exporting activity (only 18.2% of those firms exporting directly report any missing value) and the education of the manager (28.5% of firms with a manager with a university education report missing values, while 35.1% of the remaining firms report missing values). These two variables indicate that the level of competitiveness of the firm is another important factor explaining the pattern of missing values. However, other variables that are known to correlate strongly with competitiveness and productivity, such as FDI or the introduction of new technologies and products,, do not provide any further information on the MDM.

Table 5.1: INDIA, Proportion of observations with missing values in production function (PF) variables by key IC determinants

Key IC variables		Proportion of Establishments with: complete information on PF variables	at least one missing value in PF variables
Whole sample		67.2	32.8
1. Generator	Establishments not using own generator	68.6	31.4
	Establishments using own generator	66.3	33.7
2. Power outages	Establishments that do not experience power outages	61	39
	Establishments experiencing power outages	69.4	30.6
3. Water outages	Establishments that do not experience water outages	66.9	33.1
	Establishments experiencing water outages	71.5	28.5
4. E-mail	Establishments that do not use e-mail	62.2	37.8
	Establishments using e-mail	70.6	29.4
5. Web page	Establishments that do not use web page	66.8	33.2
	Establishments using web page	68.3	31.7
6. Informality (I)	Establishments reporting all sales to IRS authorities	76.4	23.6
	Establishments that hide some share of sales from the IRS	63.5	36.5
7. Informality (II)	Establishments reporting all workforce to IRS authorities	78.1	21.9
	Establishments that hide some share of workforce from the IRS	62	38
8. Corruption (I)	Establishments that do not pay bribes to deal with bureaucracy	63.4	36.6
	Establishments paying bribes to deal with bureaucracy	71.6	28.4
9. Corruption (II)	Establishments that do not pay bribes to obtain contracts with the gov.	64.6	35.4
	Establishments paying bribes to obtain contracts with the government	74.3	25.7
10. Crime	Establishments that do not suffer losses due to crime	67.7	32.3
	Establishments suffering losses due to crime	58.4	41.6
11. Security	Establishments without security expenses	67.1	32.9
	Establishments with security expenses	68.2	31.8
12. Loan	Establishments without access to a loan	67.5	32.5
	Establishments with access to a loan	67.2	32.8
13. Credit line	Establishments without access to a credit line	59.6	40.4
	Establishments with access to a credit line	73.8	26.2
14. Auditory	Establishments with annual statements reviewed by external auditory	49.8	50.2
	Establishments without annual statements reviewed by external auditory	70.4	29.6
15. Innovation (I)	Establishments without ISO certification	67	33
	Establishments with ISO certification	67.8	32.2
16. Innovation (II)	Establishments that do not introduce new products	66.4	33.6
	Establishments introducing new products	68.7	31.3
17. Innovation (III)	Establishments that do not introduce new technologies		
	Establishments introducing new technologies		
18. Training	Establishments that do not provide training	71.4	28.6
	Establishments providing training	65.1	34.9
19. Manager skills	Managers with less than a university education	64.9	35.1
	Managers with more than a university education	71.5	28.5
20. Exporting activity	Establishments that do not export	68.9	31.1
	Establishments exporting	81.8	18.2
21. FDI inflows	Establishments that do not receive FDI inflows	67.2	32.8
	Establishments receiving FDI inflows	60.7	39.3
22. Incorporated company	Establishments not in an incorporated company	66.8	33.2
	Establishments in an incorporated company	67.9	32.1
23. Holding	Establishments not in a holding		
	Establishments in a holding		
24. Capacity utilization	Establishments that do not use all their capacity	67.2	32.8
	Establishments using all their capacity	68.6	31.4

Within production function variables we include labor (labor cost), capital, sales and materials.
Source: Authors calculations with IC data.

Table 5.2: TURKEY, Proportion of observations with missing values in production function (PF) variables by key IC determinants

Key IC variables		Proportion of Establishments with: complete information on PF variables	at least one missing value in PF variables
Whole sample		52.4	47.6
1. Generator	Establishments not using own generator		
	Establishments using own generator		
2. Power outages	Establishments that do not experience power outages	41.1	58.9
	Establishments experiencing power outages	55.7	44.3
3. Water outages	Establishments that do not experience water outages	53.7	46.3
	Establishments experiencing water outages	44.7	55.3
4. E-mail	Establishments that do not use e-mail	56.0	44.0
	Establishments using e-mail	51.5	48.5
5. Web page	Establishments that do not use web page	51.8	48.2
	Establishments using web page	52.6	47.4
6. Informality (I)	Establishments reporting all sales to IRS authorities	47.1	52.9
	Establishments that hide some share of sales from IRS	55.2	44.8
7. Informality (II)	Establishments reporting all workforce to IRS authorities	47.6	52.4
	Establishments that hide some share of workforce from IRS	57.0	43.0
8. Corruption (I)	Establishments that do not pay bribes to deal with bureaucracy	48.0	52.0
	Establishments paying bribes to deal with bureaucracy	76.2	23.8
9. Corruption (II)	Establishments that do not pay bribes to obtain contracts with the gov	47.7	52.3
	Establishments paying bribes to obtain contracts with the government	63.9	36.1
10. Crime	Establishments that do not suffer losses due to crime	52.2	47.8
	Establishments suffering losses due to crime	54.4	45.6
11. Security	Establishments without security expenses	32.0	68.0
	Establishments with security expenses	93.0	7.0
12. Loan	Establishments without access to a loan	47.6	52.4
	Establishments with access to a loan	56.4	43.6
13. Credit line	Establishments without access to a credit line	45.5	54.5
	Establishments with access to a credit line	60.4	39.6
14. Auditory	Establishments with annual statements reviewed by external auditory	56.2	43.8
	Establishments without annual statements reviewed by external auditory	47.1	52.9
15. Innovation (I)	Establishments without ISO certification	51.0	49.0
	Establishments with ISO certification	54.4	45.6
16. Innovation (II)	Establishments that do not introduce new products	50.6	49.4
	Establishments introducing new products	55.5	44.5
17. Innovation (III)	Establishments that do not introduce new technologies	44.0	56.0
	Establishments introducing new technologies	64.0	36.0
18. Training	Establishments that do not provide training	47.5	52.5
	Establishments providing training	56.6	43.4
19. Manager skills	Managers with less than a university education	52.0	48.0
	Managers with more than a university education	53.9	46.1
20. Exporting activity	Establishments that do not export	54.3	45.7
	Establishments exporting	50.2	49.8
21. FDI inflows	Establishments that do not receive FDI inflows	52.8	47.2
	Establishments receiving FDI inflows	43.1	56.9
22. Incorporated company	Establishments not in an incorporated company	51.9	48.1
	Establishments in an incorporated company	62.1	37.9
23. Holding	Establishments not in a holding	53.1	46.9
	Establishments in a holding	42.5	57.5
24. Capacity utilization	Establishments that do not use all their capacity	55.5	44.5
	Establishments using all their capacity	38.0	62.0

Within production function variables we include labor (labor cost), capital, sales and materials.
Source: Authors calculations with IC data.

Table 5.3: SOUTH AFRICA, Proportion of observations with missing values in production function (PF) variables by key IC determinants

Key IC variables		Proportion of Establishments with: complete information on PF variables	at least one missing value in PF variables
Whole sample		72	28
1. Generator	Establishments not using own generator	71.8	28.2
	Establishments using own generator	73.3	26.7
2. Power outages	Establishments that do not experience power outages	63.4	36.6
	Establishments experiencing power outages	76.6	23.4
3. Water outages	Establishments that do not experience water outages	64.9	35.1
	Establishments experiencing water outages	89.4	10.6
4. E-mail	Establishments that do not use e-mail	33.3	66.7
	Establishments using e-mail	72.5	27.5
5. Web page	Establishments that do not use web page	71.9	28.1
	Establishments using web page	72.0	28.0
6. Informality (I)	Establishments reporting all sales to IRS authorities	59.3	40.7
	Establishments that hide some share of sales from IRS	74.3	25.7
7. Informality (II)	Establishments reporting all workforce to IRS authorities		
	Establishments that hide some share of workforce from IRS		
8. Corruption (I)	Establishments that do not pay bribes to deal with bureaucracy	73.4	26.6
	Establishments paying bribes to deal with bureaucracy	33.3	66.7
9. Corruption (II)	Establishments that do not pay bribes to obtain contracts with the gov	73.7	26.3
	Establishments paying bribes to obtain contracts with the government	40.0	60.0
10. Crime	Establishments that do not suffer losses due to crime	70.9	29.1
	Establishments suffering losses due to crime	72.9	27.1
11. Security	Establishments without security expenses	62.1	37.9
	Establishments with security expenses	74.5	25.5
12. Loan	Establishments without access to a loan	73.9	26.1
	Establishments with access to a loan	68.8	31.2
13. Credit line	Establishments without access to a credit line	72.6	27.4
	Establishments with access to a credit line	71.6	28.4
14. Auditory	Establishments with annual statements reviewed by external auditory	38.9	61.1
	Establishments without annual statements reviewed by external auditory	73.0	27.0
15. Innovation (I)	Establishments without ISO certification	70.9	29.1
	Establishments with ISO certification	73.6	26.4
16. Innovation (II)	Establishments that do not introduce new products	62.4	37.6
	Establishments introducing new products	76.4	23.6
17. Innovation (III)	Establishments that do not introduce new technologies	67.6	32.4
	Establishments introducing new technologies	74.9	25.1
18. Training	Establishments that do not provide training	73.5	26.5
	Establishments providing training	71.1	28.9
19. Manager skills	Managers with less than a university education	63.7	36.3
	Managers with more than a university education	75.3	24.7
20. Exporting activity	Establishments that do not export	69.8	30.2
	Establishments exporting	75.4	24.6
21. FDI inflows	Establishments that do not receive FDI inflows	71.9	28.1
	Establishments receiving FDI inflows	72.4	27.6
22. Incorporated company	Establishments not in an incorporated company	72.9	27.1
	Establishments in a incorporated company	51.4	48.6
23. Holding	Establishments not in a holding	72.4	27.6
	Establishments in a holding	69.0	31.0
24. Capacity utilization	Establishments that do not use all their capacity	72.8	27.2
	Establishments using all their capacity	66.7	33.3

Within production function variables we include labor (labor cost), capital, sales and materials.
Source: Authors calculations with IC data.

Table 5.4: TANZANIA, Proportion of observations with missing values in production function (PF) variables by key IC determinants

Key IC variables		Proportion of Establishments with: complete information on PF variables	at least one missing value in PF variables
Whole sample		44.8	55.2
1. Generator	Establishments not using own generator	44.9	55.1
	Establishments using own generator	45.1	54.9
2. Power outages	Establishments that do not experience power outages	42.1	57.9
	Establishments experiencing power outages	45.7	54.3
3. Water outages	Establishments that do not experience water outages	42.5	57.5
	Establishments experiencing water outages	50.4	49.6
4. E-mail	Establishments that do not use e-mail	43.2	56.8
	Establishments using e-mail	46.4	53.6
5. Web page	Establishments that do not use web page	43.4	56.6
	Establishments using web page	50.0	50.0
6. Informality (I)	Establishments reporting all sales to IRS authorities	45.6	54.4
	Establishments that hide some share of sales from IRS	44.3	55.7
7. Informality (II)	Establishments reporting all workforce to IRS authorities		
	Establishments that hide some share of workforce to IRS		
8. Corruption (I)	Establishments that do not pay bribes to deal with bureaucracy	41.3	58.7
	Establishments paying bribes to deal with bureaucracy	50.0	50.0
9. Corruption (II)	Establishments that do not pay bribes to obtain contracts with the gov	42.8	57.2
	Establishments paying bribes to obtain contracts with the government	54.2	45.8
10. Crime	Establishments that do not suffer losses due to crime	58.9	41.1
	Establishments suffering losses due to crime	0.0	0.0
11. Security	Establishments without security expenses	45.0	55.0
	Establishments with security expenses	47.6	52.4
12. Loan	Establishments without access to a loan	51.8	48.2
	Establishments with access to a loan	61.0	39.0
13. Credit line	Establishments without access to a credit line	42.1	57.9
	Establishments with access to a credit line	50.2	49.8
14. Auditory	Establishments with annual statements reviewed by external auditory	32.7	67.3
	Establishments without annual statements reviewed by external auditory	48.9	51.1
15. Innovation (I)	Establishments without ISO certification	43.4	56.6
	Establishments with ISO certification	57.6	42.4
16. Innovation (II)	Establishments that do not introduce new products	44.9	55.1
	Establishments introducing new products	47.0	53.0
17. Innovation (III)	Establishments that do not introduce new technologies	48.3	51.7
	Establishments introducing new technologies	39.9	60.1
18. Training	Establishments that do not provide training	44.5	55.5
	Establishments providing training	47.9	52.1
19. Manager skills	Managers with less than a university education		
	Managers with more than a university education		
20. Exporting activity	Establishments that do not export	44.6	55.4
	Establishments exporting	51.6	48.4
21. FDI inflows	Establishments that do not receive FDI inflows	43.9	56.1
	Establishments receiving FDI inflows	47.5	52.5
22. Incorporated company	Establishments not in an incorporated company	45.1	54.9
	Establishments in an incorporated company	38.1	61.9
23. Holding	Establishments not in a holding	46.4	53.6
	Establishments in a holding	33.3	66.7
24. Capacity utilization	Establishments that do not use all their capacity	45.5	54.5
	Establishments using all their capacity	36.1	63.9

Within production function variables we include labor (labor cost), capital, sales and materials.
Source: Authors calculations with IC data.

The case of Turkey is represented in Tables 5.2. The patterns are similar to those observed in India. TPower outages experienced, e-mail usage, informalities and corruption are good indicators of the pattern of missing values. Again the proportion of missing values within firms having access to credit and to an external auditory is larger relative to those that do not, which all corroborates the explanation of accountability as a determinant of the MDM. Other variables with important

implications for the MDM are exports, the FDI, the introduction of new technologies, the legal status of the firm (am incorporated company or not) and the percentage of capacity utilization.

Similar conclusions can be obtained for South Africa in Table 5.3. Missingness in this country appears to be associated with water outages, use of e-mail, informality and corruption, accountability, and the legal status, and, to a lesser extent, with power outages, security expenses and the introduction of new products and technologies.

These patterns are even more pronounced in Tanzania. Table 5.4 illustrates that, for instance, in those firms with access to a loan, 39% report missing values, while in those firms without loans the percentage rises to 48.2%. The same holds for informality, corruption, quality, technology, exporting activity, legal status, holdings or capacity utilization.

4.4 More on the relationship between the MDM and the investment climate variables

Continuing with the analysis presented so far and in order to go into more depth regarding the relationship between the probability of observing a missing value in TFP and the IC variables, we propose the following model for the probability of observing data on TFP in terms of IC and D variables

$$\Pr(s_i^a = 1 \mid D_i, IC_i) = \varphi(\rho_0^a + \rho_2^{a\prime} D_i + \rho_3^{a\prime} IC_i + \upsilon_i^a),$$

where s_{it}^a is a dichotomous variable of value 1 if we observe all sales, labor, materials and capital and zero otherwise. Symmetrically, in the case of sales, we have the following equation

$$\Pr(s_i^b = 1 \mid D_i, IC_i) = \varphi(\rho_0^b + \rho_2^{b\prime} D_i + \rho_3^{b\prime} IC_i + \upsilon_i^b),$$

where in this case s_{it}^b takes value 1 if we observe data for sales.

Tables 6.1 to 6.4 present the estimated results by applying a LPM to model the probability of having a missing value conditional on the investment climate faced by firms. Concretely, we propose four models for each country. First we consider missing values in TFP conditioning in two different vectors of IC variables. The first specification includes the same set of IC variables as that included in equation (5); that is, the set of covariates statistically significant in the extended production function, before imputing missing values by the ICA method. The second specification chooses the set of significant correlates starting from the whole set of IC variables and applying a general-to-specific procedure of selection of variables. The case of sales is symmetrical in the sense that model [3] uses the same set of IC variables as in equation (5), while the specification shown in column [4] selects the set of variables as we did in the case of column [2].

Table 6.1: INDIA, Linear probability models for the probability of observing TFP and sales

Dependent variables:	Missing on TFP [a]		Missing on sales [b]	
	[1]	[2]	[3]	[4]
Explanatory variables:	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.
Infrastructures:				
Longest # of days to clear customs for export (a)	-0.0279 [0.0112]**		-0.0108 [0.0082]	
Dummy for own generator	-0.0066 [0.0165]		0.0072 [0.0134]	
Water supply from public sources (b)	0.0001 [0.0002]		0.0000 [0.0002]	
Shipment losses in the domestic market (b)	-0.0044 [0.0015]***		-0.0028 [0.0014]**	
Dummy for own transport	-0.0083 [0.0208]		0.0122 [0.0199]	
Dummy for web page	0.0153 [0.0177]		0.0191 [0.0207]	
Losses due to power outages (b)		-0.0023 [0.0010]**		-0.0025 [0.0007]***
Dummy for e-mail (b)		0.0282 [0.0166]*		0.031 [0.0183]*
Shipment losses, domestic (b)		-0.0043 [0.0014]		-0.0028 [0.0011]**
Losses due to transport outages (b)		-0.0033 [0.0018]***		-0.0035 [0.0015]**
Red tape, corruption and crime:				
Dummy for security	0.0146 [0.0188]		0.0033 [0.0157]	
Sales reported to taxes (b)	0.0006 [0.0006]		0.0005 [0.0005]	
Workforce reported f taxes (b)	-0.0004 [0.0004]		-0.0001 [0.0004]	
Dummy for payments to speed up bureaucracy	0.0347 [0.0137]**		0.0359 [0.0122]***	
Dummy for interventionist labor regulation	-0.0327 [0.0180]*	-0.0379 [0.0187]**	-0.0383 [0.0185]**	-0.0409 [0.0189]**
Absenteeism (b)	-0.0165 [0.0074]**		-0.0122 [0.0057]**	
Dummy for payments to deal with bur. issues (b)		0.0222 [0.0140]		0.0261 [0.0136]*
Finance:				
Dummy for external audit	0.0121 [0.0174]	0.0538 [0.0252]**	0.0086 [0.0140]	0.0423 [0.0161]***
Dummy for trade association	-0.0002 [0.0002]		0.0003 [0.0002]	
Working capital financed by domestic private banks (b)	0.0234 [0.0146]		0.0231 [0.0134]*	
Dummy for loan (b)	0.0337 [0.0209]		0.0319 [0.0159]**	
Largest shareholder (b)		-0.0003 [0.0002]		-0.0004 [0.0002]**
Dummy for loan with collateral (b)		-0.0802 [0.0318]**		-0.0573 [0.0252]**
Loans denominated in foreign currency (b)		-0.0011 [0.0003]***		-0.0008 [0.0003]***
Quality, innovation and labor skills:				
Dummy for R&D (a)	0.0016 [0.1084]	0.0153 [0.0147]	-0.04 [0.0666]	0.0296 [0.0130]**
Dummy for product innovation	-0.0073 [0.0157]		-0.0099 [0.0133]	
Dummy for foreign license (b)	0.0481 [0.0314]		0.0572 [0.0297]*	
Dummy for internal training (b)	0.0025 [0.0197]		0.0001 [0.0186]	
Unskilled workforce (a)	0.0021 [0.0012]*		0.0017 [0.0011]	
Workforce with computer	0.0006 [0.0004]		0.0001 [0.0003]	
Dummy for ISO quality certification (b)		0.0148 [0.0173]		0.0325 [0.0156]***
Dummy for outsourcing (b)		0.0457 [0.0174]		0.0213 [0.0135]
Dummy for external training (b)		-0.0334 [0.0235]		-0.0256 [0.0164]
Other control variables:				
Dummy for incorporated company	0.0185 [0.0146]		0.0308 [0.0139]**	
Age	0.0077 [0.0103]		0.0097 [0.0095]	
Share of exports (b)	0.0002 [0.0002]		0.0002 [0.0002]	
Trade union (b)	0.0007 [0.0004]*	0.0008 [0.0003]	0.0006 [0.0003]*	0.0008 [0.0003]***
Strikes (b)	-0.0165 [0.0133]		-0.0037 [0.0158]	
Constant	*Yes*	*Yes*	*Yes*	*Yes*
Industry/region/size dummies	*Yes*	*Yes*	*Yes*	*Yes*
Observations	2048	2277	2048	2277
R-squared	0.23	0.23	0.18	0.18

(a) Missing in TFP takes value 1 if we observe all sales, materials, labor and capital, and 0 otherwise.
(b) Missing in TFP takes value 1 if we observe sales, and 0 otherwise.
[1] Model of the probability of observing a missing value in TFP conditional the IC and C variables significant in equation (1).
[2] Model of the probability of observing a missing value in TFP and the matrices IC* and C*, selected from the whole set of IC and C variables.
[1] Model of the probability of observing a missing value in sales conditional on in the IC and C variables significant in equation (1).
[2] Model of the probability of observing a missing value in sales and the matrices IC* and C*, selected from the whole set of IC and C variables.
Significance given by robust standard errors allowing for clustering by industry and region *** 1%, **5%, * 10%.
Source: Authors' estimations with ICSs data.

Table 6.2: TURKEY, Linear probability models for the probability of observing TFP and sales

Dependent variables:	Missing on TFP		Missing on sales	
	[1]	[2]	[3]	[4]
Explanatory variables:	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.
Infrastructures:				
Days to clear customs for imports (a)	0.019 [0.0592]		0.0189 [0.0669]	
Losses due to power outages (b)				-0.0029 [0.0016]*
Losses due to water outages (b)				0.0035 [0.0010]***
Shipment losses (b)				-0.0038 [0.0017]**
Dummy for e-mail (b)	0.021 [0.0341]		0.0811 [0.0378]**	0.1088 [0.0377]***
Electricity from generator (b)		0.0009 [0.0004]**		
Red tape, corruption and crime:				
Crime losses (b)		0.0024 [0.0005]***		0.0035 [0.0004]***
Security expenses (b)	0.1273 [0.0350]***		0.1322 [0.0403]***	
Manager's time spent on bur. issues (b)		-0.003 [0.0009]***		-0.0025 [0.0012]**
Dummy for consultant to help deal with bur. issues		-0.0693 [0.0175]***		-0.0713 [0.0270]**
Number of inspections (B)	-0.0036 [0.0022]	-0.0221 [0.0129]*	0.0003 [0.0002]	
Payments to deal with bureaucratic issues (a)	0.00001 [0.0002]	0.0013 [0.0004]***	0.0092 [0.0037]**	0.0019 [0.0004]***
Sales declared for taxes (a)	0.0087 [0.0035]**	-0.0011 [0.0004]**	-0.003 [0.0022]	-0.0013 [0.0004]***
Payments to obtain a contract with the government (b)	-0.0309 [0.0132]**	-0.0156 [0.0022]***	-0.0276 [0.0170]	-0.0136 [0.0028]***
Production lost due to absenteeism (b)	-0.0149 [0.0024]***		-0.0136 [0.0027]***	
Dummy for informal competition (b)	-0.0332 [0.0177]*		-0.0368 [0.0176]**	
Delay in obtaining a water supply (a)	-0.0282 [0.0214]		-0.033 [0.0238]	
Dummy for lawsuit (b)		-0.0494 [0.0218]**		-0.0728 [0.0293]**
Finance:				
Dummy for credit line	-0.0763 [0.0243]***		-0.0908 [0.0247]***	-0.0778 [0.0232]***
Dummy for external auditory (a)	0.0443 [0.0194]**	-0.0548 [0.0234]**	0.0327 [0.0230]	
Loans in foreign currency (b)	-0.0005 [0.0003]*		-0.0006 [0.0005]	
Dummy for new land purchased		-0.0528 [0.0313]*		
Dummy for loan denominated in Turkish Lira (b)		-0.1216 [0.0238]***		-0.1645 [0.0238]***
Dummy for loan denominated in foreign currency (b)		-0.1001 [0.0317]***		-0.1472 [0.0379]***
Dummy for long-term loan (b)				0.1261 [0.0356]***
Quality, innovation and labor skills:				
Dummy for ISO quality certification (b)		0.0869 [0.0192]***		0.0696 [0.0206]***
Dummy for new technology (b)		-0.1027 [0.0223]***		-0.0987 [0.0260]***
Dummy for foreign licensed technology (b)				0.0607 [0.0244]**
Staff with university education (b)	0.0001 [0.0010]	0.0016 [0.0007]**	0.001 [0.0010]	
Staff-part time workers	0.0018 [0.0007]**		0.0014 [0.0009]	0.0012 [0.0008]
Other control variables:				
Dummy for incorporated company		-0.092 [0.0557]		-0.0851 [0.0394]**
Age				-0.0457 [0.0220]**
Market share				0.0008 [0.0007]
Production lost due to strikes (b)	-0.0408 [0.0180]**		-0.0056 [0.0246]	
Dummy for recently privatized firm	0.0222 [0.0949]		-0.0344 [0.0877]	
Dummy for competition against imported products	-0.0472 [0.0441]		-0.0261 [0.0393]	
Constant	Yes	Yes	Yes	Yes
Industry/region/size dummies	Yes	Yes	Yes	Yes
Observations	1323	1323	1323	1323
R-squared	0.2	0.31	0.24	0.3

See footnotes in Table 6.1.
Source: Authors' estimations with ICSs data.

Table 6.3: SOUTH AFRICA, Linear probability models for the probability of observing TFP and sales

Dependent variables:	Missing on TFP	Missing on sales		
	[1]	[2]	[3]	[4]
Explanatory variables:	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.
Infrastructures:				
Days to clear customs for imports (a)	-0.018 [0.0587]		-0.0782 [0.0509]	
Sales lost due to power outages (b)	-0.0061 [0.0044]	-0.0068 [0.0036]*	-0.0059 [0.0026]**	-0.0051 [0.0022]**
Water outages (b)	0.0166 [0.0231]	0.016 [0.0032]***	0.0021 [0.0196]	
Average duration of transport failures (a)	-0.0206 [0.0467]		0.0064 [0.0445]	
Wait for electric supply (a)	0.0193 [0.0313]		0.0202 [0.0342]	
Dummy for email (b)		0.1795 [0.0686]**		
Dummy for internet				0.0356 [0.0138]**
Sales lost due to delivery delays (b)	0.0103 [0.0040]**	0.0115 [0.0034]***	0.003 [0.0028]	0.0039 [0.0027]
Red tape, corruption and crime:				
Manager's time spent on bur. issues (b)	0.0022 [0.0010]**		0.001 [0.0007]	
Payments to deal with bureaucratic issues (b)	-0.0011 [0.0007]	-0.0015 [0.0005]***	-0.0011 [0.0008]	
Sales declared for taxes (a)	0.0006 [0.0028]		-0.0022 [0.0029]	-0.0027 [0.0016]*
Payments to obtain a contract with the gov. (b)	0.0119 [0.0078]		0.0199 [0.0093]**	0.0199 [0.0079]**
Security expenses (a)	0.0033 [0.0102]	0.0078 [0.0024]***	0.0084 [0.0082]	
Crime losses (a)				0.0241 [0.0201]
Illegal payments in protection (b)	-0.0324 [0.0595]		-0.0003 [0.0424]	
Crime losses (a)	0.023 [0.0404]		0.0472 [0.0368]	
Finance:				
Percentage of credit unused (b)	0.0002 [0.0003]		0.0004 [0.0003]	0.0004 [0.0002]*
Dummy for loan	-0.0025 [0.0329]		0.0017 [0.0213]	
Dummy for credit line (b)				-0.0193 [0.0143]
Value of the collateral (b)	0.00001 [0.0002]		-0.0001 [0.0001]	
Loans in foreign currency (b)	0.0002 [0.0008]		-0.0005 [0.0004]	-0.0006 [0.0003]*
Charge to clear a check (a)	-0.0094 [0.0279]		-0.037 [0.0252]	-0.0307 [0.0162]*
Largest shareholder	0.0002 [0.0004]		0.0003 [0.0004]	
Working capital fin. by foreign commercial banks (b)	0.003 [0.0026]		0.0046 [0.0026]*	0.0045 [0.0026]*
Working capital financed by informal sources (b)	0.0011 [0.0008]		0.0002 [0.0003]	
Dummy for external auditory (b)		-0.1669 [0.0911]*		-0.1817 [0.0812]**
Quality, innovation and labor skills:				
Dummy for ISO quality certification (b)	0.0375 [0.0258]		0.0304 [0.0175]*	0.036 [0.0180]*
Dummy for new product (b)	-0.0234 [0.0310]		0.007 [0.0205]	
Dummy for discontinued product line (b)	-0.0316 [0.0264]		-0.0185 [0.0143]	
Dummy for outsourcing (b)		-0.0421 [0.0192]**		-0.0267 [0.0138]*
Staff - management	0.0009 [0.0012]		0.0013 [0.0010]	
Staff - non-production workers	-0.0009 [0.0007]		-0.0008 [0.0006]	
Dummy for training (b)				-0.0231 [0.0146]
Training for unskilled workers (a)	0.0015 [0.0023]		0.00001 [0.0020]	
University staff (b)	-0.0007 [0.0007]		-0.0012 [0.0005]**	-0.0013 [0.0005]**
Manager's experience (b)	0.002 [0.0102]		-0.0063 [0.0073]	
Dummy for closed plant		-0.0463 [0.0210]**		
Other control variables:				
Age (b)	-0.0004 [0.0005]		-0.0002 [0.0003]	
Share of the local market (b)	0.0002 [0.0004]		0.0002 [0.0003]	
Capacity utilization (b)		-0.0018 [0.0009]**		
Constant	Yes	Yes	Yes	Yes
Industry/region/size dummies	Yes	Yes	Yes	Yes
Observations	586	594	586	594
R-squared	0.22	0.25	0.24	0.24

See footnotes in Table 6.1.
Source: Authors' estimations with ICSs data.

Table 6.4: TANZANIA, Linear probability models for the probability of observing TFP and sales

Dependent variables:	Missing on TFP		Missing on sales	
	[1]	[2]	[3]	[4]
Explanatory variables:	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.	Coeff. Std. Err.
Infrastructures:				
Electricity from own generator (b)	-0.0007 [0.0014]		-0.0007 [0.0015]	
Losses due to power outages (b)	0.0035 [0.0050]	0.0049 [0.0023]**	0.0021 [0.0031]	
Losses due to water outages (b)				
Water from own well or water infrastructure (a)	0.00001 [0.0030]		0.001 [0.0023]	
Losses due to phone outages (a)	-0.0308 [0.0158]*		-0.0219 [0.0157]	
Transport outages (a)	-0.0125 [0.0349]		-0.0406 [0.0264]	
Losses due to transport delay (b)				-0.0067 [0.0020]***
Dummy for own roads (b)	-0.1213 [0.0768]		-0.0904 [0.0977]	
Dummy for webpage (b)	0.061 [0.0795]		0.0322 [0.0775]	
Wait for a water supply (a)	0.0192 [0.0249]		-0.0178 [0.0271]	
Low quality supplies (a)	-0.0035 [0.0109]		-0.0053 [0.0087]	-0.0025 [0.0013]*
Days of inventory of main supply				0.0358 [0.0175]**
Red tape, corruption and crime:				
Gift to obtain an operating license (b)	-0.0519 [0.0754]		-0.0152 [0.1104]	
Payments to deal with bureaucratic issues (b)	-0.0592 [0.0227]**	-0.0803 [0.0147]***	-0.045 [0.0267]	-0.0648 [0.0151]***
Days in inspections (b)	-0.0509 [0.0378]	-0.0788 [0.0403]*	-0.0241 [0.0387]	
Payments to obtain a contract with the gov. (b)	-0.0092 [0.0039]**	-0.0117 [0.0034]***	-0.0063 [0.0046]	-0.01 [0.0040]**
Security expenses (b)	-0.0023 [0.0026]		-0.0035 [0.0028]	
Illegal payments for protection (b)	-0.0075 [0.0224]		-0.0385 [0.0072]***	-0.0405 [0.0095]***
Finance:				
Dummy for credit line (b)				-0.1182 [0.0657]*
Interest rate of the loan (a)	0.0033 [0.0076]		-0.0017 [0.0061]	
Loans denominated in foreign currency (b)				-0.0014 [0.0009]
Dummy for current or saving account (b)		0.1616 [0.0856]*		0.2347 [0.0706]***
Working capital financed by commercial banks (b)	-0.0009 [0.0007]		-0.0011 [0.0010]	
Working capital financed by leasing (b)	-0.0059 [0.0023]**		-0.0059 [0.0013]***	
Inputs bought on credit (b)		-0.0016 [0.0008]*		
Sales bought on credit (b)	0.0007 [0.0012]		0.0007 [0.0011]	
Delay in clearing a domestic currency wire (a)	0.2385 [0.1403]*		0.196 [0.1479]	
Quality, innovation and labor skills:				
Dummy for new product (b)	0.0087 [0.0501]		0.002 [0.0462]	
Dummy for foreign license (b)				-0.2748 [0.0649]***
Dummy for upgraded product (b)				-0.1705 [0.0752]**
Dummy for new technology (b)		0.1973 [0.0631]***		0.3095 [0.0721]***
Dummy for joint venture (b)		-0.2179 [0.0796]**		
Dummy for outsourcing (b)		-0.2066 [0.0960]**		
Dummy for brought in house (b)		-0.2265 [0.0707]***		
Staff - skilled workers (b)	0.0007 [0.0004]*		0.0009 [0.0003]***	
Staff - professional workers (b)		-0.0055 [0.0033]		-0.0075 [0.0040]*
Workforce with computer (b)	0.003 [0.0017]*	0.0055 [0.0017]***	-0.0007 [0.0014]	0.0026 [0.0015]*
Dummy for training (b)				-0.0954 [0.0596]
Other control variables:				
Dummy for incorporated company (b)	0.012 [0.1990]		-0.075 [0.1534]	
Dummy for FDI (b)	0.1112 [0.0636]*	0.1255 [0.0549]**	0.1049 [0.0618]*	0.1717 [0.0586]***
Dummy for industrial zone (b)		0.121 [0.0737]		0.1274 [0.0668]*
Constant	Yes	Yes	Yes	Yes
Industry/region/size dummies	Yes	Yes	Yes	Yes
Observations	262	262	262	262
R-squared	0.18	0.22	0.16	0.3

See footnotes in Table 6.1.
Source: Authors' estimations with ICSs data.

Besides gathering evidence to show which are the variables empirically associated with the MDM, the main motivation for these models, is to know to what extent we need to control for IC variables in the estimation of equation (5). Bear in mind that even when the MDM is assumed to be MAR, we still need the following moment condition:

$$E(s_{it}u_{it} \mid s_{it}\log L_{it}, s_{it}\log M_{it}, s_{it}\log K_{it}, s_{it}IC_i, s_{it}D_{it}) = s_{it}E(u_i \mid s_{it}\log L_{it}, s_{it}\log M_{it}, s_{it}\log K_{it}, s_{it}IC_i, s_{it}D_{it}) = 0,$$

and therefore independence between the set of IC variables we are interested in (those of equations (1) and (5)) and the MDM is achieved only before controlling for any variable correlated with the MDM. At this point, in setting up our model, the question is whether it is enough to use the matrix of IC variables of equations (1) and (5) or, on the contrary, we have to find a better model for the MDM.

The results illustrate the clear relation between the MDM and the IC. Whether we use missingness in TFP (model [2]) or in sales (model [4]), those IC variables are able to explain a large proportion of the variance of the MDM. Furthermore, the results come to confirm the analysis of section 4.3, auditing, innovative activity, financing, capacity, corruption or informality among others are significant covariates of the pattern of missing data in all the countries, even after controlling for size, industry and region effects.

Moreover, the IC variables used as covariates of equation (1) present high correlation with the MDM, especially in Turkey (see specifications [1] and [3]), supporting the assumption of exogenous sampling selection, with the IC variables influencing the data generating process. Thereby, controlling for those IC variables becomes a requisite.

The question that arises at this point is whether it is enough to control for the IC variables of equation (1)—those of specifications [1] and [3]—, or rather should we select the set correlates of the MDM from the whole set of IC variables, as in specifications [2] and [4]?. In this respect, we argue that models [1] and [3] incorporate most of the information we require on the IC. In order to test it, we perform likelihood-ratio tests between model [1] on the one hand and [1] plus [2] on the other. Symmetrically, in the case of sales, we compare model [3] with [3] plus [4]. In addition, we also compare the R^2, AIC and BIC criterions of model [1] with that of model [1] plus [2] ([3] with [3] plus [4] for sales). Given these results, in the remaining part of the paper we only control for the IC variables included in equation (1).[30]

4.5 Some exhibits on the plausible correlation of PF variables and MDM

The descriptive analysis of the MDM is completed in figures 3.1 to 3.4. These figures compare the probability of picking an establishment with complete information for all production function variables with the probability of selecting an establishment with information for sales (panel A) and at least one missing value in the remaining PF variables. Panels B, C and D, simply change sales for materials, capital and employment respectively. The aim of these figures is to determine to what extent the pattern of missing values is correlated with PF variables. If the probability mass of picking a firm with a missing value is accumulated around low values of sales, materials, capital and employment, it could indicate that having a missing value is negatively related to the level of sales, materials, labor and/or capital. In other words, the probability of randomly drawing a firm with

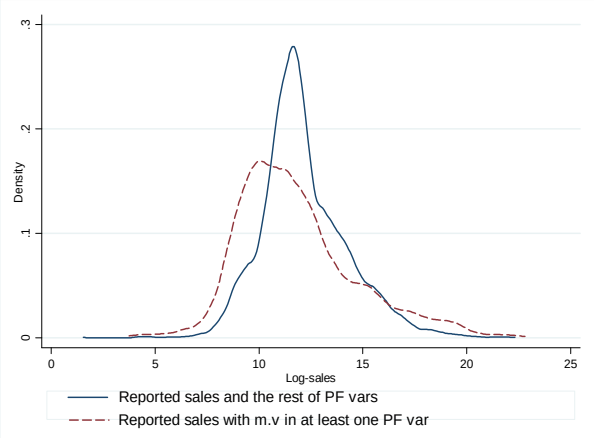
[30] We also believe that there exists a clear trade-off between parsimony and simplicity in the specification and adding further controls for the MDM

information for sales and with, at least, one PF variable missing is higher in firms with low sales. The same holds for materials and employment. The probability is lower for the case of capital. The same pattern is observed in India, Turkey, South Africa and Tanzania.

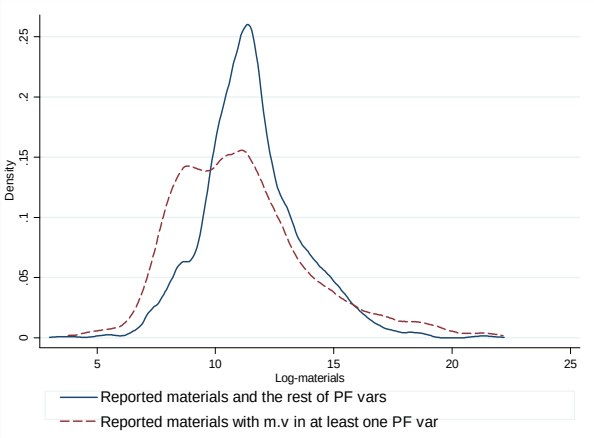
Figure 3.1: INDIA, Kernel density estimates of PF variables

(without M.V in PF variables and with M.V in any PF variable)

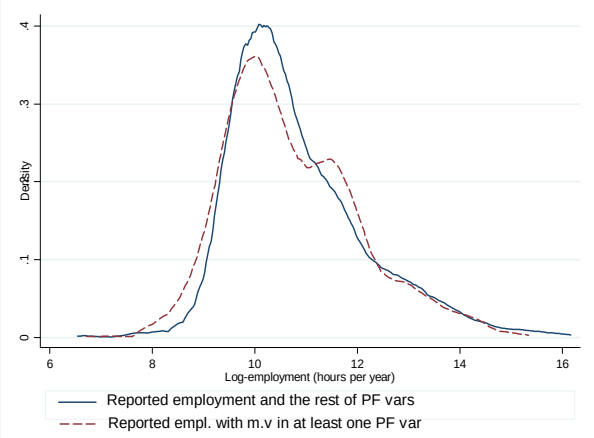
A. Sales



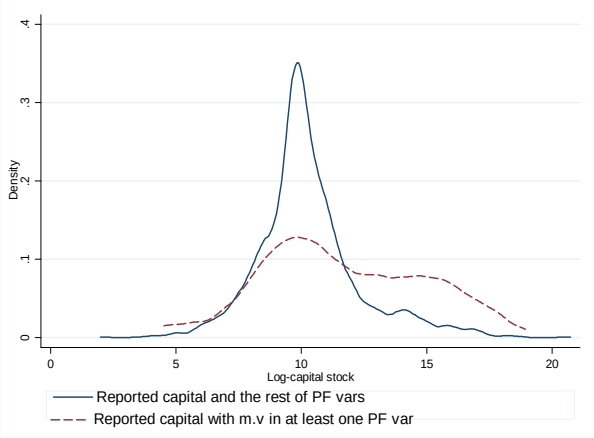
B. Materials



C. Labor



D. Capital



Notes:

Reported X and the rest of PF variables is the distribution of those establishments reporting all PF variables

Reported X with m.v in at least one of the rest of P.F is the distribution of those establishments reporting the corresponding PF variable and also reporting at least one missing value in the remaining PF variables

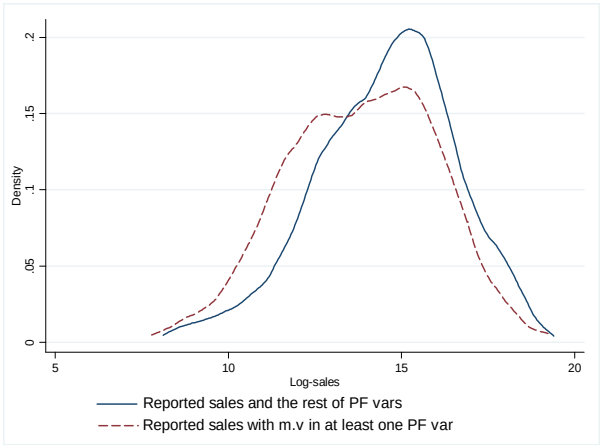
Epanechnikov kernel. Each point estimated within a range of 300 values.

Source: Authors' estimations with ICSs data.

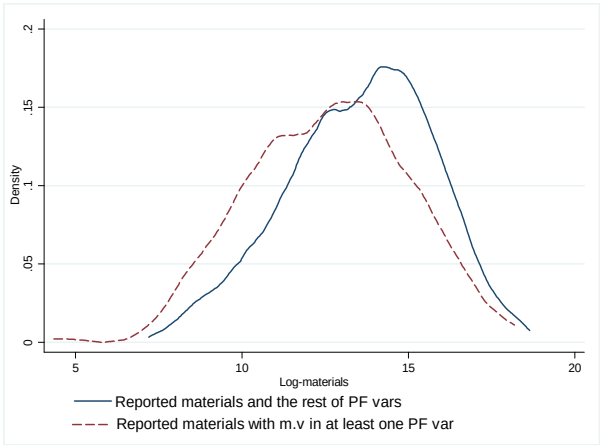
Figure 3.2: TURKEY, Kernel density estimates of PF variables

(without M.V in PF variables and with M.V in any PF variable)

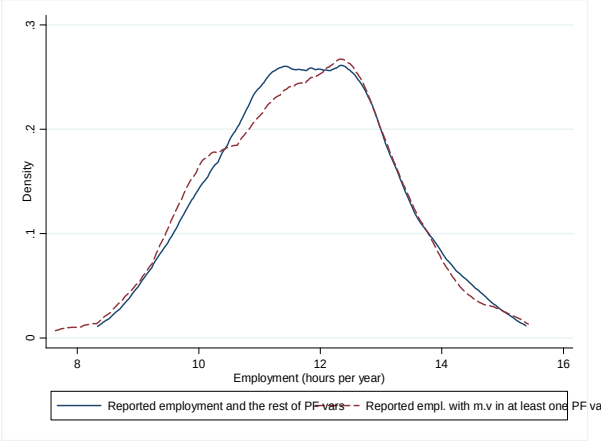
A. Sales



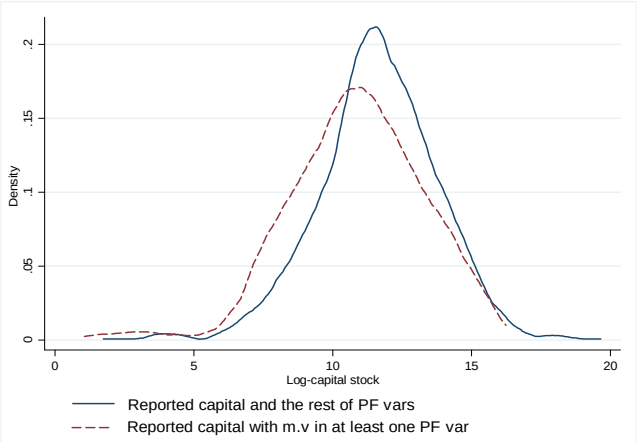
B. Materials



C. Labor



D. Capital

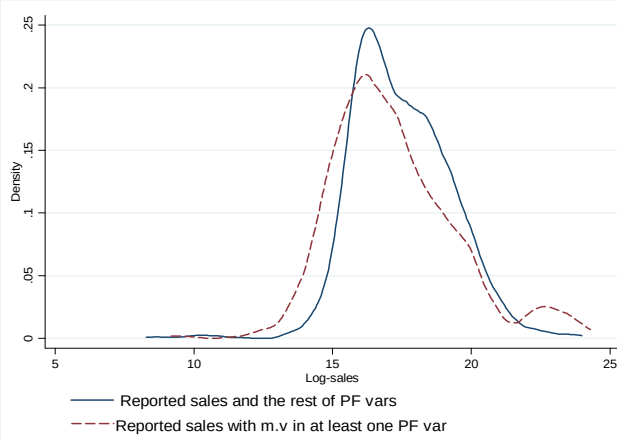


Notes:
Reported X and the rest of PF variables is the distribution of those establishments reporting all PF variables
Reported X with m.v in at least one of the rest of P.F is the distribution of those establishments reporting the corresponding PF variable and also reporting at least one missing value in the remaining PF variables
Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

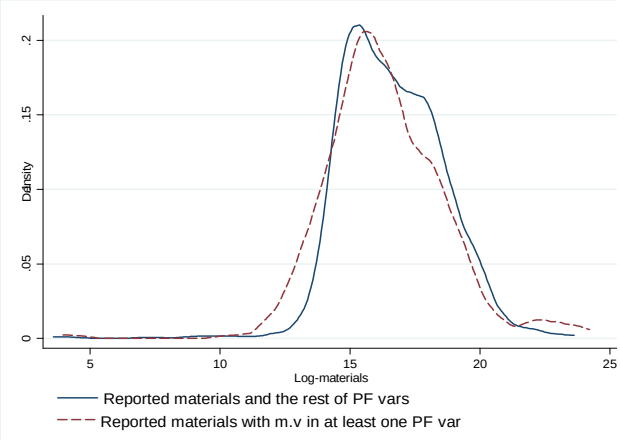
Figure 3.3: SOUTH AFRICA, Kernel density estimates of PF variables

(without M.V in PF variables and with M.V in any PF variable)

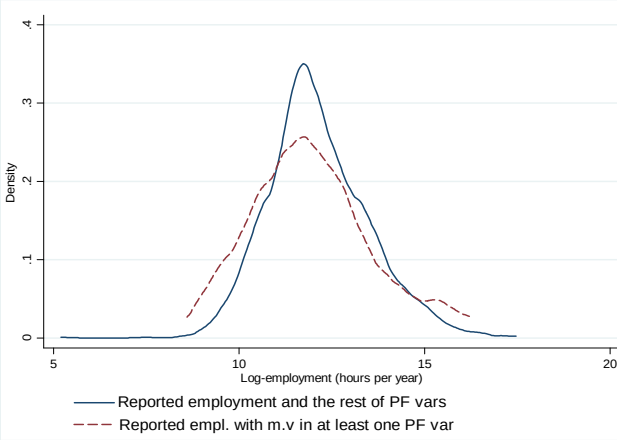
A. Sales



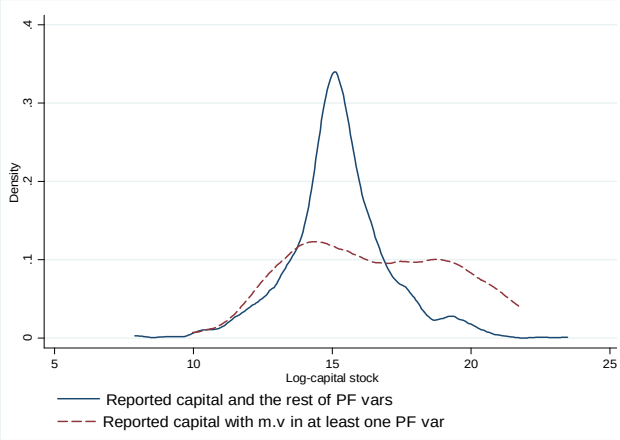
B. Materials



C. Labor



D. Capital



Notes:

Reported X and the rest of PF variable sis the distribution of those establishments reporting all PF variables

Reported X with m.v in at least one of the rest of P.F is the distribution of those establishments reporting the corresponding PF variable and also reporting at least one missing value in the remaining PF variables

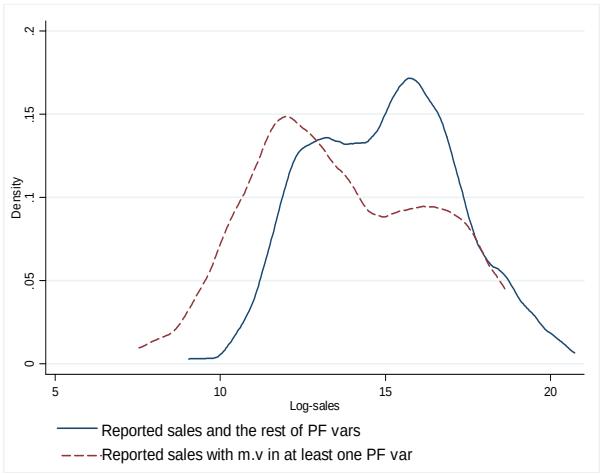
Epanechnikov kernel. Each point estimated within a range of 300 values.

Source: Authors' estimations with ICSs data.

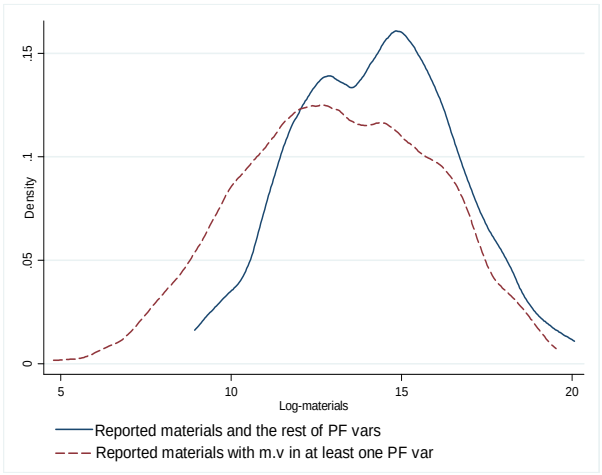
Figure 3.4: TANZANIA, Kernel density estimates of PF variables in Tanzania

(without M.V in PF variables and with M.V in any PF variable)

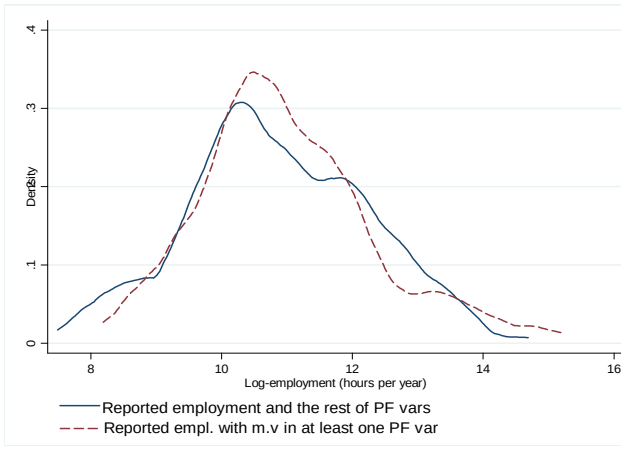
A. Sales



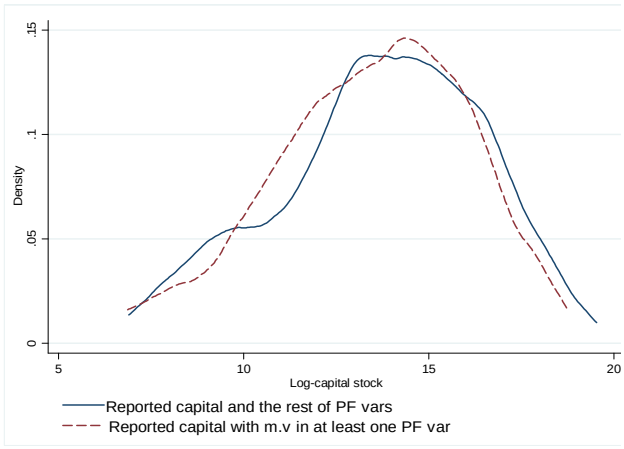
B. Materials



C. Labor



D. Capital



Notes:
Reported X and the rest of PF variables is the distribution of those establishments reporting all PF variables
Reported X with m.v in at least one of the rest of P.F is the distribution of those establishments reporting the corresponding PF variable and also reporting at least one missing value in the remaining PF variables
Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

Figures 3.1 to 3.4 support the story of weaker firms reporting more missing values. However, the story is not yet conclusive. Firms with low sales (and materials, capital and employment) do not usually need proper accountability also tend to operate in more corrupt environments and are less innovative and dynamic. In addition, as most of the firms are accumulated around low values, it is easy to infer that the probability of picking a firm with any missing value in the PF variables will be higher within this range of values as well. From these figures we cannot conclude that low sales do not imply weakness or low productivity, and therefore higher probability of having missing values.

4.6 Can we relate the MDM and our endogenous variables by means of the ICSs?

So far we know that the MDMs in the countries analyzed are, in some way, related with a number of firms' attributes, such as accountability, corruption, openness, informality or size. However, we are not still able to conclude whether the MDM is determined independently of sales and TFP. The debate would probably end if we were able to construct a model of the probability of having a missing value and productivity (or sales) as RHS variable. Unfortunately, this is not possible because, obviously, we do not observe either productivity or sales when we observe a missing value. However, we can still take advantage of the particular structure of the pattern of missing values to relate it with productivity or sales. Since the number of missing values reported increases when we move backwards in time, we can construct a model relating the probability of having a missing value in any PF variable in period t and productivity (tfp) in period t+1 plus other controls. That is, assuming that information in t+1 is better than in period t—bearing in mind that establishments report *recall* data—we propose the model below for the probability of having a missing value

$$\Pr(s_{it}^{a}=1 \mid tfp_{it+1}, D_{it}, IC_{i}) = \varphi(\delta_0^{a} + \delta_1^{a} tfp_{it+1} + \delta_2^{a\prime} D_{it} + \delta_3^{a\prime} IC_{i} + \zeta_{it}^{a}),$$

where s^a takes value 1 if we observe all sales, labor, materials and capital and 0 otherwise.[31] Or alternatively we can also use the following model for sales

$$\Pr(s_{it}^{b}=1 / y_{it+1}, D_{it}, IC_{i}) = \varphi(\delta_0^{b} + \delta_1^{b} y_{it+1} + \delta_2^{b\prime} D_{it} + \delta_3^{b\prime} IC_{it} + \zeta_{it}^{b}),$$

where s^b takes value 0 if we do not observe sales and y is the logarithm of firms' sales.

The question we are trying to answer with these kinds of models is whether the probability of observing a missing value in period t-1 is correlated with the level of sales (productivity or TFP) in period t. Or, in other words, are more productive/profitable firms more likely to keep track of their input/output accountability? Obviously, these models do not imply contemporaneous correlations but we think they might still be a good indicator of the actual relation between the level of sales/TFP and the MDM. On the other hand, an additional consideration should be noted; there is a selection bias in the models as we are only able to use those observations with observable sales or TFP in t+1, so the resulting sub-sample is likely to be biased toward those responding firms. In order to reduce the degree of the bias, we use those imputed values of sales or TFP in period t+1.[32]

[31] In addition, if we assume a first order Markov process for productivity, $\Pr(tfp_{t+1}/\ tfp_t, tfp_{t-1},\ldots) = \Pr(tfp_{t+1}/\ tfp_t)$ and therefore *tfp* in t+1 is a good proxy of *tfp* in period t the model is reduced to $\Pr(s_{it}^{a}=1/tfp_{it}, D_{it}) = \varphi(\delta_0 + \delta_1 tfp_{it} + \delta_2' D_{it} + \upsilon_{it})$.

[32] Although by applying this strategy we reduce the degree of sample bias, the problem remains to some extent. Nonetheless, we still believe that the models can be very informative about the relation of the plausible endogeneity of the MDM.

Table 7: Linear probability models for the effect of TFP and sales on the probability of observing a missing value in t+1

A. Missing in TFP [1]				
Dependent variables: for each country a dummy taking value 1 if we observe all labor, materials, capital and sales				
Explanatory variables	India	Turkey	South Africa	Tanzania
log TFP (t+1)	0.0168*	0.0183**	0.0212	0.0281
	[0.0091]	[0.0084]	[0.0180]	[0.0250]
IC variables [3]	Yes	Yes	Yes	Yes
Constant	Yes	Yes	Yes	Yes
Industry/region/size dummies	Yes	Yes	Yes	Yes
Observations	1476	426	454	87
R-squared	0.27	0.07	0.19	0.32

B. Missing in sales [2]				
Dependent variables: for each country a dummy taking value 1 if we observe sales				
Explanatory variables	India	Turkey	South Africa	Tanzania
log sales (t+1)	0.0063*	0.0069	0.0079	0.0033
	[0.0033]	[0.0043]	[0.0083]	[0.0144]
IC variables [3]	Yes	Yes	Yes	Yes
Constant	Yes	Yes	Yes	Yes
Industry/region/size dummies	Yes	Yes	Yes	Yes
Observations	1894	677	564	155
R-squared	0.17	0.05	0.16	0.14

[1] Missing in TFP takes value 1 if we observe all sales, materials, labor and capital, and 0 otherwise.
[2] Missing in sales takes value 1 if we observe sales and 0 otherwise.
[3] The set of IC variables of equation (1) is also included.
Both TFP and sales are used before imputing missing values.
Significance given by robust standard errors allowing for clustering by industry and region *** 1%, **5%, * 10%.
Source: Authors' estimations with ICSs data.

The results of both equations for missingness in TFP and sales are in Table 7. Under endogenous sampling when the pattern of missing values is correlated with sales or TFP and if we were able to observe everything, we should expect a positive relation between contemporaneous TFP/sales and the missingness problem before controlling for other determinants such as IC and D variables. As a consequence, the relation between missingness 'yesterday' and TFP/sales 'today' should also be positive. Table 7 supports this view for TFP (see Table 7 panel A) and for the cases of India and Turkey, where the $\hat{\delta}_1^a$ is positive and therefore more productive firms in year t+1 are associated with a higher probability of being able to keep track of proper accountability on output and inputs in past years. Note that we find this relation even before controlling for IC and D effects. However, the $\hat{\delta}_1^a$ for South Africa and Tanzania do not indicate any significant association between TFP and missingness in these countries. On the other hand, in the case of sales (panel B) we only observe a positive and significant effect of $\hat{\delta}_1^a$ in India, although the effect in Turkey is no longer significant. In South Africa and Tanzania the effect remains non-significant.

Therefore, Table 7 points to a plausible endogenous selection problem between missingness and TFP in India and Turkey, with the endogenous sampling selection problem corroborated in the case of sales in India but not in Turkey. On the opposite side, the analysis does not support this view in South Africa and Tanzania, neither in the case of sales nor TFP. Nonetheless, Table 7 does not allow us to conclude that there is a self-selection problem in India and Turkey, nor that the MDM is MAR in South Africa and Tanzania. At this point caution is a requisite. All we are able to say is that we have four different patterns of data generating mechanisms. For some of them we find evidence of a more likely self-selection problem and under which we can test the performance of the various imputation methods, including the Heckman models.

4.7 Conclusions on the nature of the MDM

The question at the core of the analysis of this section is whether the MDM in these countries is governed only by the level of sales or TFP (weakness) or if the MDM can be explained by a number of firms attributes, such as the level of competitiveness, dynamism, corruption, informality, accountability and other indicators relating to the firms' capacity: *MAR versus non ignorable missing data assumption*s.

According to the descriptive analysis presented, the MDM mechanism has to do with informality and corruption and also with the capacity of the firms. More dynamic firms engaged in R&D, quality, innovation of new products, technologies and operating in more exigent and competitive export markets tend to report fewer missing values. Accountability can by itself explain a large share of missing data too. Much of these variables indicate that weaker firms tend to avoid reporting PF figures, and size is in some cases a good indicator of weakness as section 4.1 indicated. All these patterns are, to a greater or lesser extent, common to all the countries analyzed.

Notwithstanding this clear relation between IC and MDM, we cannot reject the hypotheses of non-ignorability in any of the cases. As already pointed out, this assumption is untestable from the available data. The preliminary descriptive analysis of section 4.5 points to a relation between the level of usage of inputs and output and missingness. Furthermore, previous econometric analyses of section 4.6 report a plausible relation between TFP and sales and missingness in t-1, especially in the cases of India and Turkey. In either MAR or non-ignorable MDM, we believe that according to the analysis presented, controlling for those IC and D variables related with the missingness mechanism is a requisite, as can be shown from the LPM models presented for the probability of observing the required data to construct sales or TFP measures. This is the way we proceed in the rest of the paper.

The aim of the following sections is to explore the dichotomy "*MAR versus non-ignorability"* of the MDM and their effects on the imputation mechanism proposed by comparing the sensitivity of the results of estimating the extended production function (1) under two assumptions: first, MDM is ignorable and therefore it may be explained by a number of exogenous firms' characteristics; and second, the MDM is endogenous and intimately linked to the level of sales and TFP of the firms. We also take advantage of the heterogeneity of the aprioristic relations observed between the MDM and their determinant in the four countries considered. This will allow us to illustrate how sensitive the results are under very different assumptions.

In addition, besides testing the non-ignorable MDM, the analysis we present in what follows also allows us to study how the sensitivity of the imputations from the ICA method responds to: first, additional assumptions, such as randomness, or the amount of information embodied in the ICA method, all of them requiring the MAR assumption; and second, to different patterns of missing data: Turkey and Tanzania with a response rate for sales and TFP lower than 40% and India and South Africa with more than 70% of observations reported.

5. Robustness analysis

As indicated, the aim of the paper is to compare the results of estimating equation (1) under the ICA method and several alternative imputation procedures. The methods presented to test the robustness of the results have their origins in two distinct bodies of statistical literature. The first one is related with likelihood-based inference with incomplete data, in particular, the EM algorithm. The second concerns the techniques of Markov Chain Monte Carlo (MCMC), generally referred to as multiple

imputation. We also consider extensions of the ICA method, allowing for additional randomness in the imputation procedure and the selection of the explanatory variables in equation (3). Lastly, we consider the estimation of (1) by sample selection estimation, such as different Heckman models.[33]

The literature on missing data points to the advantages of modern imputation mechanisms—EM-type algorithms and MCMC simulations—over other simpler methods based on basic standard regression techniques (such as the ICA method presented), see Allison (2001) and Little and Rubin (1987) for a review. Nonetheless, while most of these techniques have been widely evaluated under univariate missing data patterns (missingness for only one variable), or simple patterns of missingness in some of the variables of the dataset, the patterns of missing data observed in ICSs are very complex and unbalanced, even if we only consider PF variables and not the remaining IC variables. As an additional objective, it raises the possibility of evaluating the performance of modern imputation mechanisms under the complex and very different patterns of missing data observed in ICSs.

5.1 The *ICA Method* as an EM type algorithm

The EM algorithm has been widely applied in a broad range of applications, from missing data to latent variables models. Here we present several EM algorithms that will serve as a benchmark to be compared with the *ICA method* proposed.

In particular, the aim is to test the sensitivity of the results obtained from the ICA method compared with other more sophisticated imputation mechanisms allowing for an additional randomness and amount of information embodied in the imputation mechanism. EM-type algorithms are based on an underlying likelihood function of the process generating data, and as a consequence imputed missing data is based on draws from the posterior predictive distributions of the postulated missing data mechanism (or data generating process). A key issue under these mechanisms is whether the MDM may be considered as MAR or not.

5.1.1 EM-Algorithm on size, industry and region

Let J denote the vector dependent variable of interest, determined by the underlying unobserved vector variable J_{Mis}. Let $f^*(J_{Mis} \mid \mathbf{X}, \theta) = 0$ be the joint density of the latent variables conditional on the matrix of observed regressors $\mathbf{X}$, and let $f(J \mid \mathbf{X}, \theta) = 0$ be the joint density of the observed variables. In essence, the maximum likelihood estimator (MLE) in this case maximizes

$$Q_N(\theta) = \frac{1}{N} L_N(\theta) = \frac{1}{N} \ln f^*(J_{Mis} \mid \mathbf{X}, \theta) - \frac{1}{N} \ln f(J_{Mis} \mid J, \mathbf{X}, \theta) .^{34} \quad (6)$$

[33] Note that although in this section we only analyze the behavior of PF variables as if they were the only set of imputed variables, IC variables are in all the cases imputed by the ICA method.

[34] Note that J* uniquely determines J but the inverse is not true, that is, J does not uniquely determine J*; from the Bayes Rule it follows that $f(J \mid \mathbf{X}, \theta) = f^*(J^* \mid \mathbf{X}, \theta) / f^*(J^* \mid J, \mathbf{X}, \theta)$ (see Cameron and Trivedi, 2005).

The first term is not observed and therefore it is ignored. The second term is replaced by its expected value which does not involve J_{Mis}. The process is iterative; at the *r-th* round the expectation of the second term is evaluated at $\theta = \hat{\theta}_r$. The Expectation step of the algorithm therefore calculates

$$Q_N(\theta \mid \hat{\theta}_r) = -E\left[\frac{1}{N}\ln f(J_{Mis} \mid J, \mathbf{X}, \theta) \mid J, \mathbf{X}, \hat{\theta}_r\right]. \tag{7}$$

The Maximization step simply maximizes $Q_N(\theta \mid \hat{\theta}_r)$ to compute $\hat{\theta}_{r+1}$. Note that the iterative process continues until convergence is achieved.

In this paper, we follow Cameron and Trivedi (2005) and propose the next EM type algorithm with our model rewritten as

$$\begin{bmatrix} J_1 \\ J_{Mis} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{8}$$

Where N_1 are the available observations and N_2 the missing observations and **X** denotes the explanatory variables, the EM algorithm consists of (1) estimating $\hat{\beta}$ using the N_1 available observations; (2) generating $\hat{J}_{Mis} = \mathbf{X}_2\hat{\beta}$; (3) in order to mimic the distribution of J_1 generating adjusted values of $\hat{J}^a_{Mis} = (\hat{\mathbf{V}}^{-1/2}\hat{J}_{Mis}) \otimes \mathbf{u}_m$, where $\mathbf{u}_m$ is a Monte Carlo draw from the $N(0, s^2)$ distribution, being s^2 the variance of u_1 and a estimate of V can be obtained as $\hat{\mathbf{V}}(\hat{J}_{Mis}) \equiv \hat{\mathbf{V}}(\hat{J} \mid \mathbf{X}_2) = s^2(I_{N_2} + \mathbf{X}_2[\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_2')$, and $\otimes$ denotes element by element multiplication; (4) using the augmented sample obtain a revised estimate of $\hat{\beta}$; (5) repeating steps (1) to (4) until convergence is achieved, in the sense that the change in the sum of the square residuals becomes arbitrarily small.

Note that steps (3) and (4) are simply random draws from the conditional distributions of J given β in the case of step (3), and of β given s^2 in the case of step (4). In this first case, by means of direct comparisons with the ICA method, we include in the matrix **X** only the industry, region and size dummies. We also exclude from the imputation those observations with all production function variables missing.

Note the advantages of the EM algorithms over the ICA method. Since the EM algorithm works on the posterior predictive density, after each replication the new estimation of $\hat{\beta}$ improves the previous one—because in each iteration we are approaching the postulated distribution of the mechanism generating data. In addition, theoretically the estimates of s^2 improve the ones obtained in the *ICA method,* as those are likely to be downward biased as they do not make allowance for the uncertainty inherent in J_{Mis}. Obviously, these advantages greatly depend on the specification (model) chosen for the EM algorithm.

5.1.2 Extended EM-Algorithm on PF variables

The first alternative model for the EM algorithm is to extend matrix **X** to contain industry, region size, dummies and production function variables. The imputation now has two iterative processes. The first iteration process is the iterative EM algorithm per se, while the second one consists of replacing missing cells conditional on the information available for the remaining production function variables and the patterns of missing values observed (see Figures 1 to 4). We start by

replacing the production function variable with the larger amount of missing values where **X** contains the remaining PF variables. We continue by applying the EM algorithm to the remaining PF variables.

5.1.3 Extended EM-Algorithm on PF and IC variables

In order to check the sensitivity of the results to the matrix **X** used, and therefore to the amount of information embodied in the EM algorithm, we include in this case industry/region/size dummies, PF variables and a large set of IC variables. Concretely, the set of IC variables comes from the significant IC variables of equation (1). The idea is to check how the EM algorithm responds to the amount of information incorporated in the imputation mechanism. Different results with respect to EM algorithms in sections 5.1.1 and 5.1.2 would pose some doubts about the validity of the ICA method, as it does not incorporate enough information in the imputation mechanism.

5.2 Further extensions of the *ICA method*

We now extend the ICA method to meet additional assumptions on the MDM. In particular we develop the ICA method to incorporate some degree of randomness in the imputation. We also propose an ICA method in which the dependent variable of the model (sales or logY) is excluded from the imputation procedure.

5.2.1 Random industry-region-size replacement: *random ICA Method*

Under the two assumptions mentioned in section 3 (normality of replaced variables and linearity, apart from the MAR assumption) the ICA method leads to consistent estimation of the parameters of equation (1). However, it could be argued that a more efficient method might be used. Notice that by imputing missing values we are modifying the population distribution of replaced variables. In particular, if the two conditions mentioned in section 3 hold the sample average of the modified distribution of the variable it converges with the population expectation. Unfortunately, this is not true in the case of the standard deviation. With the replacement strategy we are reducing the variability of the distribution of those variables with missing values and therefore any statistical inference will be based on downward biased standard errors. Moreover, the bias in the standard errors will be higher as the proportion of missing values increases and the sample size decreases.

This problem will arise whenever we use imputed data as if it were real data. It has to do with the lack of uncertainty in the estimation of the parameters of estimating regressors equations and reflects the fact that conventional formulas to compute standard errors do not correct for imputed data.

The ICA method, although deterministic, introduces variability in the imputation of missing data by replacing missing cells for industries, regions and sizes with the variability given by *I*R*S* being *I, R* and *S* the numbers of industries, regions and sizes respectively. A good question is therefore whether this variation is enough or if the ICA method leads to downward biased standard errors. To answer this, we propose an alternative variation of the ICA method which consists of adding a random part to each imputed value.

The new replacement strategy is again based on the expectation of equation (3), but in this case a random term is added in order to embody uncertainty to the imputation mechanism

$$\tilde{J}_{it} = \hat{\rho}_0 + \hat{\rho}_{R,J} D_{R,it} + \hat{\rho}_{I,J} D_{I,it} + \hat{\rho}_{S,J} D_{S,it} + \hat{\sigma}_{J,\varepsilon} \xi_{J,it} \qquad J = Y, L, M, K \tag{9}$$

where $\hat{\sigma}_{J,\varepsilon}$ is the standard error of the residual $\varepsilon_{J,it}$ from

$$J_{it} = \rho_0 + \rho_{R,J} D_{R,it} + \rho_{I,J} D_{I,it} + \rho_{S,J} D_{S,it} + \varepsilon_{J,it} \qquad J = Y, L, M, K$$

and $\xi_{J,it}$ is a random draw from $\varepsilon_{J,it}$. In particular, we take 100 random draws from $\varepsilon_{J,it}$ constructing 100 candidate values to replace each missing cell in the data matrix. To make the definite replacement we compute the average across the 100 candidate values.

5.2.2 Random industry-region-size replacement: *bootstrap ICA Method*

Another problem arising from the lack of uncertainty inherent in deterministic imputation methods is that, generally, when certain instruments and/or regressors are estimated in a first stage (in our case for production function variables) the asymptotic variance needs to be adjusted because of the generated instruments, see Pagan (1984), Newey (1984), Murphy and Topel (1985) and Newey and McFadden (1994).[35]

A plausible solution for this problem is to compute the bootstrap estimate of the standard errors of the estimated coefficients of equation (5). The idea is to create '*r*' replications of the original sample using as strata industry and region. In the next step and for each replication, we apply equation (4) to replace the missing data and to estimate equation (5). The result will be a bootstrap distribution of the estimators of equation (4) under different replacements of missing data that can be used to compute the bootstrap estimates of the standard errors.

5.2.3 *ICA method* on the inputs

One can also look at the imputation of missing data in the dependent variable of equation (1), sales. In this respect, it can be argued that the MDM may be correlated with the dependent variable of (1), so imputing missing values in sales and estimate (2) by OLS or standard econometric techniques is not a valid solution. In this case, when *s* depends on log*Y*, it is clear that *s* and *u* are no longer uncorrelated, even though we control for *IC* and *D* variables. In particular when *s* is correlated with *logY* in equation (2) there is a self-selection problem that should be handled with other sample selection corrections, such as the Heckman model, as we shall see later on.

Here we propose the same replacement mechanism as in section 3, but in this case excluding the sales of the replacement process. The extended production function to be estimated is therefore

$$s_{it}^{**} \log Y_{it} = s_{it}^{**} (\alpha_0 + \alpha_L \log \tilde{L}_{it} + \alpha_M \log \tilde{M}_{it} + \alpha_K \log \tilde{K}_{it} + \alpha'_{IC} IC_i + \alpha'_D D_{it}) + s_{it}^{**} \tilde{\tilde{u}}_{it}, \tag{12}$$

with identification conditions symmetrical to those of equation (5).

Note that when there is no sample selection, incomplete data is MAR, the incompleteness of logY is not so large that it makes the complete case unrepresentative of the real population and we are not concerned with efficiency, estimating (12) by standard techniques is equivalent to estimating

[35] More precisely, the problem appears when testing the null hypotheses $H_0 : \psi = 0$, where $\psi = \alpha, \beta, \delta, \omega$ are the coefficients of generated regressors (see equation 1). Before including the generated regressors in (1), the usual test statistic on ψ has a limiting standard normal distribution under H_0. However, when $\psi \neq 0$ standard t statistics will not be asymptotically valid and an adjustment is needed for the asymptotic variances of all estimators of generated regressors.

(5) or (2). On the contrary, when there is a sample selection problem, the point of reference to compare with (12) would be the Heckman selection model.

5.3 Multiple imputation via switching regression

The aim now is to propose different imputation mechanisms to compare their performance with the ICA method and its variations. The following imputation mechanism was first proposed by van Buuren, Boshuizen and Knook (1999) and it has been chosen because it fits very well with datasets with a large amount of missing values in many variables, such as IC datasets. See also Schafer (1999) for a tutorial on multiple imputation, and Schafer (1997) and Gelman, King and Liu (1998) for applications.

The basic idea is to create a small number of data copies, each of which has the missing values suitably imputed. Each imputed dataset is then analyzed independently. Estimates of the parameters of interest are properly averaged across the data copies, while standard errors are computed according to *'Rubin rules'*, see Rubin (1987). In particular, this multiple imputation mechanism is accomplished in the following steps:

1. Specify the posterior predictive density of incomplete data as $p(J_{MIS}|X,s)$ given that the non-response mechanism is $p(\ s \mid J, IC, C, D)$ and the complete data model is p(J, *IC, C, D*), where X is the set of covariates used in the imputation mechanism and s is the pattern of missing values. The posterior predictive density is generally given by

$$p(J_{MIS} \mid X,s) = \int p(J_{MIS} \mid X,s,\theta) p(\theta \mid X,s) d\theta \tag{13}$$

   where the standard procedure to impute missing data consists of first, drawing a value of θ^* from $p(\theta \mid X,s)$ and second, drawing a value J_{Mis}* from $p(J_{MIS} \mid X,s,\theta = \theta^*)$.
2. The next step is to draw imputations from this density to produce *m* complete datasets. Here we follow van Buuren et al. (1999) and we produce m=5 datasets.
3. Estimate equation (1) *m* times.
4. Pool the *m* results.

This imputation mechanism involves choosing the form of the linear model and the predictor variables. In particular, we use a linear regression of each J_{MIS}= *Y, L, M and K* on a set X of predictor variables, where the set of predictor variables is given by *X=Y, K, L, M,* and *D*. Note that each *J* is used as a predictor variable and as an imputed variable in (10), while *D* are used only as predictor variables.

5.4 Sample selection correction (I): Heckman on complete case

If the pattern of missing values is endogenously determined (it is correlated with output (*logY*) in equation (4)), thereby giving rise to a self-selection problem, the ICA method may lead to inconsistent estimates of parameters of (1). In these cases one has to implement the Heckman (1976) or Heckit method to correct for self-selection, since OLS applied either to the complete case or to the sample with replacement is inconsistent. In particular, the Heckman model over the complete case is given by

$$E(\log Y_{it} \mid \log L_{it}, \log M_{it}, \log K_{it}, IC_i^H, D_{it}, s_{it}=1) = \alpha_0 + \alpha_L \log L_{it} + \alpha_M \log M_{it} + \alpha_K \log K_{it} + \beta' IC_i + \omega' D_{it} + \rho\lambda(\gamma_L \log L_{it} + \gamma_M \log M_{it} + \gamma_K \log K_{it} + \gamma'_{IC} IC_i^H + \gamma' D_{it}) \quad (14)$$

where as usual $\rho\lambda(.)$ is simply the inverse of Mills ratio or Heckman's lambda given by the following Probit

$$\Pr(s=1 \mid l_{it}, m_{it}, k_{it}, IC_i^H, D_{it}) = \Phi(\gamma_L \log K_{it} + \gamma_M \log M_{it} + \gamma_K \log K_{it} + \gamma'_{IC} IC_i^H + \gamma' D_{it}), \quad (15)$$

with the following moment condition $E(u \mid \log L_{it}, \log M_{it}, \log K_{it}, IC_i, IC_i^H, D_{it}) = 0$.

The Heckman method is highly sensitive to model choice, requiring a good knowledge of the nature of the missing data mechanism. For this reason, the selection of the Probit model in (12) goes from the general to the specific, to select the variables with a significant effect on the probability of having a missing value. Concretely, the selection of variables starts with a wide set of more than 120 IC and D variables in each country. Eventually, the final set of significant variables is reduced to a number around 15 and 25.

5.5 Sample selection correction (II): Heckman imputing inputs with the *ICA method*

In 3.4 the selection of Heckman model is based on the complete case. In this section, we propose performing the same model on the sample after replacing missing values in employment, materials and capital according to equations (10) and (11). The Heckman model in this case is given by

$$E(\log Y_{it} \mid \log \tilde{L}_{it}, \log \tilde{M}_{it}, \log \tilde{K}_{it}, IC_i^H, D_{it}, s_{it}=1) = \alpha_0 + \alpha_L \log \tilde{L}_{it} + \alpha_M \log \tilde{M}_{it} + \alpha_K \log \tilde{K}_{it} + \beta' IC_i + \omega' D_{it} + \rho\lambda(\gamma_L \log \tilde{L}_{it} + \gamma_M \log \tilde{M}_{it} + \gamma_K \log \tilde{K}_{it} + \gamma'_{IC} IC_i^H + \gamma' D_{it}) \quad (16)$$

with Heckman's Lambda and moment condition obtained symmetrical to the previous sub-section. Note that equation (17) is directly comparable with equation (12).

In addition, in sections 5.2.1 and 5.2.2, we introduced the problem of lack of uncertainty in the estimation of the standard errors of estimating regressors equations. A solution proposed was to obtain the bootstrap standard errors under replacement of missing values in each resampling. The solution here is similar: we obtain the bootstrap standard errors to make statistical inference and to correct the aforementioned problem. More precisely, we will compare the standard errors from the estimating sample with the bootstrap estimator of the standard errors, which will give us a benchmark on how serious this issue is in our case.

6. Empirical results

The objective of this section is to evaluate to what extent the results obtained from the ICA method are influenced by different assumptions on the MDM. In particular, as we pointed out in section 5, under the ICA method we have to consider two different key assumptions on the patterns of missing data. First, if we can assume MDM as MAR, in which case then we test the goodness-of-fit of the ICA method against other more sophisticated mechanisms that are supposed to work better, as they consider the randomness issue and are able to include more information in the imputation mechanisms. And second, the MDM is non-ignorable and therefore we are forced to apply sample selection corrections such as Heckman models.

The evaluation of the ICA method is based on the kernel estimates of inputs and output and the underlying TFP densities under all the imputation mechanism proposed. We also present the empirical results from estimating the extended production function (1) under different imputation methods. In all the cases, we use the ICA method as a benchmark for comparison purposes. In all the regressions, outliers, defined as those observations with ratios of labor cost to sales and/or materials to sales greater than one, are excluded.

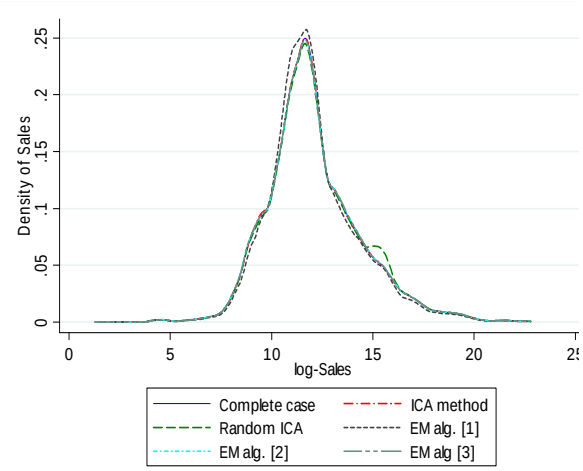
6.1 Evaluation of imputation mechanism: Comparison of estimated inputs and output densities

The kernel densities of $\log \tilde{Y}_{it}$, $\log \tilde{L}_{it}$, $\log \tilde{M}_{it}$, $\log \tilde{K}_{it}$ for each country and for the complete case, the ICA method, the random ICA method and the three EM-type algorithms considered are in figures 4.1 to 4.4. In turn, the descriptive statistics of the variables under each imputation mechanism are in tables 8.1 to 8.4.

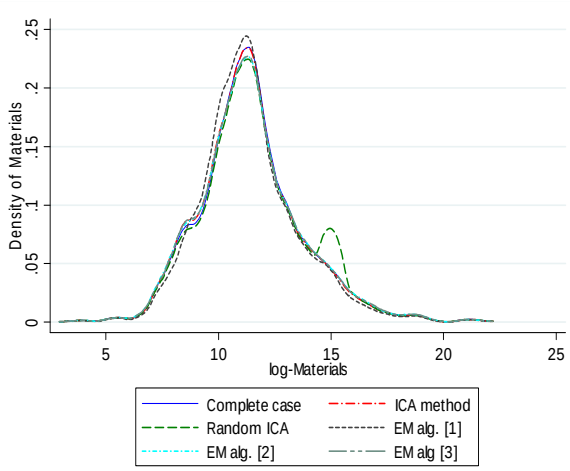
Figure 4.1: INDIA, comparison of the ICA method and other imputation mechanisms for PF variables

I. Kernel[1] estimates of output and input densities

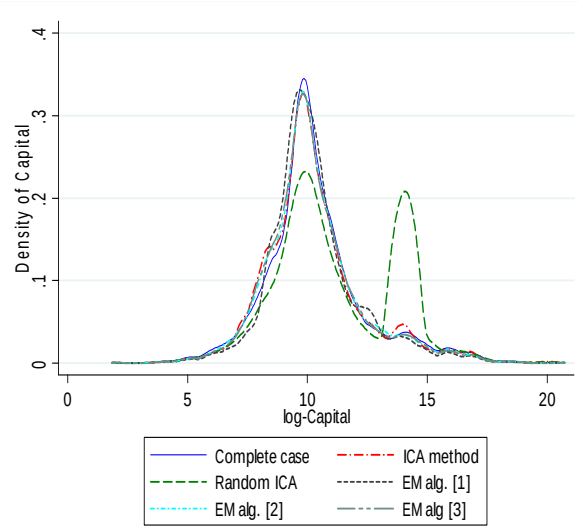
A. Sales (log)



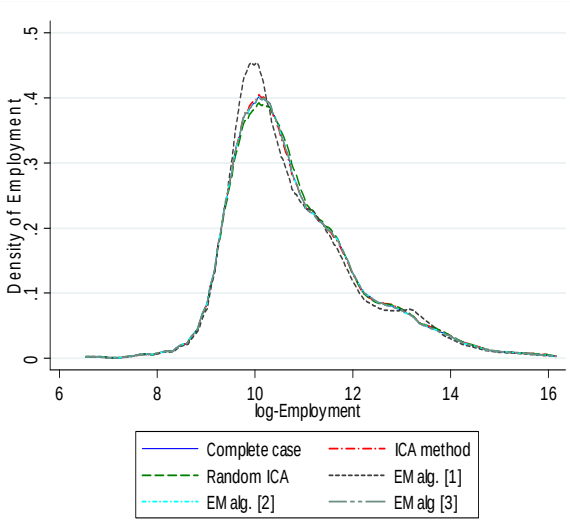
B. Materials (log)



C. Capital stock (log)



D. Employment (log)



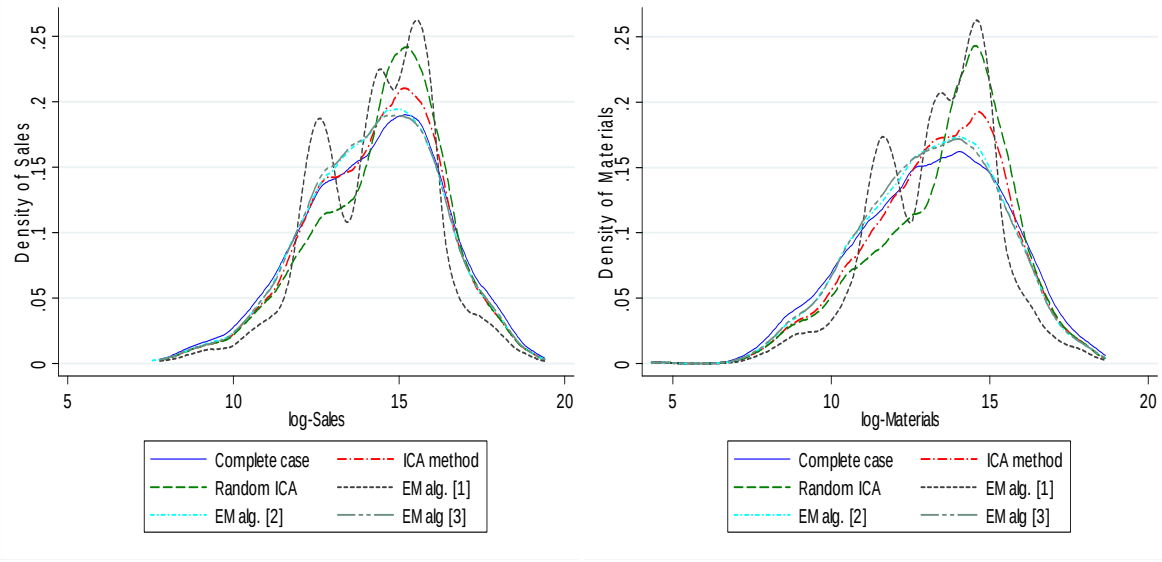
Notes:
1 Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

Figure 4.2: TURKEY, comparison of the ICA method and other imputation mechanisms for PF variables

I. Kernel[1] estimates of output and inputs densities

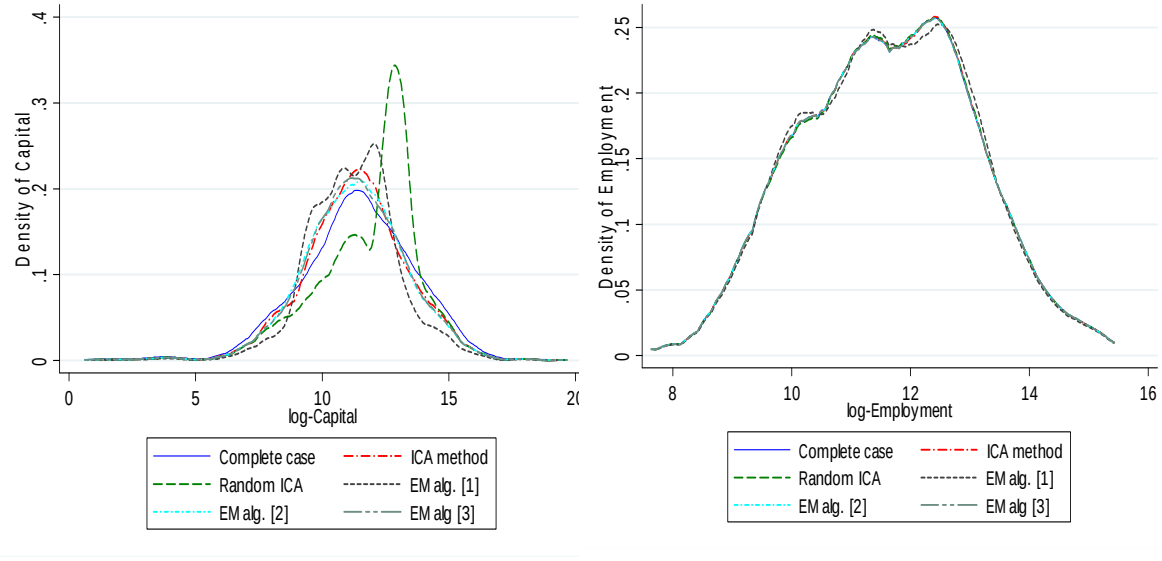
A. Sales (log)

B. Materials (log)



C. Capital stock (log)

D. Employment (log)

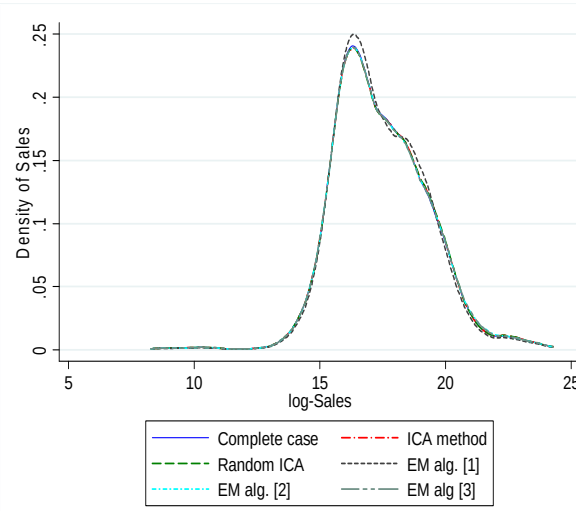


Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

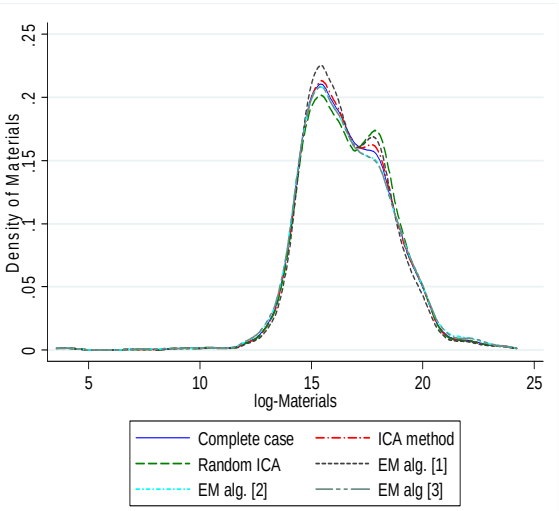
Figure 4.3: SOUTH AFRICA, comparison of the ICA method and other imputation mechanisms for PF variables

I. Kernel[1] estimates of output and inputs densities

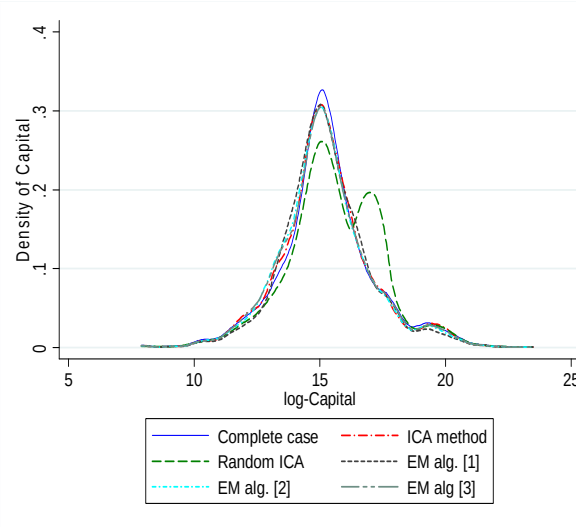
A. Sales (log)



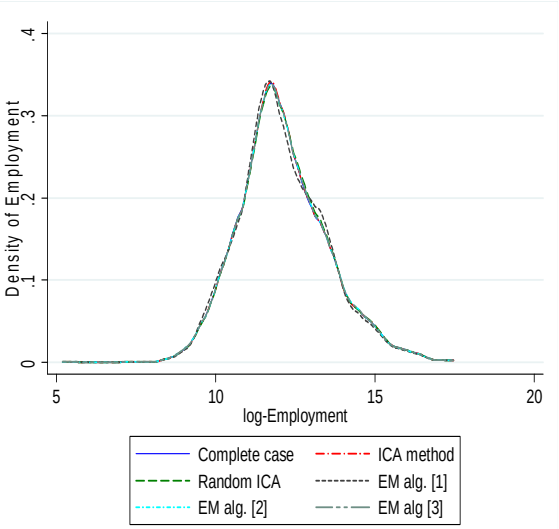
B. Materials (log)



C. Capital stock (log)



D. Employment (log)

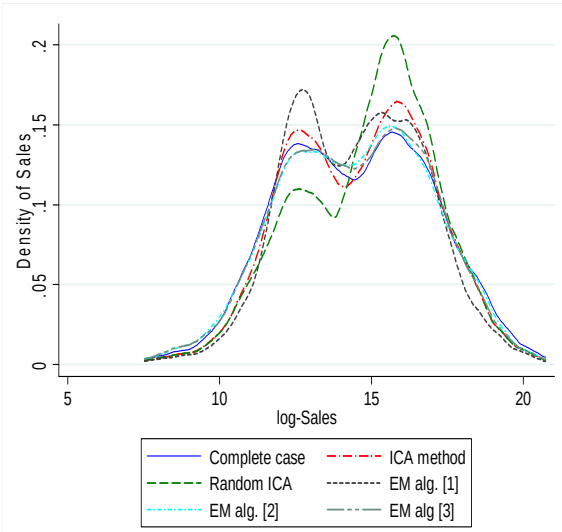


Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

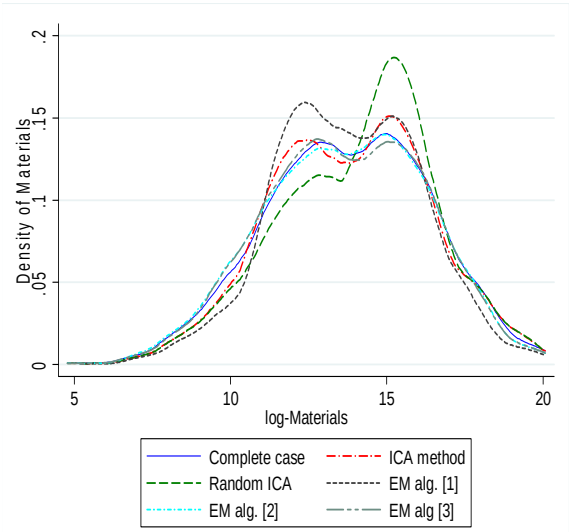
Figure 4.4: TANZANIA, comparison of the ICA method and other imputation mechanisms for PF variables

I. Kernel[1] estimates of output and input densities

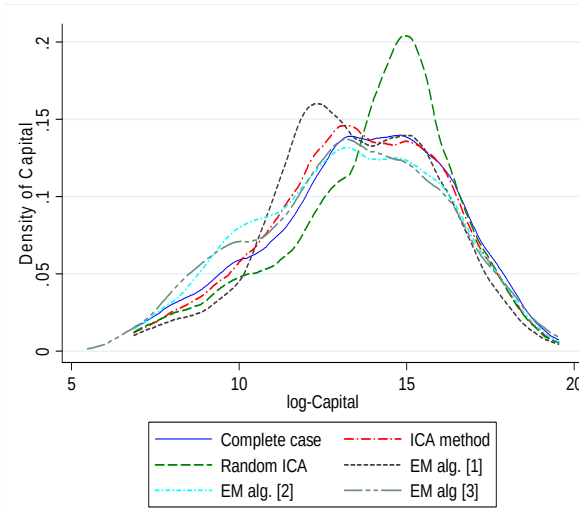
A. Sales (log)



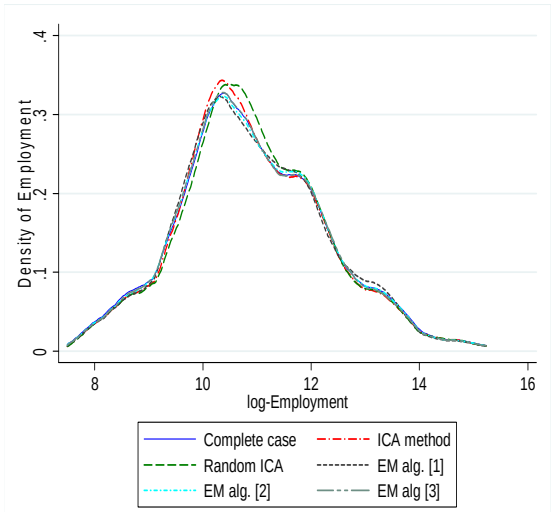
B. Materials (log)



C. Capital stock (log)



D. Employment (log)



Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
Source: Authors' estimations with ICSs data.

Table 8.1 INDIA, Descriptive statistics of production function variables under different imputation mechanism

	Variable	#Obs. (#imputed)	Mean	Std. Dev.	Min	Max
Sales	Complete case	5841.00	12.08	2.30	1.30	22.79
	ICA method	5935 (94)	12.07	2.29	1.30	22.79
	Random ICA meth.	5935 (94)	12.13	2.32	1.30	22.79
	EM alg. [1]	6848 (1007)	12.02	2.19	1.30	22.79
	EM alg. [2]	5882 (41)	12.08	2.30	1.30	22.79
	EM alg. [3]	5882 (41)	12.08	2.30	1.30	22.79
Materials	Complete case	5597.00	11.44	2.30	2.94	22.20
	ICA method	5933 (336)	11.40	2.28	2.94	22.20
	Random ICA meth.	5933 (336)	11.57	2.35	2.94	22.20
	EM alg. [1]	6848 (1251)	11.35	2.17	2.94	22.20
	EM alg. [2]	5906 (309)	11.42	2.32	2.94	22.20
	EM alg. [3]	5906 (336)	11.42	2.32	2.94	22.20
Capital	Complete case	4555.00	10.31	2.11	1.85	20.73
	ICA method	5918 (1363)	10.28	2.10	1.85	20.73
	Random ICA meth.	5918 (1363)	11.20	2.47	1.85	20.73
	EM alg. [1]	6848 (2293)	10.26	1.89	1.85	20.73
	EM alg. [2]	5807 (1252)	10.25	2.04	1.85	20.73
	EM alg. [3]	5807 (1252)	10.23	2.02	1.85	20.73
Employment	Complete case	6164.00	10.82	1.33	6.54	16.16
	ICA method	6321 (157)	10.82	1.34	6.54	16.16
	Random ICA meth.	6321 (157)	10.84	1.34	6.54	16.16
	EM alg. [1]	6849 (687)	10.78	1.31	6.54	16.16
	EM alg. [2]	6164 (0)	10.82	1.33	6.54	16.16
	EM alg. [3]	6164 (0)	10.82	1.33	6.54	16.16

Source: Authors' estimations with ICSs data.

Table 8.2 TURKEY, Descriptive statistics of production function variables under different imputation mechanism

	Variable	#Obs. (#imputed)	Mean	Std. Dev.	Min	Max
Sales	Complete case	1497	14.24	2.10	7.78	19.40
	ICA method	1821 (324)	14.30	1.99	7.78	19.40
	Random ICA meth.	1821 (324)	14.44	1.97	7.78	19.40
	EM alg. [1]	2646 (1149)	14.27	1.78	7.78	19.40
	EM alg. [2]	1808 (311)	14.22	2.02	7.55	19.40
	EM alg. [3]	1808 (311)	14.22	2.01	7.78	19.40
Materials	Complete case	1293	13.19	2.31	4.33	18.65
	ICA method	1822 (529)	13.37	2.13	4.34	18.65
	Random ICA meth.	1822 (529)	13.59	2.12	4.34	18.65
	EM alg. [1]	2646 (1353)	13.31	1.86	4.33	18.65
	EM alg. [2]	1802 (509)	13.18	2.18	4.33	18.65
	EM alg. [3]	1802 (509)	13.15	2.18	4.33	18.65
Capital	Complete case	1289	11.39	2.26	0.63	19.65
	ICA method	1816 (527)	11.32	2.05	1.05	19.65
	Random ICA meth.	1816 (527)	11.86	2.05	1.05	19.65
	EM alg. [1]	2646 (1357)	11.22	1.79	0.63	19.65
	EM alg. [2]	1807 (518)	11.28	2.05	0.63	19.65
	EM alg. [3]	1807 (518)	11.30	2.04	0.63	19.65
Employment	Complete case	2529	11.63	1.45	7.64	15.42
	ICA method	2548 (19)	11.63	1.45	7.64	15.42
	Random ICA meth.	2548 (19)	11.63	1.44	7.64	15.42
	EM alg. [1]	2646 (117)	11.63	1.44	7.64	15.42
	EM alg. [2]	2539 (10)	11.63	1.45	7.64	15.42
	EM alg. [3]	2539 (10)	11.63	1.45	7.64	15.42

Source: Authors' estimations with ICSs data.

Table 8.3 SOUTH AFRICA, Descriptive statistics of production function variables under different imputation mechanism

	Variable	#Obs. (#imputed)	Mean	Std. Dev.	Min	Max
Sales	Complete case	1578	17.43	1.86	8.28	24.29
	ICA method	1587 (9)	17.44	1.87	8.28	24.29
	Random ICA meth.	1587 (9)	17.44	1.87	8.28	24.29
	EM alg. [1]	1789 (211)	17.42	1.81	8.28	24.29
	EM alg. [2]	1587 (9)	17.44	1.87	8.28	24.29
	EM alg. [3]	1587 (9)	17.44	1.87	8.28	24.29
Materials	Complete case	1508	16.59	2.03	3.56	24.21
	ICA method	1587 (79)	16.60	2.00	3.56	24.21
	Random ICA meth.	1587 (79)	16.66	2.01	3.56	24.21
	EM alg. [1]	1789 (281)	16.58	1.93	3.56	24.21
	EM alg. [2]	1586 (78)	16.59	2.08	3.56	24.21
	EM alg. [3]	1586 (78)	16.59	2.08	3.56	24.21
Capital	Complete case	1337	15.29	1.89	7.90	23.48
	ICA method	1586 (249)	15.25	1.86	7.90	23.48
	Random ICA meth.	1586 (249)	15.60	1.90	7.90	23.48
	EM alg. [1]	1786 (449)	15.24	1.75	7.90	23.48
	EM alg. [2]	1583 (246)	15.20	1.84	7.90	23.48
	EM alg. [3]	1580 (243)	15.22	1.87	7.90	23.48
Employment	Complete case	1664	12.12	1.40	5.19	17.47
	ICA method	1685 (21)	12.12	1.40	5.19	17.47
	Random ICA meth.	1685 (21)	12.13	1.40	5.19	17.47
	EM alg. [1]	1784 (120)	12.10	1.40	5.19	17.47
	EM alg. [2]	1680 (16)	12.13	1.40	5.19	17.47
	EM alg. [3]	1680 (16)	12.13	1.40	5.19	17.47

The null hypothesis of the one-sample Kolmogorov-Smirnov Test is that the cumulative distribution differs from the hypothesized theoretical normal distribution.
Source: Authors' estimations with ICSs data.

Table 8.4 TANZANIA, Descriptive statistics of production function variables under different imputation mechanism

	Variable	#Obs. (#imputed)	Mean	Std. Dev.	Min	Max
Sales	Complete case	511	14.52	2.43	7.54	20.73
	ICA method	667 (156)	14.60	2.30	7.54	20.73
	Random ICA meth.	667 (156)	14.85	2.25	7.54	20.73
	EM alg. [1]	801 (290)	14.51	2.18	7.54	20.73
	EM alg. [2]	647 (136)	14.48	2.42	7.54	20.73
	EM alg. [3]	647 (136)	14.48	2.41	7.54	20.73
Materials	Complete case	539	13.76	2.58	4.78	20.07
	ICA method	667 (128)	13.82	2.52	4.78	20.07
	Random ICA meth.	667 (128)	14.08	2.49	4.78	20.07
	EM alg. [1]	803 (264)	13.74	2.32	4.78	20.07
	EM alg. [2]	646 (107)	13.67	2.58	4.78	20.07
	EM alg. [3]	646 (107)	13.67	2.57	4.78	20.07
Capital	Complete case	529	13.59	2.69	6.86	19.54
	ICA method	664 (135)	13.54	2.57	6.86	19.54
	Random ICA meth.	664 (135)	13.91	2.51	6.86	19.54
	EM alg. [1]	806 (277)	13.46	2.40	6.86	19.54
	EM alg. [2]	654 (125)	13.26	2.74	6.86	19.54
	EM alg. [3]	654 (125)	13.26	2.81	5.47	19.54
Employment	Complete case	730	10.92	1.37	7.50	15.23
	ICA method	788 (58)	10.91	1.34	7.50	15.23
	Random ICA meth.	788 (58)	10.94	1.34	7.50	15.23
	EM alg. [1]	790 (60)	10.92	1.36	7.50	15.23
	EM alg. [2]	758 (28)	10.92	1.36	7.50	15.23
	EM alg. [3]	768 (38)	10.92	1.36	7.50	15.23

Source: Authors' estimations with ICSs data.

We find that the proportion of missing values is an important factor in the observed underlying distributions after imputing missing values. Therefore, by means of explanation it is useful to discuss the results by groups of countries. The first group, with India and South Africa, comprises those countries with the largest response rate of PF variables, 65% in India and 70% in South Africa. The second group includes Tanzania and Turkey, whose response rates are only 40 and 30% respectively.

As shown in the kernel densities, the response rate dramatically determines the shape of the densities after imputing missing values. In India (see Figure 4.1), where the response rate is reasonably high in all the variables except capital, all the methods lead to estimated densities similar to those of the complete case. However, in the case of capital where the response rate is considerably lower, we observe a dramatic change in the distribution of the imputed values by the Random ICA method. Concretely, the distribution appears to have two modes, moving a considerable proportion of density from the center of the distribution to the right. This misleading behavior is already indicated in the case of materials, although to a lesser extent.

Regarding the estimated distributions of the remaining imputation mechanism, all of them lead to results similar to those of the complete case, including the ICA method and EM

algorithms. Nonetheless, in terms of descriptive statistics, it is noticeable that, in spite of the uncertainty inherent in the EM algorithm [1], it slightly reduces the estimated standard deviation of all PF variables, even with respect to the ICA method case. This is probably due to the higher number of imputed cells than under other mechanisms. Nonetheless, it must also be pointed out that the reduction of the standard deviation is only of the order of one decimal point. In this sense, the Random ICA method, and the remaining EM algorithms increase, to some extent, the estimated standard errors with respect to the ICA method.

The case of South Africa is virtually symmetrical to that of India. Again the Random ICA method performs badly in the case of capital. Likewise, due to the larger proportion of missing values imputed, the EM algorithm [1] leads to estimated standard errors that slightly reduce those of the complete case.

As the response rate of PF variables decreases, the estimated densities obtained from the EM algorithms and Random ICA method tend to be different from those of the complete case and the standard ICA method, especially in the case of the Random ICA method. This is illustrated in the cases of Turkey and Tanzania in figures 4.2 and 4.4. Nonetheless, the estimated descriptive statistics are quite homogeneous among imputation methods, as shown in tables 8.2 and 8.4. The estimated means are virtually equal in all the cases, and the standard errors show great consistency across specifications, except in the EM algorithm [1] where, again due to the larger proportion of values imputed, the standard errors are slightly lower.

It is useful to recapitulate the main conclusions of this subsection before introducing the results of estimating equation (1). Overall, there are small differences in the imputation of PF variables. Nonetheless, these differences become more marked as the number of missing values increases and when the variables are far from being normally distributed.

6.2 Evaluation of imputation mechanism: Comparison of estimating results of equation (1)

6.2.1 Comparison of the *ICA method* and other EM algorithms

Tables 9.1, 9.2, 9.3 and 9.4 show the results of estimating equation (5) after imputing missing values by the ICA method and by the three EM algorithms proposed in section 5.1. A key conclusion is that when the proportion of missing values is not large enough there are no remarkable differences between applying the ICA method or the EM algorithm [1], neither in the point estimates of the input-output (I-O) elasticities, nor in the standard errors (recall that uncertainty is a key issue under EM algorithms). Another interesting observation is that we do not gain much by extending the EM algorithm to include the IC variables among the information set.

Table 9.1: INDIA, Extended production function and comparison of ICA method with EM algorithms

Dependent variable: Log of total sales		ICA Method[1]		EM Algorithms[2]		
				[1]	[2]	[3]
Category	**Variable**	**Coeff. std. err.**	**Boot. s.e**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.1027 [0.0341]***	(0.0306)***	0.0976 [0.0331]***	0.0516 [0.0250]**	0.0527 [0.0250]**
	Log-materials	0.7989 [0.0185]***	(0.0462)***	0.8362 [0.0186]***	0.8607 [0.0176]***	0.8628 [0.0177]***
	Log-capital	0.0676 [0.0239]***	(0.0153)***	0.0629 [0.0225]***	0.0537 [0.0146]***	0.0502 [0.0147]***
Infrastructure	Longest # of days to clear customs for exports (a)	-0.0125 [0.0263]	(0.0376)	-0.0039 [0.0275]	-0.0158 [0.0209]	-0.0156 [0.0208]
	Dummy for own generator	0.0538 [0.0422]	(0.0424)	0.0378 [0.0396]	0.015 [0.0247]	0.0131 [0.0249]
	Water supply from public sources (b)	0.0014 [0.0005]***	(0.0008)*	0.0013 [0.0004]***	0.0009 [0.0003]***	0.0008 [0.0003]**
	Shipment losses in the domestic market (b)	-0.0047 [0.0039]	(0.0128)	-0.0023 [0.0035]	-0.0017 [0.0030]	-0.0016 [0.0030]
	Dummy for own transport	0.0238 [0.0475]	(0.0861)	-0.0084 [0.0464]	-0.003 [0.0340]	-0.0023 [0.0341]
	Dummy for web page	0.0402 [0.0394]	(0.0264)	0.0047 [0.0378]	0.0013 [0.0310]	0.0008 [0.0313]
	Dummy for security	0.0467 [0.0423]	(0.1407)	0.0426 [0.0403]	0.0497 [0.0285]*	0.0505 [0.0285]*
Red tape, corruption and crime	Sales reported for taxes (b)	0.0006 [0.0014]	(0.0052)	0.0009 [0.0013]	0.0008 [0.0010]	0.0009 [0.0010]
	Workforce reported for taxes (b)	-0.0015 [0.0012]	(0.0042)	-0.0015 [0.0010]	-0.0009 [0.0008]	-0.0009 [0.0008]
	Dummy for payments to speed up bureaucracy	-0.0464 [0.0336]	(0.0526)	-0.0443 [0.0292]	0.0041 [0.0255]	0.0083 [0.0259]
	Dummy for interventionist labor regulation	-0.036 [0.0361]	(0.0211)*	-0.0317 [0.0340]	-0.0259 [0.0330]	-0.028 [0.0331]
	Absenteeism (b)	-0.0299 [0.0222]	(0.0571)	-0.0204 [0.0195]	-0.0069 [0.0156]	-0.0071 [0.0160]
Finance and corporate governance	Dummy for trade association	0.0785 [0.0455]*	(0.0456)*	0.0756 [0.0408]*	0.024 [0.0297]	0.0194 [0.0300]
	Working capital financed by domestic private banks (b)	0.0002 [0.0007]	(0.0005)	-0.0002 [0.0007]	0.0003 [0.0006]	0.0003 [0.0006]
	Dummy for external audit	0.0691 [0.0395]*	(0.0452)	0.0662 [0.0362]*	0.0633 [0.0283]**	0.0655 [0.0282]**
	Dummy for loan (b)	0.1102 [0.0473]**	(0.0637)*	0.0892 [0.0464]*	0.0121 [0.0331]	0.006 [0.0327]
Quality, innovation and labor skills	Dummy for R&D (a)	0.1787 [0.2382]	(0.2347)	0.2041 [0.2534]	0.0702 [0.1322]	0.0638 [0.1320]
	Dummy for product innovation	-0.0073 [0.0360]	(0.0710)	-0.0153 [0.0332]	-0.025 [0.0244]	-0.0265 [0.0246]
	Dummy for foreign license (b)	0.204 [0.1053]*	(0.1302)	0.1425 [0.1033]	0.086 [0.0847]	0.0801 [0.0852]
	Dummy for internal training (b)	0.0579 [0.0533]	(0.0516)	0.0578 [0.0511]	0.0702 [0.0443]	0.0703 [0.0442]
	Unskilled workforce (a)	0.0013 [0.0036]	(0.0016)	0.0013 [0.0036]	-0.0034 [0.0030]	-0.0039 [0.0031]
	Workforce with computer	0.0017 [0.0011]	(0.0015)	0.0016 [0.0010]	0.0012 [0.0009]	0.0011 [0.0008]
Other control variables	Dummy for incorporated company	0.0265 [0.0396]	(0.0901)	0.0162 [0.0368]	0.0272 [0.0301]	0.0261 [0.0300]
	Age	0.0534 [0.0267]**	(0.0214)**	0.0438 [0.0251]*	0.0456 [0.0174]**	0.0487 [0.0174]***
	Share of exports (b)	0.001 [0.0009]	(0.0005)**	0.0006 [0.0009]	0.00004 [0.0006]	-0.0001 [0.0006]
	Trade union (b)	0.0008 [0.0012]	(0.0010)	0.0008 [0.0012]	0.0009 [0.0009]	0.0007 [0.0009]
	Strikes (b)	-0.0683 [0.0449]	(0.0821)	-0.0475 [0.0380]	-0.0112 [0.0307]	-0.0107 [0.0314]
	Constant	0.7377 [0.3449]**		0.4456 [0.3504]	1.0108 [0.2499]***	1.0335 [0.2492]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes
	Observations	5211		5216	5175	5176
	R-squared	0.88		0.9	0.94	0.94

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions.

[1] ICA method is in section 3 of main text. Significance is given by clustered and White-robust standard errors in brackets; *** 1%, **5%, * 10%. In parentheses are bootstrap standard errors after 1000 replications (see section 5.2.2 on the motivation of using bootstrap standard errors). Correlation by clusters is also considered.

[2] EM algorithms are explained in section 5.1. EM alg [1] includes as covariates of the imputation mechanism industry/region/size/time (I/R/S/T) dummies (see section 5.1.1); EM alg [2] includes I/R/S/T dummies and production function variables (see section 5.1.2); EM alg [3] also includes a set of IC variables (see section 5.1.3). Significance is given by clustered White-robust standard errors. (a) IC variables instrumented with industry/region average variables. (b) missing values in IC variables replaced by means of *ICA method*.

Source: Authors' calculations with ICSs

Table 9.2: TURKEY, Extended production function and comparison of ICA method with EM algorithms

	Dependent variable: Log of total sales	ICA Method[1]		EM Algorithms[2]		
				[1]	[2]	[3]
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.416 [0.0492]***	(0.1088)***	0.3743 [0.0434]***	0.3421 [0.0459]***	0.3323 [0.0467]***
	Log-materials	0.4184 [0.0404]***	(0.0249)***	0.4829 [0.0429]***	0.6075 [0.0369]***	0.6052 [0.0370]***
	Log-capital	0.0371 [0.0165]**	(0.0428)	0.0548 [0.0199]***	0.0801 [0.0190]***	0.0783 [0.0184]***
Infrastructures	Days to clear customs for imports (a)	-0.0707 [0.0686]	(0.0688)	-0.1497 [0.0578]**	-0.1206 [0.0516]**	-0.1399 [0.0462]***
	Dummy for e-mail	0.2866 [0.0920]***	(0.1365)**	0.1659 [0.0789]**	0.1648 [0.0726]**	0.188 [0.0720]**
Red tape, corruption and crime	Security expenses (b)	-0.0246 [0.0828]	(0.0011)***	-0.0117 [0.0520]	-0.0504 [0.0456]	-0.0647 [0.0416]
	Payments to deal with bureaucratic issues (a)	-0.011 [0.0020]***	(0.0077)	-0.0092 [0.0013]***	-0.0065 [0.0011]***	-0.0072 [0.0012]***
	Sales declared for taxes (a)	-0.0226 [0.0057]***	(0.0045)***	-0.0234 [0.0046]***	-0.0148 [0.0042]***	-0.0177 [0.0040]***
	Number of inspections (b)	0.0046 [0.0044]	(0.0597)	-0.0002 [0.0026]	0.0001 [0.0026]	0.0007 [0.0023]
	Payments to obtain a contract with the government (b)	-0.0373 [0.0315]	(0.0058)***	-0.0524 [0.0217]**	-0.0274 [0.0175]	-0.0514 [0.0159]***
	Production lost due to absenteeism (b)	-0.0054 [0.0043]	(0.0367)	-0.0122 [0.0037]***	-0.0082 [0.0028]***	-0.0094 [0.0029]***
	Dummy for informal competition (b)	0.0044 [0.0295]	(0.1203)	-0.0055 [0.0236]	-0.0013 [0.0189]	-0.0059 [0.0196]
	Delay in obtaining a water supply (a)	-0.1325 [0.0634]**	(0.0993)	-0.1388 [0.0565]**	-0.0746 [0.0559]	-0.0935 [0.0600]
Finance	Dummy for credit line	0.068 [0.0868]	(0.1383)	0.1157 [0.0702]	0.0744 [0.0660]	0.0778 [0.0674]
	Dummy for external auditory (a)	0.0863 [0.0753]	(0.1117)	0.0655 [0.0461]	0.0627 [0.0397]	0.0935 [0.0406]**
	Loans in foreign currency (b)	0.0018 [0.0009]**	(0.0010)*	0.0013 [0.0005]**	0.0008 [0.0006]	0.0007 [0.0006]
Quality, innov. and labor skills	Staff with university education (b)	0.0095 [0.0026]***	(0.0018)***	0.0087 [0.0029]***	0.0064 [0.0029]**	0.0081 [0.0029]***
	Staff-part time workers	-0.008 [0.0030]**	(0.0222)	-0.0046 [0.0023]*	-0.0059 [0.0016]***	-0.0058 [0.0018]***
Other control variables	Production lost due to strikes (b)	-0.1689 [0.0634]**	(0.0351)***	-0.1596 [0.0435]***	-0.124 [0.0322]***	-0.1072 [0.0323]***
	Dummy for recently privatized firm	1.0606 [0.2812]***	(0.2511)***	0.8692 [0.2579]***	0.6825 [0.2478]***	0.6644 [0.2508]**
	Dummy for competition against imported products	0.2069 [0.0962]**	(0.2737)	0.1595 [0.0736]**	0.0951 [0.0603]	0.0755 [0.0607]
	Constant	3.5299 [0.7190]***		3.6661 [0.5851]***	1.6872 [0.3782]***	1.9648 [0.3791]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes
	Observations	1684		1679	1733	1733
	R-squared	0.73		0.81	0.86	0.86

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions.

[1, 2] See footnotes in Table 9.1.

(a) IC variables instrumented with industry/region average variables. (b) missing values in IC variables replaced by means of *ICA method*.

Source: Authors' calculations with ICSs.

Table 9.3: SOUTH AFRICA, Extended production function and comparison of ICA method with EM algorithms

	Dependent variable: Log of total sales	ICA Method[1]		EM Algorithms[2]		
				[1]	[2]	[3]
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.3226 [0.0711]***	(0.0365)***	0.3144 [0.0676]***	0.2285 [0.0667]***	0.2261 [0.0666]***
	Log-materials	0.5195 [0.1017]***	(0.0214)***	0.5355 [0.0942]***	0.5781 [0.0947]***	0.574 [0.0943]***
	Log-capital	0.1247 [0.0300]***	(0.0118)***	0.1287 [0.0370]***	0.123 [0.0373]***	0.1282 [0.0386]***
Infrastructure	Days to clear customs for imports (a)	-0.1188 [0.1125]	(0.1233)	0.1291 [0.1320]	0.0193 [0.0935]	0.0322 [0.0975]
	Sales lost due to power outages (b)	-0.0171 [0.0114]	(0.0047)***	-0.0128 [0.0101]	-0.0112 [0.0077]	-0.0096 [0.0073]
	Water outages (b)	-0.1477 [0.0527]***	(0.0942)	-0.1287 [0.0438]***	-0.1482 [0.0533]***	-0.1611 [0.0562]***
	Average duration of transport failures (a)	-0.0439 [0.0806]	(0.0379)	0.06 [0.0893]	0.0021 [0.0628]	-0.0156 [0.0611]
	Wait for electric supply (a)	-0.0867 [0.0553]	(0.0173)***	-0.1368 [0.0337]***	-0.0921 [0.0272]***	-0.0863 [0.0258]***
	Sales lost due to delivery delays (b)	-0.0099 [0.0083]	(0.0073)	-0.0148 [0.0084]*	-0.0097 [0.0072]	-0.0077 [0.0065]
Red tape, corruption and crime	Manager's time spent on bur. issues (b)	0.007 [0.0051]	(0.0016)***	0.0077 [0.0050]	0.0077 [0.0057]	0.0084 [0.0058]
	Payments to deal with bureaucratic issues (b)	-0.0045 [0.0024]*	(0.3604)	-0.005 [0.0026]*	-0.0042 [0.0023]*	-0.0121 [0.0038]***
	Sales declared for taxes (a)	0.0056 [0.0046]	(0.0022)**	0.0056 [0.0042]	0.0059 [0.0025]**	0.0058 [0.0027]**
	Payments to obtain a contract with the government (b)	-0.0144 [0.0185]	(0.1975)	-0.0119 [0.0175]	-0.0161 [0.0146]	-0.015 [0.0144]
	Security expenses (a)	0.1407 [0.0511]**	(0.0069)***	-0.0023 [0.0148]	-0.0075 [0.0109]	-0.0056 [0.0113]
	Illegal payments in protection (b)	0.3969 [0.2428]	(0.1128)***	0.3754 [0.2492]	0.3882 [0.2202]*	0.4761 [0.2187]**
	Crime losses (a)	-0.0502 [0.0788]	(0.1374)	-0.0541 [0.0948]	0.0099 [0.0621]	0.0193 [0.0662]
Finance and corporate governance	Percentage of credit unused (b)	0.0014 [0.0010]	(0.0013)	0.0016 [0.0008]*	0.0019 [0.0010]*	0.002 [0.0010]*
	Dummy for loan	0.0715 [0.0492]	(0.0327)**	0.0841 [0.0479]*	0.0762 [0.0406]*	0.0761 [0.0407]*
	Value of the collateral (b)	-0.0008 [0.0002]***	(0.0009)	-0.0007 [0.0002]***	-0.0006 [0.0002]***	-0.0007 [0.0002]***
	Loans in foreign currency (b)	0.0018 [0.0022]	(0.0024)	0.0007 [0.0018]	-0.0002 [0.0012]	0.0001 [0.0012]
	Charge to clear a check (a)	-0.1164 [0.0503]**	(0.0253)***	-0.0861 [0.0520]	-0.0995 [0.0387]**	-0.1068 [0.0384]***
	Largest shareholder	0.0006 [0.0010]	(0.0008)	0.0011 [0.0010]	0.001 [0.0007]	0.0011 [0.0007]
	Working capital financed by foreign commercial banks (b)	0.0106 [0.0083]	(0.0084)	0.0072 [0.0072]	0.0057 [0.0060]	0.0044 [0.0063]
	Working capital financed by informal sources (b)	-0.0022 [0.0023]	(0.0001)***	-0.0018 [0.0021]	-0.0027 [0.0018]	-0.0026 [0.0018]
Quality, innovation and labor skills	Dummy for ISO quality certification (b)	0.1603 [0.0766]**	(0.0365)***	0.1521 [0.0732]**	0.0838 [0.0404]**	0.0782 [0.0390]*
	Dummy for new product (b)	0.091 [0.0494]*	(0.0113)***	0.1083 [0.0530]**	0.1053 [0.0461]**	0.1001 [0.0460]**
	Dummy for discontinued product line (b)	-0.1007 [0.0610]	(0.0384)**	-0.1029 [0.0560]*	-0.0874 [0.0541]	-0.0805 [0.0534]
	Staff - management	0.004 [0.0028]	(0.0009)***	0.0036 [0.0028]	0.0034 [0.0030]	0.0032 [0.0030]
	Staff - non-production workers	-0.0034 [0.0022]	(0.0025)	-0.0032 [0.0022]	-0.0023 [0.0022]	-0.0024 [0.0022]
	Training for unskilled workers (a)	0.001 [0.0026]	(0.0030)	-0.0001 [0.0038]	0.0018 [0.0019]	0.0008 [0.0021]
	University staff (b)	0.0049 [0.0015]***	(0.0007)***	0.0055 [0.0016]***	0.0039 [0.0014]***	0.0038 [0.0014]**
	Manager's experience (b)	0.0391 [0.0249]	(0.0217)*	0.0369 [0.0222]	0.028 [0.0187]	0.0271 [0.0184]
Other control variables	Age (b)	0.0018 [0.0015]	(0.0016)	0.0014 [0.0013]	0.0023 [0.0013]*	0.0023 [0.0013]*
	Share of the local market (b)	0.0032 [0.0008]***	(0.0004)***	0.0035 [0.0009]***	0.0027 [0.0007]***	0.0028 [0.0007]***
	Constant	2.7174 [0.8932]***	(0.0365)***	2.0109 [0.8200]**	2.5368 [0.7330]***	2.5977 [0.7464]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes
	Observations	1483		1528	1552	1550
	R-squared	0.89		0.89	0.9	0.9

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions. [1, 2,] See footnotes in Table 9.1.
Source: Authors' calculations with ICSs.

Table 9.4: TANZANIA, Extended production function and comparison of ICA method with EM algorithms

	Dependent variable: Log of total sales	ICA Method[1]		EM Algorithms[2]		
				[1]	[2]	[3]
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.1655 [0.0853]*	(0.0512]***	0.1142 [0.0919]	0.0584 [0.0459]	0.0207 [0.0501]
	Log-materials	0.4252 [0.0581]***	(0.0340]***	0.4867 [0.0677]***	0.7201 [0.0435]***	0.724 [0.0401]***
	Log-capital	0.1589 [0.0323]***	(0.0208]***	0.1628 [0.0317]***	0.1326 [0.0286]***	0.1171 [0.0288]***
Infrastructure	Electricity from own generator (b)	0.0021 [0.0016]	(0.0053]	0.0035 [0.0016]**	0.0036 [0.0011]***	0.0027 [0.0011]**
	Losses due to water outages (b)	-0.0112 [0.0058]*	(0.0162]	-0.0172 [0.0049]***	-0.0087 [0.0033]**	-0.0082 [0.0034]**
	Water from own well or water infrastructure (a)	0.0001 [0.0051]	(0.0011]	-0.0044 [0.0044]	0.0013 [0.0031]	0.0029 [0.0035]
	Losses due to phone outages (a)	-0.0322 [0.0198]	(0.0071]***	-0.0066 [0.0268]	0.0115 [0.0232]	0.0159 [0.0260]
	Transport outages (a)	-0.0047 [0.0703]	(0.1168]	0.0366 [0.0623]	0.0069 [0.0295]	-0.0287 [0.0280]
	Dummy for own roads (b)	0.289 [0.1488]*	(0.0581]***	0.3789 [0.1279]***	0.2864 [0.1176]**	0.4766 [0.1189]***
	Dummy for webpage (b)	0.1578 [0.1212]	(0.1994]	0.0972 [0.1533]	0.1054 [0.1346]	0.2051 [0.1243]
	Wait for a water supply (a)	-0.1814 [0.0427]***	(0.0702]**	-0.1354 [0.0533]**	-0.093 [0.0262]***	-0.1649 [0.0235]***
	Low quality supplies (a)	-0.0163 [0.0128]	(0.0041]***	-0.0351 [0.0141]**	-0.0165 [0.0105]	-0.0202 [0.0112]*
Red tape, corruption and crime	Gift to obtain an operating license (b)	-0.4983 [0.1935]**	(0.1066]***	-0.3964 [0.1550]**	0.0537 [0.1051]	-0.0553 [0.0983]
	Payments to deal with bureaucratic issues (b)	0.0939 [0.0299]***	(0.0164]***	0.0808 [0.0272]***	0.0512 [0.0503]	0.085 [0.0396]**
	Days in inspections (b)	-0.1045 [0.0735]	(0.0494]**	-0.0735 [0.0703]	0.0027 [0.0379]	0.0005 [0.0362]
	Payments to obtain a contract with the government (b)	-0.0114 [0.0066]*	(0.0091]	-0.0026 [0.0059]	-0.0082 [0.0040]*	-0.0079 [0.0044]*
	Security expenses (b)	-0.0119 [0.0042]***	(0.0092]	-0.0081 [0.0040]*	-0.0023 [0.0031]	-0.0005 [0.0035]
	Illegal payments in protection (b)	-0.0827 [0.0170]***	(0.1019]	-0.0518 [0.0140]***	-0.031 [0.0144]**	-0.0489 [0.0206]**
Finance and corporate governance	Interest rate of the loan (a)	-0.0109 [0.0145]	(0.0099]	-0.0139 [0.0127]	0.0033 [0.0073]	0.0117 [0.0078]
	Working capital financed by commercial banks (b)	-0.0009 [0.0018]	(0.0012]	-0.0015 [0.0016]	-0.0016 [0.0011]	-0.001 [0.0011]
	Working capital financed by leasing (b)	-0.0794 [0.0282]***	(0.0054]***	-0.118 [0.0279]***	-0.015 [0.0038]***	-0.0893 [0.0428]**
	Sales bought on credit (b)	-0.0014 [0.0012]	(0.0011]	0 [0.0011]	0.0006 [0.0010]	0.0006 [0.0010]
	Delay in clearing a domestic currency wire (a)	-0.3418 [0.3273]	(0.0935]***	-0.0439 [0.2600]	0.1691 [0.1544]	0.0498 [0.1606]
Quality, innovation and labor skills	Dummy for new product (b)	0.0429 [0.1063]	(0.2036]	-0.0053 [0.1090]	-0.0481 [0.0782]	-0.0897 [0.0632]
	Staff - skilled workers (b)	0.0026 [0.0023]	(0.0050]	0.0025 [0.0021]	0.0036 [0.0014]**	0.0038 [0.0014]**
	Workforce with computer (b)	0.0066 [0.0030]**	(0.0056]	0.0071 [0.0034]**	0.0001 [0.0049]	0.003 [0.0041]
Other control variables	Dummy for incorporated company (b)	0.2914 [0.2023]	(0.5683]	-0.0777 [0.4506]	0.1645 [0.1868]	0.0871 [0.2324]
	Dummy for FDI (b)	0.1397 [0.1445]	(0.2844]	0.0825 [0.1397]	-0.0662 [0.0792]	-0.0859 [0.0768]
	Constant	7.2978 [1.0168]***		6.3827 [0.8512]***	2.4414 [0.5932]***	3.296 [0.6161]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes
	Observations	559		560	603	597
	R-squared	0.88		0.88	0.94	0.94

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions.

[1, 2,] See footnotes in Table 9.1.

Source: Authors' calculations with ICSs.

Table 9.1 focuses on the case of India, in which the ICA method and the EM algorithm on industry, region and size variables (EM algorithm [1]) lead to similar results in terms of input-output elasticities. However, there are divergences in the input-output elasticities estimated for the remaining two EM-algorithms. Concretely, the employment coefficient decreases from 0.1 in the ICA method and EM algorithm [1], to 0.05 in the EM algorithms [2] and [3]. Similarly, it is worth mentioning that the estimates of the standard errors of the coefficients of the input-output elasticities do not improve in the EM algorithm [1] with respect to the ICA method, and are even lower in the EM algorithms [2] and [3].

It is important to note that most of the differences between the ICA method and the EM algorithm [1] on the one hand and the EM algorithms [2] and [3] on the other can be explained by the greater amount of information embodied in the imputation process: production function variables in the EM [2] and production function, IC, and D variables in EM [3]; and not by the iterative process based on posterior predictive densities as in the EM algorithms. When the pattern of missing data is very unbalanced and we are able to observe only one or two PF variables for each cross-sectional observation, those EM algorithms including additional variables, beyond the region/industry/size dummies, are more likely to lead to heterogeneous results as they include a different amount of information for each cross-section. This becomes more patent in the case of the EM algorithm [3], in which we also include IC variables in the imputation.

Apart from this observation, the elasticities and semi-elasticities of IC variables show a reasonable robustness to the imputation mechanism used. In general terms, the ICA method is more consistent with the results from the EM algorithm [1], whereas EM algorithms [2] and [3] show more differences. For example, out of 6 IC variables significant in the ICA method case, 5 are also significant in the EM algorithm [1], while only 3 in the EM algorithms [2] and [3] (see Table 12). Nonetheless, the changes observed are only in the magnitude of the coefficients estimated, and never in the direction of the effects. All the estimated IC coefficients move within a reasonable range of values in the four cases.

Table 12: Summary of results from estimating equation (1) under different imputation methods with respect to the ICA method case

			Complete case	ICA method & variations				EM algorithms			Multiple imputation	Heckman models		
				ICA met.	ICA met. (boot. s. e.)	Random ICA met.	ICA met. on inputs	EM alg. [1]	EM alg. [2]	EM alg. [3]		Heckman on complete case	Heckman replacing inputs	Heckman (boot. s.e)
India: Tables 9.1, 10.1 & 11.1	Input-output elasticities	Significant change in estimated elasticity?[3]	No	-	-	No	No	No	Yes (L, M)	Yes (L, M)	Yes (L)	No	No	-
		Change in significance?[3]	No	-	No	No	No	No	No	No	No	No	No	No
	IC variables [27 vars.]	Significant variables[1]	4, (3)	**6**	6, (2)	6, (3)	4, (1)	5, (0)	4, (1)	4, (1)	5, (2)	11, (7)	11, (6)	15, (10)
		Non-significant variables[2]	23, (5)	**21**	21, (2)	21, (4)	23, (3)	22, (1)	23, (3)	23, (3)	22, (2)	16, (2)	16, (1)	12, (1)
		Change in the direction of the effect?[3]	No	-	-	No	No	No	No	No	No	No	No	No
	Number of observations		3943	**5211**	-	5063	5134	5216	5175	5176	5262	4233	5407	-
	Significant Heckman's Lambda?		-	-	-	-	-	-	-	-	-	No	No	-
Turkey: Table 9.2, 10.2 & 11.2	Input-output elasticities	Significant change in estimated elasticity?[3]	Yes (M)	-	-	No	Yes (L, M, K)	Yes (M, L)	Yes (L, M, K)	Yes (L, M, K)	Yes (L, M)	Yes (L. K)	Yes (L. K)	-
		Change in significance?[3]	No	-	Yes (L)	Yes (L)	No	No	No	No	No	No	No	No
	IC variables [18 vars.]	Significant variables[1]	9, (3)	**10**	8, (2)	9, (0)	9, (0)	13, (3)	9, (2)	11, (4)	10, (2)	9, (3)	11, (1)	16, (6)
		Non-significant variables[2]	9, (4)	**8**	10, (4)	9, (1)	9, (1)	5, (0)	9, (2)	7, (2)	8, (2)	9, (2)	7, (0)	2, (0)
		Change in the direction of the effect?[3]	No	-	-	No	No	No	No	No	No	No	No	No
	Number of observations		792	**1684**	-	1684	1360	1679	1733	1733	1646	1941	2509	-
	Significant Heckman's Lambda?		-	-	-	-	-	-	-	-	-	No	No	-
South Africa: Table 9.3, 10.3 & 11.3	Input-output elasticities	Significant change in estimated elasticity?[3]	No	-	-	Yes (K)	No	No	Yes (L)	Yes (L)	Yes (L)	No	No	-
		Change in significance?[3]	No	-	No	No	No	No	No	No	No	No	No	No
	IC variables [31 vars.]	Significant variables[1]	10, (3)	**9**	16, (10)	12, (3)	9, (0)	12, (5)	14, (6)	14, (5)	15, (7)	15, (8)	19, (11)	18, (10)
		Non-significant variables[2]	21, (2)	**22**	15, (3)	19, (0)	22, (0)	19, (2)	17, (1)	17, (1)	16, (1)	16, (2)	12, (1)	13, (1)
		Change in the direction of the effect?[3]	No	-	-	No	No	No	No	No	No	No	No	-
	Number of observations			**1483**				1528	1552	1550		1443	1657	-
	Significant Heckman's Lambda?		-	-	-	-	-	-	-	-	-	No	No	-
Tanzania: Table 9.4, 10.4 & 11.4	Input-output elasticities	Significant change in estimated elasticity?[3]	Yes (M)	-	-	Yes (L, K)	Yes (L, M)	Yes (L)	Yes (M, L)	Yes (M, L)	Yes (L, M, K)	Yes (M)	Yes (M)	-
		Change in significance?[3]	No	-	No	No	No	Yes (L)	Yes (L)	Yes (L)	No	No	No	No
	IC variables [25 vars.]	Significant variables[1]	10, (4)	**10**	9, (4)	11, (4)	10, (2)	11, (2)	8, (2)	10, (3)	8, (2)	14, (9)	9, (5)	7, (5)
		Non-significant variables[2]	15, (3)	**15**	16, (5)	14, (3)	15, (2)	14, (1)	17, (4)	15, (3)	17, (4)	11, (5)	16, (6)	18, (6)
		Change in the direction of the effect?[3]	No	-	-	No	No				No	No	No	-
	Number of observations		291	**559**	-	557	442	560	603	597	570	581	771	-
	Significant Heckman's Lambda?		-	-	-	-	-	-	-	-	-	No	No	-

[1] In parenthesis: variables non-significant in the ICA method that became significant under other imputation mechanisms.
[2] In parenthesis: variables significant in the ICA method and no longer significant under other imputation mechanisms.
[3] With respect to the ICA method.
A more detailed description of the results is in Tables 8.1 to 8.4.
Source: Authors' calculations with ICSs data.

The case of South Africa in Table 9.3, with a pattern of missing values similar to that of India, leads to analogous conclusions. Again the I-O elasticities estimated under the ICA method are rather similar to those we get under the EM algorithm [1], whereas the EM algorithms [2] and [3] diverge in the sense that the estimated I-O elasticity for employment is almost one percent point lower than in the ICA method and EM algorithm [1]. The patterns observed for the standard errors estimated are the same as those of India: almost equal standard errors between the ICA method and the rest of EM algorithms, so no improvements of efficiency can be observed from using the EM algorithms in this case. Concretely, from Table 12 there are 10 significant IC variables under the ICA method, and the same variables are significant again under the EM algorithm [1] (plus another three new significant IC variables). In the EM algorithms [2] and [3] only 7 IC variables out of 10 repeat significance.

The patterns observed in India and South Africa are not supported by the Turkish case in Table 9.2. Recall that the proportion of missing values among PF variables reaches 70%, and therefore the effects of the imputation mechanism used will be quite different from those applied to patterns of missing data with only a 20% or 30% response rate. In this case, it is remarkable that I-O elasticities in the EM algorithms [1], [2] and [3] are closer to constant returns to scale (CRS) than the ICA method is. In this sense, and in terms of I-O elasticities, the results from the ICA method are different from the EM algorithms, with materials and capital elasticities significantly lower than in the remaining cases. However, the estimated standard errors do not change much and the significance of the PF variables is not modified in any of the cases. In spite of these changes in the I-O elasticities, it is important to note that again the IC parameters appear to be robust to the imputation method used. Ten IC variables turned out to be significant in the ICA method case, 12 in the EM algorithm [1] and 14 in the EM algorithms [2] and [3]. Apart from minor changes in the magnitude of the coefficients, and in some cases in the significance of some variables, we do not observe changes in the estimated directions of the effects of the IC variables.

Finally, the case of Tanzania is presented in Table 9.4. The proportion of missing values in PF variables in this country is more than 70% of the original sampling frame, similar to that of Turkey. However, unlike the Turkish case, EM algorithms [2] and [3] do not improve the results obtained from the ICA method. Again, the ICA method and EM algorithm show symmetrical behavior with similar I-O elasticities, whereas in EM algorithms [2] and [3] the estimated elasticity for employment is three times lower than in the ICA method, increasing in turn the elasticity of materials. On the other hand, almost all of those IC variables significant in the ICA method repeat significance in the EM algorithms, and what is more important, the coefficients are robust to all the imputation mechanisms, apart from marginal differences in some variables (see Table 12).

6.2.2 Comparison of the ICA method with complete case, extensions of the ICA method and multiple imputation

In this section, we compare the results obtained from the ICA method with those from the complete case, other extensions of the ICA method (see section 5.2) and multiple imputation

(see section 5.3) in tables 10.1 to 10.4. Table 10.1 focuses on the case of India. The fourth column comprises the results of the complete case, for which the number of observations is considerably reduced with respect to the ICA method case, from 5211 to 3943. In spite of the reduced number of observations used, there are not significant changes either in the estimated I-O elasticities, or in their level of significance. Referring to the IC parameters, it is worth mentioning that, although there are no changes in the directions of the estimated effects, and the coefficients are rather robust in both specifications, some of the variables lost their significance in the complete case, with respect to the ICA method. Thus, out of the 6 significant IC variables in the ICA method, only 1 is also significant in the complete case.

Table 10.1: INDIA, Extended production function and comparison of ICA method with extensions

Dependent variable: log of total sales		ICA method and extensions				Complete case[4]	Multiple imputation (Switching regr.)[5]
		Original ICA meth.[1]		Random ICA m.[2]	ICA m. on inputs[3]		
Category	**Variable**	**Coeff. std. err.**	**Boot. s.e**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.1027 [0.0341]***	(0.0306)***	0.1051 [0.0346]***	0.0922 [0.0343]***	0.1168 [0.0317]***	0.0659 [0.0245]***
	Log-materials	0.7989 [0.0185]***	(0.0462)***	0.8135 [0.0186]***	0.8054 [0.0192]***	0.7994 [0.0236]***	0.8560 [0.0169]***
	Log-capital	0.0676 [0.0239]***	(0.0153)***	0.0438 [0.0143]***	0.0722 [0.0248]***	0.0504 [0.0170]***	0.0452 [0.0128]***
Infrastructure	Longest # of days to clear customs for export (a)	-0.0125 [0.0263]	(0.0376)	-0.01 [0.0317]	-0.0167 [0.0266]	-0.0432 [0.0268]	-0.0155 [0.0213]
	Dummy for own generator	0.0538 [0.0422]	(0.0424)	-0.0083 [0.0453]	0.0516 [0.0431]	0.0424 [0.0293]	0.0198 [0.0254]
	Water supply from public sources (b)	0.0014 [0.0005]***	(0.0008)*	0.0009 [0.0006]	0.0014 [0.0005]***	0.0013 [0.0004]***	0.0008 [0.0003]**
	Shipment losses in the domestic market (b)	-0.0047 [0.0039]	(0.0128)	-0.0075 [0.0034]**	-0.0037 [0.0038]	-0.0023 [0.0054]	-0.0020 [0.0029]
	Dummy for own transport	0.0238 [0.0475]	(0.0861)	0.0013 [0.0459]	0.0334 [0.0482]	0.0465 [0.0369]	-0.0038 [0.0347]
	Dummy for web page	0.0402 [0.0394]	(0.0264)	0.0516 [0.0427]	0.0329 [0.0382]	0.0098 [0.0327]	0.0067 [0.0316]
	Dummy for security	0.0467 [0.0423]	(0.1407)	0.045 [0.0392]	0.0573 [0.0429]	0.0564 [0.0293]*	0.0582 [0.0293]**
Red tape, corruption and crime	Sales reported for taxes (b)	0.0006 [0.0014]	(0.0052)	0.002 [0.0012]*	0.0009 [0.0014]	0.0002 [0.0010]	0.0010 [0.0009]
	Workforce reported for taxes (b)	-0.0015 [0.0012]	(0.0042)	-0.0021 [0.0009]**	-0.0014 [0.0012]	0.0005 [0.0008]	-0.0010 [0.0007]
	Dummy for payments to speed up bureaucracy	-0.0464 [0.0336]	(0.0526)	-0.0148 [0.0265]	-0.0416 [0.0335]	0.0072 [0.0247]	0.0004 [0.0254]
	Dummy for interventionist labor regulation	-0.036 [0.0361]	(0.0211)*	-0.0372 [0.0369]	-0.0275 [0.0368]	-0.031 [0.0330]	-0.0303 [0.0322]
	Absenteeism (b)	-0.0299 [0.0222]	(0.0571)	-0.0233 [0.0256]	-0.0263 [0.0216]	-0.0011 [0.0193]	-0.0108 [0.0158]
Finance and corporate governance	Dummy for trade association	0.0785 [0.0455]*	(0.0456)*	0.094 [0.0480]*	0.0734 [0.0454]	0.022 [0.0388]	0.0263 [0.0302]
	Working capital financed by domestic private banks (b)	0.0002 [0.0007]	(0.0005)	0.0005 [0.0006]	0.0002 [0.0008]	0.0003 [0.0008]	0.0002 [0.0005]
	Dummy for external audit	0.0691 [0.0395]*	(0.0452)	0.0541 [0.0440]	0.0627 [0.0386]	0.0392 [0.0300]	0.0689 [0.0294]**
	Dummy for loan (b)	0.1102 [0.0473]**	(0.0637)*	0.0851 [0.0538]	0.1107 [0.0492]**	-0.0397 [0.0409]	0.0188 [0.0337]
Quality, innovation and labor skills	Dummy for R&D (a)	0.1787 [0.2382]	(0.2347)	0.0959 [0.1637]	0.1885 [0.2400]	0.0862 [0.1313]	0.1143 [0.1353]
	Dummy for product innovation	-0.0073 [0.0360]	(0.0710)	-0.0331 [0.0392]	-0.0079 [0.0366]	-0.0528 [0.0262]**	-0.0285 [0.0276]
	Dummy for foreign license (b)	0.204 [0.1053]*	(0.1302)	0.2384 [0.1181]**	0.1555 [0.1013]	0.1401 [0.0939]	0.1032 [0.0835]
	Dummy for internal training (b)	0.0579 [0.0533]	(0.0516)	0.0744 [0.0649]	0.0631 [0.0537]	0.0884 [0.0458]	0.0717 [0.0440]*
	Unskilled workforce (a)	0.0013 [0.0036]	(0.0016)	0.0038 [0.0042]	0.0003 [0.0037]	-0.001 [0.0033]	-0.0030 [0.0029]
	Workforce with computer	0.0017 [0.0011]	(0.0015)	0.0014 [0.0009]	0.0019 [0.0011]*	0.0007 [0.0007]	0.0012 [0.0008]
Other control variables	Dummy for incorporated company	0.0265 [0.0396]	(0.0901)	0.056 [0.0358]	0.0127 [0.0423]	0.0494 [0.0282]*	0.0280 [0.0311]
	Age	0.0534 [0.0267]**	(0.0214)**	0.0352 [0.0287]	0.0525 [0.0271]*	0.0322 [0.0208]	0.0392 [0.0182]**
	Share of exports (b)	0.001 [0.0009]	(0.0005)**	0.001 [0.0010]	0.001 [0.0009]	0.0002 [0.0005]	-0.0001 [0.0005]
	Trade union (b)	0.0008 [0.0012]	(0.0010)	0.0015 [0.0013]	0.001 [0.0013]	0.0001 [0.0008]	0.0007 [0.0008]
	Strikes (b)	-0.0683 [0.0449]	(0.0821)	-0.0557 [0.0470]	-0.0707 [0.0457]	0.0248 [0.0439]	-0.0112 [0.0321]
	Constant	0.7377 [0.3449]**		0.7174 [0.3636]*	0.7182 [0.3455]**	1.0943 [0.2692]***	0.9976 [0.2528]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes	Yes
	Observations	5211		5063	5134	3943	5262
	R-squared	0.88		0.88	0.88	0.94	-

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions.

[1] See footnote 1 in Table 9.1. [2] Random ICA method is described in section 5.2.1. [3] ICA method on inputs is in section 5.2.3. [4] Complete case considers missingness in PF variables only, not in IC variables. [5] Multiple imputation via switching regression can be found in section 5.3.

In all the cases significance is given by clustered and White- robust standard errors in brackets; *** 1%, **5%, * 10%. In the case of the ICA method, in parentheses are bootstrap standard errors after 1000 replications (see section 5.2.2 on the motivation for using bootstrap standard errors). Correlation by cluster is also considered.

Source: Authors' calculations with ICSs.

Table 10.2: TURKEY, Extended production function and comparison of ICA method with extensions

Dependent variable: log of total sales		ICA method and extensions				Complete case[4]	Multiple imputation (Switching regr.)[5]
		Original ICA meth.[1]		Random ICA m.[2]	ICA m. on inputs[3]		
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.416 [0.0492]***	(0.1088)***	0.3819 [0.0501]***	0.5106 [0.0558]***	0.4002 [0.0885]***	0.3446 [0.0524]***
	Log-materials	0.4184 [0.0404]***	(0.0249)***	0.4137 [0.0392]***	0.4615 [0.0484]***	0.5332 [0.0494]***	0.5779 [0.0316]***
	Log-capital	0.0371 [0.0165]**	(0.0428)	0.0193 [0.0198]	0.0686 [0.0232]***	0.0639 [0.0271]**	0.0603 [0.0246]**
Infrastructures	Days to clear customs for imports (a)	-0.0707 [0.0686]	(0.0688)	-0.1133 [0.0776]	-0.0711 [0.0705]	-0.1594 [0.0856]*	-0.1318 [0.0660]**
	Dummy for e-mail	0.2866 [0.0920]***	(0.1365)**	0.3833 [0.1048]***	0.3072 [0.1054]***	0.0317 [0.1295]	0.1729 [0.0754]**
Red tape, corruption and crime	Security expenses (b)	-0.0246 [0.0828]	(0.0011)***	0.0137 [0.0836]	-0.0861 [0.0919]	-0.0468 [0.0786]	-0.0215 [0.0587]
	Payments to deal with bureaucratic issues (a)	-0.011 [0.0020]***	(0.0077)	-0.0108 [0.0021]***	-0.0102 [0.0021]***	-0.0084 [0.0014]***	-0.0073 [0.0011]***
	Sales declared for taxes (a)	-0.0226 [0.0057]***	(0.0045)***	-0.0197 [0.0061]***	-0.0151 [0.0065]**	-0.0184 [0.0082]**	-0.0159 [0.0051]***
	Number of inspections (b)	0.0046 [0.0044]	(0.0597)	0.001 [0.0049]	0.005 [0.0044]	-0.0019 [0.0038]	0.0000 [0.0036]
	Payments to obtain a contract with the government (b)	-0.0373 [0.0315]	(0.0058)***	-0.0345 [0.0357]	-0.0217 [0.0368]	-0.0257 [0.0360]	-0.0354 [0.0236]
	Production lost due to absenteeism (b)	-0.0054 [0.0043]	(0.0367)	-0.0079 [0.0051]	-0.005 [0.0039]	-0.0107 [0.0054]*	-0.0110 [0.0036]***
	Dummy for informal competition (b)	0.0044 [0.0295]	(0.1203)	-0.0083 [0.0323]	0.0207 [0.0279]	-0.0015 [0.0315]	-0.0062 [0.0232]
	Delay in obtaining a water supply (a)	-0.1325 [0.0634]**	(0.0993)	-0.1346 [0.0688]*	-0.1419 [0.0863]	-0.0825 [0.0785]	-0.0965 [0.0571]*
Finance	Dummy for credit line	0.068 [0.0868]	(0.1383)	0.0967 [0.0905]	0.0888 [0.1061]	0.0657 [0.0685]	0.0699 [0.0719]
	Dummy for external auditory (a)	0.0863 [0.0753]	(0.1117)	0.0992 [0.0739]	0.1012 [0.0791]	0.1385 [0.0709]*	0.0781 [0.0521]
	Loans in foreign currency (b)	0.0018 [0.0009]**	(0.0010)*	0.0015 [0.0008]*	0.0018 [0.0010]*	0.0005 [0.0009]	0.0009 [0.0008]
Quality, innov. and labor skills	Staff with university education (b)	0.0095 [0.0026]***	(0.0018)***	0.0107 [0.0028]***	0.01 [0.0040]**	0.008 [0.0035]**	0.0060 [0.0032]*
	Staff-part time workers	-0.008 [0.0030]**	(0.0222)	-0.0077 [0.0032]**	-0.0102 [0.0029]***	-0.0069 [0.0027]**	-0.0067 [0.0019]***
Other control variables	Production lost due to strikes (b)	-0.1689 [0.0634]**	(0.0351)***	-0.1063 [0.0650]	-0.1538 [0.0671]**	-0.1765 [0.0521]***	-0.1092 [0.0564]*
	Dummy for recently privatized firm	1.0606 [0.2812]***	(0.2511)***	1.0239 [0.2791]***	1.0215 [0.3100]***	1.2627 [0.3162]***	0.8012 [0.2884]***
	Dummy for competition against imported products	0.2069 [0.0962]**	(0.2737)	0.2013 [0.0962]**	0.2096 [0.1041]*	0.0156 [0.0823]	0.1021 [0.0665]
	Constant	3.5299 [0.7190]***		4.6379 [0.7023]***	1.4306 [0.5738]**	2.6911 [0.7730]***	2.6126 [0.4577]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes	Yes
	Observations	1684		1684	1360	792	1646
	R-squared	0.73		0.68	0.75	0.85	-

Notes of Table 10.1
Source: Authors' calculations with ICSs.

Table 10.3: SOUTH AFRICA, Extended production function and comparison of ICA method with extensions

Dependent variable: log of total sales		ICA method and extensions				Complete case[4]	Multiple imputation (Switching regr.)[5]
		Original ICA meth.[1]		Random ICA meth.[2]	ICA met. on inputs[3]		
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.3226 [0.0711]***	(0.0365)***	0.3822 [0.0776]***	0.3295 [0.0717]***	0.3428 [0.0541]***	0.2453 [0.0681]***
	Log-materials	0.5195 [0.1017]***	(0.0214)***	0.4914 [0.0877]***	0.5182 [0.1015]***	0.4877 [0.0961]***	0.5674 [0.0905]***
	Log-capital	0.1247 [0.0300]***	(0.0118)***	0.0791 [0.0264]***	0.124 [0.0302]***	0.1118 [0.0322]***	0.1180 [0.0345]***
Infrastructure	Days to clear customs for imports (a)	-0.1188 [0.1125]	(0.1233)	-0.14 [0.1247]	-0.1407 [0.1176]	0.018 [0.1976]	0.0423 [0.1008]
	Sales lost due to power outages (b)	-0.0171 [0.0114]	(0.0047)***	-0.0194 [0.0127]	-0.0142 [0.0104]	-0.003 [0.0085]	-0.0107 [0.0080]
	Water outages (b)	-0.1477 [0.0527]***	(0.0942)	-0.1441 [0.0591]**	-0.1405 [0.0513]**	-0.1427 [0.0659]**	-0.1393 [0.0504]***
	Average duration of transport failures (a)	-0.0439 [0.0806]	(0.0379)	-0.0065 [0.0867]	-0.074 [0.0832]	0.1229 [0.1507]	-0.0022 [0.0762]
	Wait for electric supply (a)	-0.0867 [0.0553]	(0.0173)***	-0.1075 [0.0589]*	-0.0767 [0.0573]	-0.0629 [0.0558]	-0.1014 [0.0309]***
	Sales lost due to delivery delays (b)	-0.0099 [0.0083]	(0.0073)	-0.0111 [0.0092]	-0.0119 [0.0080]	-0.0074 [0.0081]	-0.0089 [0.0072]
Red tape, corruption and crime	Manager's time spent on bur. issues (b)	0.007 [0.0051]	(0.0016)***	0.0072 [0.0051]	0.0073 [0.0052]	0.0058 [0.0043]	0.0079 [0.0056]
	Payments to deal with bureaucratic issues (b)	-0.0045 [0.0024]*	(0.3604)	-0.0063 [0.0031]*	-0.0045 [0.0023]*	-0.0008 [0.0125]	-0.0044 [0.0024]*
	Sales declared for taxes (a)	0.0056 [0.0046]	(0.0022)**	0.0015 [0.0049]	0.0064 [0.0044]	0.0091 [0.0039]**	0.0058 [0.0031]*
	Payments to obtain a contract with the government (b)	-0.0144 [0.0185]	(0.1975)	-0.0218 [0.0201]	-0.017 [0.0208]	-0.0129 [0.0112]	-0.0180 [0.0162]
	Security expenses (a)	0.1407 [0.0511]**	(0.0069)***	0.1245 [0.0586]**	0.1159 [0.0477]**	0.0227 [0.0146]	-0.0075 [0.0123]
	Illegal payments for protection (b)	0.3969 [0.2428]	(0.1128)***	0.4048 [0.2751]	0.3997 [0.2428]	0.3265 [0.3225]	0.3606 [0.2254]*
	Crime losses (a)	-0.0502 [0.0788]	(0.1374)	0.0153 [0.0855]	-0.0679 [0.0786]	0.1115 [0.0871]	-0.0121 [0.0708]
Finance and corporate governance	Percentage of credit unused (b)	0.0014 [0.0010]	(0.0013)	0.0014 [0.0010]	0.0015 [0.0010]	0.0007 [0.0006]	0.0018 [0.0010]*
	Dummy for loan	0.0715 [0.0492]	(0.0327)**	0.0678 [0.0547]	0.072 [0.0493]	0.0602 [0.0421]	0.0814 [0.0443]*
	Value of the collateral (b)	-0.0008 [0.0002]***	(0.0009)	-0.0008 [0.0002]***	-0.0008 [0.0002]***	-0.0009 [0.0002]***	-0.0007 [0.0002]***
	Loans in foreign currency (b)	0.0018 [0.0022]	(0.0024)	0.0024 [0.0023]	0.0016 [0.0021]	0.0012 [0.0011]	-0.0001 [0.0012]
	Charge to clear a check (a)	-0.1164 [0.0503]**	(0.0253)***	-0.1404 [0.0570]**	-0.1108 [0.0501]**	-0.1722 [0.0582]***	-0.0905 [0.0402]**
	Largest shareholder	0.0006 [0.0010]	(0.0008)	-0.0003 [0.0010]	0.0008 [0.0009]	0.0001 [0.0009]	0.0010 [0.0008]
	Working capital fin. by foreign commercial banks (b)	0.0106 [0.0083]	(0.0084)	0.0073 [0.0090]	0.0107 [0.0082]	0.0203 [0.0195]	0.0050 [0.0062]
	Working capital financed by informal sources (b)	-0.0022 [0.0023]	(0.0001)***	-0.0032 [0.0023]	-0.0021 [0.0023]	-0.0046 [0.0011]***	-0.0025 [0.0019]
Quality, innovation and labor skills	Dummy for ISO quality certification (b)	0.1603 [0.0766]**	(0.0365)***	0.1956 [0.0646]***	0.1578 [0.0764]**	0.121 [0.0670]*	0.1029 [0.0454]**
	Dummy for new product (b)	0.091 [0.0494]*	(0.0113)***	0.1233 [0.0587]**	0.0926 [0.0496]*	0.0461 [0.0393]	0.0948 [0.0475]**
	Dummy for discontinued product line (b)	-0.1007 [0.0610]	(0.0384)**	-0.1334 [0.0648]**	-0.099 [0.0597]	-0.0616 [0.0353]*	-0.0864 [0.0527]*
	Staff - management	0.004 [0.0028]	(0.0009)***	0.0049 [0.0027]*	0.0038 [0.0027]	0.0041 [0.0030]	0.0034 [0.0030]
	Staff - non-production workers	-0.0034 [0.0022]	(0.0025)	-0.0033 [0.0021]	-0.0033 [0.0022]	-0.0026 [0.0021]	-0.0024 [0.0021]
	Training for unskilled workers (a)	0.001 [0.0026]	(0.0030)	0.0023 [0.0028]	0 [0.0025]	-0.0047 [0.0045]	0.0011 [0.0027]
	University staff (b)	0.0049 [0.0015]***	(0.0007)***	0.0051 [0.0015]***	0.0049 [0.0014]***	0.0044 [0.0011]***	0.0043 [0.0014]***
	Manager's experience (b)	0.0391 [0.0249]	(0.0217)*	0.0412 [0.0271]	0.0387 [0.0249]	0.0325 [0.0254]	0.0292 [0.0196]
Other control variables	Age (b)	0.0018 [0.0015]	(0.0016)	0.0019 [0.0014]	0.0017 [0.0014]	0.0023 [0.0013]*	0.0021 [0.0013]*
	Share of the local market (b)	0.0032 [0.0008]***	(0.0004)***	0.0023 [0.0009]**	0.0032 [0.0008]***	0.0027 [0.0009]***	0.0029 [0.0007]***
	Constant	2.7174 [0.8932]***	(0.0365)***	3.5878 [0.8355]***	2.6721 [0.8751]***	2.6313 [0.9880]**	2.6249 [0.7400]***
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes	Yes
	Observations	1483		1483	1474	1236	1483
	R-squared	0.89		0.87	0.89	0.91	

Notes for Table 10.1
Source: Authors' calculations with ICSs.

Table 10.4: TANZANIA, Extended production function and comparison of ICA method with extensions

Dependent variable: log of total sales		ICA method and extensions				Complete case[4]	Multiple imputation (Swithching regression)[5]
		Original ICA meth.[1]		Random ICA met.[2]	ICA met. on inputs[3]		
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**	**Coeff. std. err.**
PF variables	Log-employment	0.1655 [0.0853]*	(0.0512]***	0.2643 [0.1039]**	0.2339 [0.0603]***	0.1651 [0.0681]**	0.1217 (0.0625]**
	Log-materials	0.4252 [0.0581]***	(0.0340]***	0.4008 [0.0527]***	0.6087 [0.0406]***	0.6242 [0.0468]***	0.7170 (0.0390]***
	Log-capital	0.1589 [0.0323]***	(0.0208]***	0.0975 [0.0418]**	0.1302 [0.0280]***	0.1311 [0.0312]***	0.0977 (0.0294]***
Infrastructure	Electricity from own generator (b)	0.0021 [0.0016]	(0.0053]	0.0013 [0.0017]	0.0019 [0.0016]	-0.0002 [0.0022]	0.0039 (0.0016]**
	Losses due to water outages (b)	-0.0112 [0.0058]*	(0.0162]	-0.0132 [0.0081]	-0.0058 [0.0051]	-0.0107 [0.0062]*	-0.0094 (0.0046]**
	Water from own well or water infrastructure (a)	0.0001 [0.0051]	(0.0011]	-0.0094 [0.0060]	-0.0017 [0.0046]	0.0004 [0.0056]	-0.0003 (0.0038]
	Losses due to phone outages (a)	-0.0322 [0.0198]	(0.0071]***	-0.0453 [0.0237]*	0.0003 [0.0208]	0.0089 [0.0209]	0.0078 (0.0238]
	Transport outages (a)	-0.0047 [0.0703]	(0.1168]	0.0785 [0.0940]	0.0243 [0.0573]	-0.0859 [0.0567]	0.0054 (0.0322]
	Dummy for own roads (b)	0.289 [0.1488]*	(0.0581]***	0.1502 [0.1582]	0.4010 [0.1164]***	0.4073 [0.1249]***	0.3117 (0.1422]**
	Dummy for webpage (b)	0.1578 [0.1212]	(0.1994]	0.1453 [0.1280]	0.2560 [0.1038]**	0.3106 [0.1170]**	0.0977 (0.1635]
	Wait for a water supply (a)	-0.1814 [0.0427]***	(0.0702]**	-0.1769 [0.0531]***	-0.1388 [0.0411]***	-0.1252 [0.0326]***	-0.1036 (0.0356]***
	Low quality supplies (a)	-0.0163 [0.0128]	(0.0041]***	-0.0389 [0.0164]**	-0.0210 [0.0127]	-0.0285 [0.0142]*	-0.0183 (0.0120]
Red tape, corruption and crime	Gift to obtain an operating license (b)	-0.4983 [0.1935]**	(0.1066]***	-0.4607 [0.2385]*	-0.3262 [0.1439]**	-0.1671 [0.1562]	0.0694 (0.1218]
	Payments to deal with bureaucratic issues (b)	0.0939 [0.0299]***	(0.0164]***	0.0376 [0.0578]	0.1182 [0.0295]***	0.0767 [0.0192]***	0.0546 (0.0472]
	Days in inspections (b)	-0.1045 [0.0735]	(0.0494]**	-0.1172 [0.0984]	-0.0514 [0.0425]	-0.0524 [0.0643]	-0.0009 (0.0461]
	Payments to obtain a contract with the government (b)	-0.0114 [0.0066]*	(0.0091]	-0.0177 [0.0086]**	-0.0189 [0.0066]***	-0.0254 [0.0078]***	-0.0140 (0.0051]***
	Security expenses (b)	-0.0119 [0.0042]***	(0.0092]	-0.0151 [0.0055]**	-0.0072 [0.0034]**	0.008 [0.0193]	-0.0042 (0.0032]
	Illegal payments for protection (b)	-0.0827 [0.0170]***	(0.1019]	-0.081 [0.0329]**	-0.0774 [0.0179]***	-0.0603 [0.0251]**	-0.0392 (0.0131]***
Finance and corporate governance	Interest rate of the loan (a)	-0.0109 [0.0145]	(0.0099]	-0.0028 [0.0182]	-0.0038 [0.0094]	0.0111 [0.0113]	-0.0021 (0.0090]
	Working capital financed by commercial banks (b)	-0.0009 [0.0018]	(0.0012]	-0.0008 [0.0021]	-0.0013 [0.0014]	0.0007 [0.0013]	-0.0014 (0.0012]
	Working capital financed by leasing (b)	-0.0794 [0.0282]***	(0.0054]***	-0.1362 [0.0450]***	-0.0489 [0.0305]	-0.0304 [0.0329]	-0.0129 (0.0069]*
	Sales bought on credit (b)	-0.0014 [0.0012]	(0.0011]	-0.0036 [0.0017]**	-0.0003 [0.0011]	-0.0021 [0.0014]	-0.0002 (0.0014]
	Delay in clearing a domestic currency wire (a)	-0.3418 [0.3273]	(0.0935]***	-0.0024 [0.3738]	0.1242 [0.2583]	0.3236 [0.2952]	0.2044 (0.1717]
Quality, innovation and labor skills	Dummy for new product (b)	0.0429 [0.1063]	(0.2036]	0.1217 [0.1118]	-0.0526 [0.0945]	-0.1533 [0.1066]	-0.1045 (0.0981]
	Staff - skilled workers (b)	0.0026 [0.0023]	(0.0050]	0.0053 [0.0028]*	0.0038 [0.0022]*	0.0054 [0.0021]**	0.0039 (0.0020]*
	Workforce with computer (b)	0.0066 [0.0030]**	(0.0056]	0.0079 [0.0038]**	0.0094 [0.0039]**	0.0154 [0.0055]***	0.0037 (0.0045]
Other control variables	Dummy for incorporated company (b)	0.2914 [0.2023]	(0.5683]	0.238 [0.2648]	0.2327 [0.1841]	-0.2476 [0.1896]	0.2544 (0.2270]
	Dummy for FDI (b)	0.1397 [0.1445]	(0.2844]	0.3044 [0.1888]	0.1788 [0.1225]	0.1061 [0.1123]	-0.0255 (0.1128]
	Constant	7.2978 [1.0168]***		7.2545 [1.3295]***	2.7433 [0.8631]***	3.1164 [0.8674]***	2.4194 [0.7159]
	Industry/region/size/time dummies	Yes		Yes	Yes	Yes	Yes
	Observations	559		557	442	291	570
	R-squared	0.88		0.81	0.9300	0.95	

Notes forTable 10.1

Source: Authors' calculations with ICSs.

Especially interesting is the comparison of the ICA method with the Random ICA method—introduced in section 5.2.1—in which we introduce a random component to the imputation procedure in order to test the role played by the uncertainty inherent in the imputation mechanism. In a similar vein, another interesting point is to check the sensitivity of the significance level of the variables using bootstrap standard errors to correct for the problem of generated regressors (see section 5.2.2). Only 2 IC variables lose their significance in the ICA method with bootstrap standard error with respect to the regular case, and 2 new variables became significant. A similar pattern is observed in the Random ICA method with 6 significant IC variables, of which 3 were also significant in the ICA method (Table 12 includes the summary of significant IC variables in each case).

Finally, the ICA method on inputs and the multiple imputation cases lead to similar results in the I-O elasticities, with the exception of a slight decline in the capital elasticity. In both cases, the significance of some IC variables is lost, although the direction of the estimated effects never changes.

Similar conclusions can be drawn in the case of South Africa, the results of which are presented in Table 10.3. In this case, the number of observations used in the complete case only differs by 250 with respect to the ICA method. As expected from the larger response rate of PF variables in this country, there is no significant efficiency lost in the complete case and most IC variables remain significant. As in the case of India,, the Random ICA method and the bootstrap standard errors change the significance of some variables, and while some variables lose their significance, a small group of other IC variables become significant. Finally, both the ICA method on inputs and multiple imputation show robust results with respect to the ICA method. We only observe changes in the second or third decimals.

The cases of Turkey and Tanzania (tables 10.2 and 10.4 respectively) are rather different from the two previous ones. In both cases, using the complete case implies using less than 50% of the sample under the complete case. This implies a clear efficiency loss, which is translated into four less significant IC variables in the complete case in Turkey and three in Tanzania. By means of significance of IC variables, the results from the Random ICA, Bootstrap ICA method and ICA on inputs cases are more consistent with those from the standard ICA method. In this respect, introducing more uncertainty into the imputation procedure used in Turkey does not change the significance of 6 and 9 IC variables, depending on whether we focus on the Bootstrap ICA or on the Random ICA respectively. In Tanzania the patterns are similar: 4 IC variables lose their significance in both the Bootstrap ICA and the Random ICA. Lastly, in both cases, Turkey and Tanzania, the ICA method on inputs and the multiple imputation do not modify the results of the ICA method.

On the other hand, regarding I-O elasticities and in the case of Turkey, it is important to note that, although we only observe changes in the I-O estimate for materials, the I-O elasticity of employment is non-significant under the ICA method with bootstrap standard errors and the Random ICA method.

6.2.3 Comparison of the ICA method and the Heckman selection model

We now focus on the comparison of the ICA method and the Heckman models proposed in section 5.4 and 5.5. The estimating results are in tables 11.1 to 11.4. The main conclusions are summarized in Table 12.

Table 11.1: INDIA, Extended production function and comparison of ICA method with Heckman models

Dependent variable: log of total sales		**ICA Method[1]**		**Heckman models[2]**		
				Heckman on comp case	**Heckman replacing inputs**	
Category	**Variable**	**Coeff. std. err.**	**Boot. s.e**	**Coeff. std. err.**	**Coeff. std. err.**	**Boot. s.e**
PF variables	Log-employment	0.1027 [0.0341]***	(0.0306)***	0.1127 [0.0160]***	0.0806 [0.0184]***	(0.0452)***
	Log-materials	0.7989 [0.0185]***	(0.0462)***	0.7998 [0.0069]***	0.8121 [0.0070]***	(0.0567)***
	Log-capital	0.0676 [0.0239]***	(0.0153)***	0.0477 [0.0062]***	0.0578 [0.0070]***	(0.0168)***
Infrastructure	Longest # of days to clear customs for exports (a)	-0.0125 [0.0263]	(0.0376)	-0.0451 [0.0155]***	-0.0077 [0.0150]	(0.1542)
	Dummy for own generator	0.0538 [0.0422]	(0.0424)	0.0466 [0.0229]**	0.0769 [0.0265]***	(0.0064)***
	Water supply from public sources (b)	0.0014 [0.0005]***	(0.0008)*	0.0014 [0.0003]***	0.0012 [0.0003]***	(0.0460)***
	Shipment losses in the domestic market (b)	-0.0047 [0.0039]	(0.0128)	-0.0029 [0.0033]	-0.0022 [0.0029]	(0.1197)
	Dummy for own transport	0.0238 [0.0475]	(0.0861)	0.0438 [0.0283]	-0.0063 [0.0336]	(0.0742)
	Dummy for web page	0.0402 [0.0394]	(0.0264)	0.0061 [0.0221]	0.0212 [0.0263]	(0.0051)**
	Dummy for security	0.0467 [0.0423]	(0.1407)	0.0487 [0.0200]**	0.018 [0.0240]	(0.0035)**
Red tape, corruption and crime	Sales reported to taxes (b)	0.0006 [0.0014]	(0.0052)	-0.0001 [0.0007]	0.0011 [0.0008]	(0.0073)
	Workforce reported for taxes (b)	-0.0015 [0.0012]	(0.0042)	0.0005 [0.0007]	-0.001 [0.0007]	(0.0049)
	Dummy for payments to speed up bureaucracy	-0.0464 [0.0336]	(0.0526)	0.0079 [0.0186]	-0.0259 [0.0226]	(0.0463)
	Dummy for interventionist labor regulation	-0.036 [0.0361]	(0.0211)*	-0.0407 [0.0226]*	-0.0334 [0.0272]	(0.0658)**
	Absenteeism (b)	-0.0299 [0.0222]	(0.0571)	0.0003 [0.0112]	-0.0147 [0.0129]	(0.1783)**
Finance and corporate governance	Dummy for trade association	0.0785 [0.0455]*	(0.0456)*	0.0339 [0.0241]	0.0143 [0.0274]	(0.0762)
	Working capital financed by domestic private banks (b)	0.0002 [0.0007]	(0.0005)	0.0004 [0.0004]	0.001 [0.0004]**	(0.0006)**
	Dummy for external audit	0.0691 [0.0395]*	(0.0452)	0.0419 [0.0204]**	0.0827 [0.0245]***	(0.0408)*
	Dummy for loan (b)	0.1102 [0.0473]**	(0.0637)*	-0.0395 [0.0301]	0.1181 [0.0340]***	(0.0002)***
Quality, innovation and labor skills	Dummy for R&D (a)	0.1787 [0.2382]	(0.2347)	0.0813 [0.0933]	0.2063 [0.1112]*	(0.0010)
	Dummy for product innovation	-0.0073 [0.0360]	(0.0710)	-0.0508 [0.0200]**	-0.0081 [0.0233]	(0.0352)***
	Dummy for foreign license (b)	0.204 [0.1053]*	(0.1302)	0.141 [0.0434]***	0.1478 [0.0499]***	(0.0006)
	Dummy for internal training (b)	0.0579 [0.0533]	(0.0516)	0.0794 [0.0290]***	0.0813 [0.0338]**	(0.0093)
	Unskilled workforce (a)	0.0013 [0.0036]	(0.0016)	-0.0016 [0.0017]	-0.004 [0.0019]**	(0.1225)
	Workforce with computer	0.0017 [0.0011]	(0.0015)	0.0006 [0.0005]	0.0015 [0.0006]***	(0.0498)***
Other control variables	Dummy for incorporated company	0.0265 [0.0396]	(0.0901)	0.0566 [0.0225]**	0.016 [0.0273]	(0.0398)**
	Age	0.0534 [0.0267]**	(0.0214)**	0.0363 [0.0146]**	0.0856 [0.0181]***	(0.0431)**
	Share of exports (b)	0.001 [0.0009]	(0.0005)**	0.0001 [0.0004]	0.0003 [0.0004]	(0.0020)**
	Trade union (b)	0.0008 [0.0012]	(0.0010)	-0.00004 [0.0005]	0.0002 [0.0005]	(0.0014)**
	Strikes (b)	-0.0683 [0.0449]	(0.0821)	0.0482 [0.0301]	-0.0213 [0.0317]	(0.0043)
	Constant	0.7377 [0.3449]**		1.1579 [0.1899]***	0.8508 [0.2174]***	
	Industry/region/size/time dummies	Yes		Yes	Yes	
	Observations	5211		4323 (Cens: 5515/ Unc: 3808	5407 (Censored: 515/ Uncens: 4982)	
	R-squared	0.88				
	Heckman's Lambda (Inverse of Mills ration)			0.0130 [0.0634]	0.1221 [0.0926]	.

Estimating results of equation (1) under different imputation mechanisms for missing data. Those observations with missing values in all sales, labor (labor cost), materials and capital are excluded in all the regressions.[1] See footnote in Table 8.1. [2] Heckman models are explained in section 5.4. Heckman model on complete case considers missingness only in PF variables, not in IC variables, see section 5.4.1. Heckman replacing inputs compute the model on the sample with replacement of missing values in inputs (labor, materials and capital), see section 5.4.2.
In all the cases significance is given by clustered by industry and region White- robust standard errors in brackets; *** 1%, **5%, * 10%. In the case of the ICA method and Heckmand replacing inputs, in parentheses are bootstrap standard errors after 1000 replications (see sections and 5.2.2 5.4.2). Correlation by cluster is also considered. Source: Authors' calculations with ICSs.

Table 11.2: TURKEY, Extended production function and comparison of ICA method with Heckman models

Dependent variable: log of total sales		ICA Method[1]		Heckman models[2]		
				Heckman on complete case	Heckman replacing inputs	
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. Std.Err**	**Coeff. std. err.**	**Boot. st. er.**
PF variables	Log-employment	0.416 [0.0492]***	(0.1088)***	0.4017 [0.0423]***	0.5104 [0.0427]***	(0.0376)***
	Log-materials	0.4184 [0.0404]***	(0.0249)***	0.5306 [0.0189]***	0.4585 [0.0187]***	(0.0310)***
	Log-capital	0.0371 [0.0165]**	(0.0428)	0.063 [0.0164]***	0.067 [0.0182]***	(0.0168)***
Infrastructures	Days to clear customs for imports (a)	-0.0707 [0.0686]	(0.0688)	-0.155 [0.0835]*	-0.0648 [0.0859]	(0.0556)***
	Dummy for e-mail	0.2866 [0.0920]***	(0.1365)**	0.0193 [0.0822]	0.3121 [0.0786]***	(0.0659)**
Red tape, corruption and crime	Security expenses (b)	-0.0246 [0.0828]	(0.0011)***	-0.0379 [0.0824]	-0.0658 [0.0831]	(0.0575)**
	Payments to deal with bureaucratic issues (a)	-0.011 [0.0020]***	(0.0077)	-0.0084 [0.0009]***	-0.0101 [0.0010]***	(0.0012)***
	Sales declared to taxes (a)	-0.0226 [0.0057]***	(0.0045)***	-0.0175 [0.0075]**	-0.0131 [0.0077]*	(0.0055)***
	Number of inspections (b)	0.0046 [0.0044]	(0.0597)	-0.0017 [0.0043]	0.0049 [0.0045]	(0.0028)
	Payments to obtain a contract with the government (b)	-0.0373 [0.0315]	(0.0058)***	-0.0371 [0.0323]	-0.0363 [0.0315]	(0.0256)**
	Production lost due to absenteeism (b)	-0.0054 [0.0043]	(0.0367)	-0.0138 [0.0073]*	-0.0102 [0.0074]	(0.0042)**
	Dummy for informal competition (b)	0.0044 [0.0295]	(0.1203)	-0.011 [0.0283]	0.0046 [0.0306]	(0.0194)
	Delay in obtaining a water supply (a)	-0.1325 [0.0634]**	(0.0993)	-0.0926 [0.0588]	-0.165 [0.0603]***	(0.0467)***
Finance	Dummy for credit line	0.068 [0.0868]	(0.1383)	0.0473 [0.0621]	0.0493 [0.0644]	(0.0482)**
	Dummy for external auditory (a)	0.0863 [0.0753]	(0.1117)	0.1407 [0.0617]**	0.1075 [0.0641]*	(0.0448)***
	Loans in foreign currency (b)	0.0018 [0.0009]**	(0.0010)*	0.0003 [0.0009]	0.0016 [0.0009]*	(0.0008)*
Quality, innov. and labor skills	Staff with university education (b)	0.0095 [0.0026]***	(0.0018)***	0.0083 [0.0023]***	0.0104 [0.0024]***	(0.0018)***
	Staff-part time workers	-0.008 [0.0030]**	(0.0222)	-0.0065 [0.0027]**	-0.0093 [0.0028]***	(0.0019)***
Other control variables	Production lost due to strikes (b)	-0.1689 [0.0634]**	(0.0351)***	-0.1805 [0.0593]***	-0.153 [0.0723]**	(0.0453)***
	Dummy for recently privatized firm	1.0606 [0.2812]***	(0.2511)***	1.3287 [0.3695]***	1.0391 [0.2582]***	(0.2653)***
	Dummy for competition against imported products	0.2069 [0.0962]**	(0.2737)	0.021 [0.0724]	0.2084 [0.0730]***	(0.0634)***
	Constant	3.5299 [0.7190]***		3.0323 [0.6775]***	1.7704 [0.7084]**	(0.0376)***
	Industry/region/size/time dummies	Yes		Yes	Yes	
	Observations	1684		1941 (Censored: 1149/ Uncensored: 792)	2509 (Censored: 1149/ Uncensored: 1360)	
	R-squared	0.73				
	Heckman's Lambda			-0.1531 [0.1188]	0.0639 (0.1332]	

Notes for Table 11.1.
Source: Authors' calculations with ICSs.

Table 11.3: SOUTH AFRICA, Extended production function and comparison of ICA method with Heckman models

	Dependent variable: log of total sales	ICA Method[1]		Heckman models[2]		
				Heckman on complete case	Heckman replacing inputs	
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. Std.Err**	**Coeff. std. err.**	**Boot. st. er.**
PF variables	Log-employment	0.3226 [0.0711]***	(0.0365]***	0.3427 [0.0261]***	0.3275 [0.0250]***	(0.0452)***
	Log-materials	0.5195 [0.1017]***	(0.0214]***	0.4871 [0.0121]***	0.5184 [0.0120]***	(0.0567)***
	Log-capital	0.1247 [0.0300]***	(0.0118]***	0.1117 [0.0123]***	0.1241 [0.0129]***	(0.0168)***
Infrastructure	Days to clear customs for import s(a)	-0.1188 [0.1125]	(0.1233]	0.032 [0.1133]	-0.1728 [0.1286]	(0.1542)
	Sales lost due to power outages (b)	-0.0171 [0.0114]	(0.0047]***	-0.0059 [0.0062]	-0.0166 [0.0069]**	(0.0064)***
	Water outages (b)	-0.1477 [0.0527]***	(0.0942]	-0.1215 [0.0501]**	-0.1383 [0.0516]***	(0.0460)***
	Average duration of transport failures (a)	-0.0439 [0.0806]	(0.0379]	0.1092 [0.0936]	-0.0821 [0.0985]	(0.1197)
	Wait for electric supply (a)	-0.0867 [0.0553]	(0.0173]***	-0.0311 [0.0544]	-0.057 [0.0717]	(0.0742)
	Sales lost due to delivery delays (b)	-0.0099 [0.0083]	(0.0073]	-0.0069 [0.0054]	-0.0109 [0.0054]**	(0.0051)**
Red tape, corruption and crime	Manager's time spent on bur. issues (b)	0.007 [0.0051]	(0.0016]***	0.0065 [0.0016]***	0.0079 [0.0017]***	(0.0035)**
	Payments to deal with bureaucratic issues (b)	-0.0045 [0.0024]*	(0.3604]	-0.0028 [0.0101]	-0.0056 [0.0039]	(0.0073)
	Sales declared to taxes (a)	0.0056 [0.0046]	(0.0022]**	0.0079 [0.0041]*	0.0062 [0.0056]	(0.0049)
	Payments to obtain a contract with the government (b)	-0.0144 [0.0185]	(0.1975]	-0.0099 [0.0198]	-0.0134 [0.0228]	(0.0463)
	Security expenses (a)	0.1407 [0.0511]**	(0.0069]***	0.0308 [0.0152]**	0.1324 [0.0578]**	(0.0658)**
	Illegal payments in protection (b)	0.3969 [0.2428]	(0.1128]***	0.2767 [0.1745]	0.3686 [0.0888]***	(0.1783)**
	Crime losses (a)	-0.0502 [0.0788]	(0.1374]	0.1006 [0.0792]	-0.0561 [0.0817]	(0.0762)
Finance and corporate governance	Percentage of credit unused (b)	0.0014 [0.0010]	(0.0013]	0.0006 [0.0005]	0.0013 [0.0006]**	(0.0006)**
	Dummy for loan	0.0715 [0.0492]	(0.0327]**	0.0634 [0.0400]	0.0705 [0.0413]*	(0.0408)*
	Value of the collateral (b)	-0.0008 [0.0002]***	(0.0009]	-0.0009 [0.0002]***	-0.0008 [0.0002]***	(0.0002)***
	Loans in foreign currency (b)	0.0018 [0.0022]	(0.0024]	0.0013 [0.0012]	0.0015 [0.0012]	(0.0010)
	Charge to clear a check (a)	-0.1164 [0.0503]**	(0.0253]***	-0.1773 [0.0324]***	-0.1239 [0.0340]***	(0.0352)***
	Largest shareholder	0.0006 [0.0010]	(0.0008]	0.0000 [0.0006]	0.0008 [0.0007]	(0.0006)
	Working capital financed by foreign commercial banks (b)	0.0106 [0.0083]	(0.0084]	0.0241 [0.0070]***	0.0134 [0.0045]***	(0.0093)
	Working capital financed by informal sources (b)	-0.0022 [0.0023]	(0.0001]***	-0.0044 [0.0031]	-0.002 [0.0036]	(0.1225)
Quality, innovation and labor skills	Dummy for ISO quality certification (b)	0.1603 [0.0766]**	(0.0365]***	0.1208 [0.0359]***	0.1599 [0.0389]***	(0.0498)***
	Dummy for new product (b)	0.091 [0.0494]*	(0.0113]***	0.0322 [0.0377]	0.0807 [0.0398]**	(0.0398)**
	Dummy for discontinued product line (b)	-0.1007 [0.0610]	(0.0384]**	-0.0565 [0.0333]*	-0.0865 [0.0375]**	(0.0431)**
	Staff - management	0.004 [0.0028]	(0.0009]***	0.0047 [0.0016]***	0.0041 [0.0015]***	(0.0020)**
	Staff - non-production workers	-0.0034 [0.0022]	(0.0025]	-0.0027 [0.0011]**	-0.0033 [0.0012]***	(0.0014)**
	Training for unskilled workers (a)	0.001 [0.0026]	(0.0030]	-0.0048 [0.0032]	0.0012 [0.0041]	(0.0043)
	University staff (b)	0.0049 [0.0015]***	(0.0007]***	0.0036 [0.0015]**	0.0044 [0.0014]***	(0.0012)***
	Manager's experience (b)	0.0391 [0.0249]	(0.0217]*	0.0336 [0.0142]**	0.0369 [0.0150]**	(0.0173)**
Other control variables	Age (b)	0.0018 [0.0015]	(0.0016]	0.0016 [0.0009]*	0.0012 [0.0010]	(0.0011)
	Share of the local market (b)	0.0032 [0.0008]***	(0.0004]***	0.0028 [0.0006]***	0.0031 [0.0006]***	(0.0007)***
	Constant	2.7174 [0.8932]***		2.7155 [0.5500]***	2.7170 [0.6986]***	
	Industry/region/size/time dummies	Yes		Yes	Yes	
	Observations	1483		1443 (Censored: 2007/ Uncens: 1236)	1657 (Censored: 183/ Uncens.: 1484)	
	R-squared	0.89				
	Heckman's Lambda			-0.2747 [0.1993]	-0.2471 [0.2303]	

Notes for Table 11.1.
Source: Authors' calculations with ICSs.

Table 11.4: TANZANIA, Extended production function and comparison of ICA method with Heckman models

	Dependent variable: log of total sales	ICA Method[1]		Heckman models[2]		
				Heckman on complete case	Heckman replacing inputs	
Category	**Variable**	**Coeff. std. err.**	**Boot. st. er.**	**Coeff. std. err.**	**Coeff. std. err.**	**Boot. st. er.**
PF variables	Log-employment	0.1655 [0.0853]*	(0.0512)***	0.1422 [0.0557]**	0.1742 [0.0669]***	(0.0677)**
	Log-materials	0.4252 [0.0581]***	(0.0340)***	0.6176 [0.0274]***	0.6099 [0.0317]***	(0.0439)***
	Log-capital	0.1589 [0.0323]***	(0.0208)***	0.1427 [0.0209]***	0.1417 [0.0265]***	(0.0235)***
Infrastructure	Electricity from own generator (b)	0.0021 [0.0016]	(0.0053)	-0.001 [0.0018]	0.0041 [0.0020]**	(0.0017)**
	Losses due to water outages (b)	-0.0112 [0.0058]*	(0.0162)	-0.0081 [0.0060]	-0.0029 [0.0063]	(0.0054)
	Water from own well or water infrastructure (a)	0.0001 [0.0051]	(0.0011)	0.001 [0.0031]	0.0044 [0.0036]	(0.0042)
	Losses due to phone outages (a)	-0.0322 [0.0198]	(0.0071)***	-0.0315 [0.0284]	-0.0226 [0.0321]	(0.0291)
	Transport outages (a)	-0.0047 [0.0703]	(0.1168)	-0.1172 [0.0503]**	-0.0214 [0.0583]	(0.0499)
	Dummy for own roads (b)	0.289 [0.1488]*	(0.0581)***	0.3742 [0.1143]***	0.3416 [0.1444]**	(0.1321)***
	Dummy for webpage (b)	0.1578 [0.1212]	(0.1994)	0.3178 [0.0972]***	0.1595 [0.1208]	(0.1468)
	Wait for a water supply (a)	-0.1814 [0.0427]***	(0.0702)**	-0.1214 [0.0415]***	-0.1888 [0.0551]***	(0.0466)***
	Low quality supplies (a)	-0.0163 [0.0128]	(0.0041)***	-0.0252 [0.0116]**	-0.0323 [0.0118]***	(0.0130)**
Red tape, corruption and crime	Gift to obtain an operating license (b)	-0.4983 [0.1935]**	(0.1066)***	-0.1757 [0.1281]	0.0688 [0.1482]	(0.1589)
	Payments to deal with bureaucratic issues (b)	0.0939 [0.0299]***	(0.0164)***	0.0365 [0.0420]	0.0245 [0.0446]	(0.0495)
	Days in inspections (b)	-0.1045 [0.0735]	(0.0494)**	-0.1106 [0.0525]**	-0.0246 [0.0585]	(0.0580)
	Payments to obtain a contract with the government (b)	-0.0114 [0.0066]*	(0.0091)	-0.0332 [0.0088]***	-0.0101 [0.0074]	(0.0066)
	Security expenses (b)	-0.0119 [0.0042]***	(0.0092)	0.0068 [0.0108]	-0.0051 [0.0058]	(0.0052)
	Illegal payments in protection (b)	-0.0827 [0.0170]***	(0.1019)	-0.1209 [0.0478]**	-0.026 [0.0467]	(0.0493)
Finance and corporate governance	Interest rate of the loan (a)	-0.0109 [0.0145]	(0.0099)	0.0036 [0.0098]	-0.0074 [0.0115]	(0.0127)
	Working capital financed by commercial banks (b)	-0.0009 [0.0018]	(0.0012)	0.0000 [0.0014]	-0.003 [0.0016]*	(0.0015)**
	Working capital financed by leasing (b)	-0.0794 [0.0282]***	(0.0054)***	-0.0234 [0.0408]	-0.0473 [0.0096]***	(0.0806)
	Sales bought on credit (b)	-0.0014 [0.0012]	(0.0011)	-0.0029 [0.0012]**	0.0038 [0.0014]***	(0.0014)***
	Delay in clearing a domestic currency wire (a)	-0.3418 [0.3273]	(0.0935)***	0.4842 [0.1853]***	0.1533 [0.1876]	(0.1996)
Quality, innovation and labor skills	Dummy for new product (b)	0.0429 [0.1063]	(0.2036)	-0.1942 [0.0850]**	-0.003 [0.1014]	(0.0951)
	Staff - skilled workers (b)	0.0026 [0.0023]	(0.0050)	0.0074 [0.0020]***	0.0092 [0.0026]***	(0.0024)***
	Workforce with computer (b)	0.0066 [0.0030]**	(0.0056)	0.0183 [0.0037]***	-0.0084 [0.0032]***	(0.0070)
Other control variables	Dummy for incorporated company (b)	0.2914 [0.2023]	(0.5683)	-0.2149 [0.3207]	0.1701 [0.3050]	(0.1810)
	Dummy for FDI (b)	0.1397 [0.1445]	(0.2844)	0.1752 [0.1051]*	-0.0289 [0.1426]	(0.1326)
	Constant	7.2978 [1.0168]***		3.8725 [0.7997]***	3.1102 [0.9936]***	
	Industry/region/size/time dummies	Yes		Yes	Yes	
	Observations	559		581 (Censored: 290/ Uncens: 291)	771 (Censored: 317/ Uncens: 454)	
	R-squared	0.88				
	Heckman's Lambda			-0.2747[0.1993]	-0.2471[0.2303]	

Notes for Table 11.1
Source: Authors' calculations with ICSs.

First of all, we consider it important to note that Heckman's Lambda is significant in none of the four cases. Thereby, the plausible selection bias is not supported by the Heckman model in any country.

Besides the significance of Heckman's Lambda, the results are quite similar when we correct for the endogenous selection and when we do not. In India and South Africa there are no significant changes in the I-O elasticities. Nonetheless, the larger proportion of missing observations in Turkey and South Africa introduces some degree of heterogeneity between the results of the ICA method and the Heckman models. Even under very different estimated I-O elasticities, the IC parameters move within a reasonable range of values and there are no changes in the estimated direction of the effects. Overall, there are more IC variables significant in the Heckman model, even when we consider bootstrap standard errors.

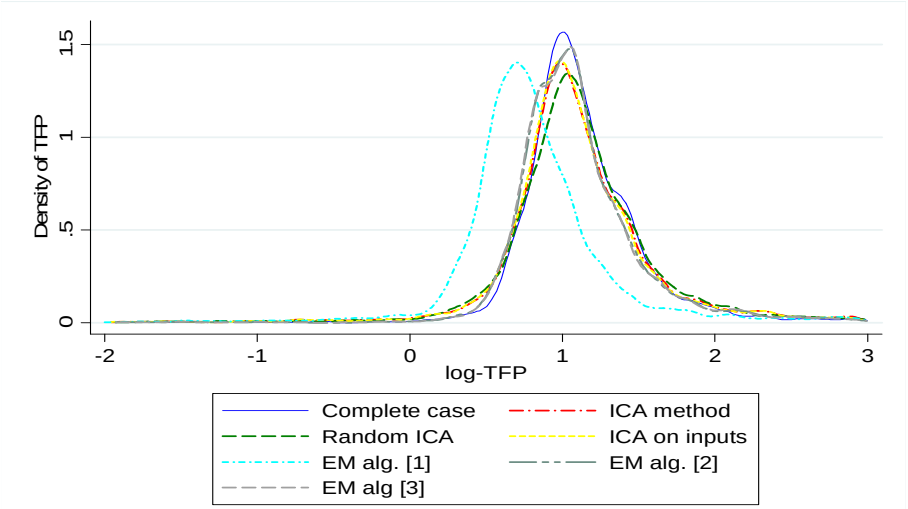
6.3 Evaluation of the imputation mechanism: Comparison of estimated TFP densities

We end this section with the evaluation of the estimated densities of the TFPs for each country. The estimated kernel densities of the different TFP measures obtained after applying the different imputation mechanism are obtained from equation (1) according to the following expression $\log TFP_{it} = s^*_{it}[\log \tilde{Y}_{it} - (\hat{\alpha}_L \log \tilde{L}_{it} + \hat{\alpha}_M \log \tilde{M}_{it} + \hat{\alpha}_K \log \tilde{K}_{it})]$, where $\log TFP_{it}$ is the measured productivity after the imputation process, $\log \tilde{Y}_{it}$, $\log \tilde{L}_{it}$, $\log \tilde{M}_{it}$, $\log \tilde{K}_{it}$ are the imputed inputs and output, the alphas with a hat on top denote the different estimated input-output elasticities after imputing missing values and s^* is the pattern of missing values in PF variables after the imputation process. The results are in figures 5.1 to 5.2, along with the descriptive statistics of each TFP measure and the correlation matrix among productivities.

Again we should differentiate between two groups of countries. In the first one, say that consisting of India and South Africa, the estimated TFP measures show a similar shape of kernel densities, although with different estimated means, especially in the case of EM algorithm [1] in India. In South Africa, this pattern is more marked, with more ostensible differences in the first moment of the distribution of the different TFP measures, although all the kernel densities have a similar shape, indicating that the standard deviations do not differ much among them, which is corroborated in panels B and C, where the descriptive statistics and the matrix of correlations are shown.

Figure 5.1: INDIA, evaluation of TFP measures under different imputation methods

I. Kernel[1] estimates of TFP densities



II. Table of descriptive statistics of TFP measures

	# Obs	Mean	Std. Dev.	Min	Max
Complete case	4327	1.15	0.68	-12.51	12.08
ICA meth.	5915	1.17	0.98	-12.53	12.19
Random ICA	5915	1.10	1.19	-12.51	12.15
ICA on inputs	5821	1.16	0.95	-12.55	12.25
Em alg [1]	6848	0.83	0.90	-12.96	12.09
Em alg [2]	5731	1.13	0.71	-12.66	12.44
Em alg [3]	5731	1.13	0.71	-12.67	12.43

III. Correlation matrix between TFP measures

	Complete case	ICA meth.	Random ICA	ICA on inputs	Em alg [1]	Em alg [2]	Em alg [3]
Complete case	1.000						
ICA meth.	0.999	1.000					
Random ICA	1.000	0.999	1.000				
ICA on inputs	0.998	1.000	0.998	1.000			
Em alg [1]	0.993	0.995	0.994	0.996	1.000		
Em alg [2]	0.990	0.991	0.993	0.993	0.997	1.000	
Em alg [3]	0.990	0.991	0.993	0.993	0.997	1.000	1.000

Notes:

[1] Epanechnikov kernel. Each point estimated within a range of 300 values.

Complete case: TFP measure from the sample without replacement of missing values; likewise, input-output elasticities are obtained from estimating equation (1) in the complete case (see I-O elasticities in Table 9.1).

ICA method: TFP measure with inputs and output replaced by *ICA method* and input-output elasticities from Table 8.1.

Random ICA: TFP measure with inputs and output replaced by *random ICA method* and input output elasticities from Table 9.1.

ICA on inputs: TFP measure when only inputs are imputed by the ICA method (not sales), the I-O elasticities and semi-elasticities used are in Table 9.1.

Em alg. [1]: TFP measure obtained under imputation of inputs and output by the EM algorithm described in section 5.1.1. Likewise, the I-O elasticities are in Table 9.1.

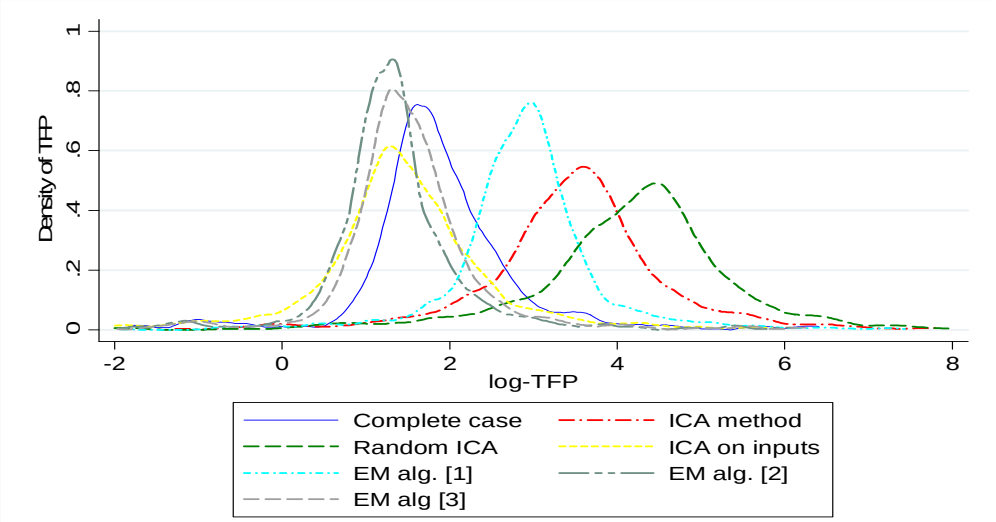
Em alg. [2]: In this case the EM algorithm used is that described in section 5.1.2. The I-O elasticities are in Table 9.1.

Em alg. [3]: The description of the EM algorithm is in section 5.1.3. The I-O elasticities are in Table 9.1.

Source: Authors' estimations with ICSs.

Figure 5.2: TURKEY, evaluation of TFP measures under different imputation methods

I. Kernel[1] estimates of TFP densities



II. Table of descriptive statistics of TFP measures

	# Obs	Mean	Std. Dev.	Min	Max
Complete case	818	1.84	1.01	-5.25	6.41
ICA meth.	1805	3.45	1.20	-3.36	7.85
Random ICA	1805	4.16	1.28	-2.67	8.91
ICA on inputs	1481	1.37	1.23	-5.64	5.87
Em alg [1]	2646	2.87	0.97	-4.05	7.44
Em alg [2]	1802	1.33	0.88	-5.84	6.13
Em alg [3]	1802	1.51	0.88	-5.65	6.31

III. Correlation matrix between TFP measures

	Complete case	ICA meth.	Random ICA	ICA on inputs	Em alg [1]	Em alg [2]	Em alg [3]
Complete case	1.000						
ICA meth.	0.969	1.000					
Random ICA	0.954	0.998	1.000				
ICA on inputs	0.992	0.974	0.956	1.000			
Em alg [1]	0.990	0.993	0.986	0.985	1.000		
Em alg [2]	0.990	0.927	0.908	0.969	0.964	1.000	
Em alg [3]	0.991	0.932	0.914	0.969	0.968	1.000	1.000

Notes:

[1] Epanechnikov kernel. Each point estimated within a range of 300 values.

Complete case: TFP measure from the sample without replacement of missing values; likewise, input-output elasticities are obtained from estimating equation (1) in the complete case (see I-O elasticities in Table 9.2).

ICA method: TFP measure with inputs and output replaced by *ICA method* and input-output elasticities from Table 8.2.

Random ICA: TFP measure with inputs and output replaced by *random ICA method* and input output elasticities from Table 9.2.

ICA on inputs: TFP measure when only inputs are imputed by the ICA method (not sales).The I-O elasticities and semi-elasticities used are in Table 9.2.

Em alg. [1]: TFP measure obtained under imputation of inputs and output by the EM algorithm described in section 5.1.1. Likewise, the I-O elasticities are in Table 9.2.

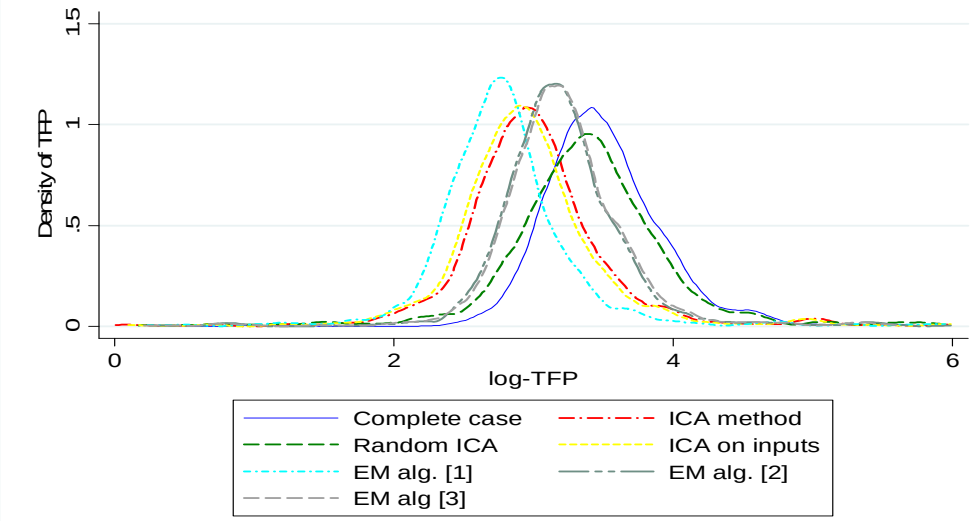
Em alg. [2]: In this case the EM algorithm used is that described in section 5.1.2. The I-O elasticities are in Table 9.2.

Em alg. [3]: The description of the EM algorithm is in section 5.1.3. The I-O elasticities are in Table 9.2.

Source: Authors' estimations with ICSs.

Figure 5.3: SOUTH AFRICA, evaluation of TFP measures under different imputation methods

I. Kernel[1] estimates of TFP densities



II. Table of descriptive statistics of TFP measures

	# Obs	Mean	Std. Dev.	Min	Max
Complete case	1265	3.50	0.70	-3.74	10.34
ICA meth.	1585	2.99	0.84	-4.34	10.28
Random ICA	1585	3.38	0.90	-4.97	10.31
ICA on inputs	1576	2.94	0.84	-4.39	10.21
Em alg [1]	1784	2.78	0.80	-4.47	10.26
Em alg [2]	1581	3.21	0.72	-4.01	11.21
Em alg [3]	1578	3.22	0.72	-4.00	11.18

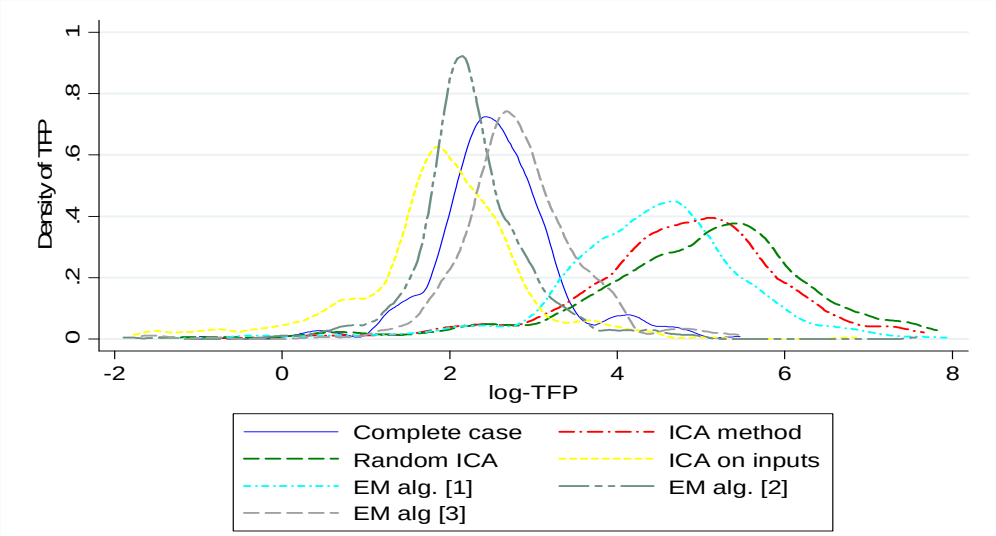
III. Correlation matrix between TFP measures

	Complete case	ICA meth.	Random ICA	ICA on inputs	Em alg [1]	Em alg [2]	Em alg [3]
Complete case	1.000						
ICA meth.	0.996	1.000					
Random ICA	0.998	0.993	1.000				
ICA on inputs	0.996	1.000	0.993	1.000			
Em alg [1]	0.992	0.999	0.988	0.999	1.000		
Em alg [2]	0.982	0.991	0.975	0.990	0.992	1.000	
Em alg [3]	0.982	0.991	0.975	0.990	0.993	1.000	1.000

Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
Complete case: TFP measure from the sample without replacement of missing values; likewise, input-output elasticities are obtained from estimating equation (1) in the complete case (see I-O elasticities in Table 9.3).
ICA method: TFP measure with inputs and output replaced by *ICA method* and input-output elasticities from Table 8.3.
Random ICA: TFP measure with inputs and output replaced by *random ICA method* and input output elasticities from Table 9.3.
ICA on inputs: TFP measure when only inputs are imputed by the ICA method (not sales),.The I-O elasticities and semi-elasticities used are in Table 9.3.
Em alg. [1]: TFP measure obtained under imputation of inputs and output by the EM algorithm described in section 5.1.1. Likewise, the I-O elasticities are in Table 9.3.
Em alg. [2]: In this case the EM algorithm used is that described in section 5.1.2. The I-O elasticities are in Table 9.3.
Em alg. [3]: The description of the EM algorithm is in section 5.1.3. The I-O elasticities are in Table 9.3.
Source: Authors' estimations with ICSs.

Figure 5.4: TANZANIA, evaluation of TFP measures under different imputation methods

I. Kernel[1] estimates of TFP densities



II. Table of descriptive statistics of TFP measures

	# Obs	Mean	Std. Dev.	Min	Max
Complete case	313	2.53	0.87	-3.21	5.47
ICA meth.	661	4.79	1.30	-0.75	9.72
Random ICA	661	4.98	1.50	-1.47	10.03
ICA on inputs	505	1.81	1.14	-3.68	6.85
Em alg [1]	790	4.39	1.18	-1.30	8.92
Em alg [2]	628	2.25	0.80	-3.82	7.64
Em alg [3]	638	2.81	0.86	-3.21	8.35

III. Correlation matrix between TFP measures

	Complete case	ICA meth.	Random ICA	ICA on inputs	Em alg [1]	Em alg [2]	Em alg [3]
Complete case	1.000						
ICA meth.	0.913	1.000					
Random ICA	0.877	0.991	1.000				
ICA on inputs	0.997	0.904	0.869	1.000			
Em alg [1]	0.948	0.994	0.975	0.937	1.000		
Em alg [2]	0.981	0.829	0.779	0.971	0.884	1.000	
Em alg [3]	0.979	0.849	0.804	0.963	0.901	0.996	1.000

Notes:
[1] Epanechnikov kernel. Each point estimated within a range of 300 values.
Complete case: TFP measure from the sample without replacement of missing values; likewise, input-output elasticities are obtained from estimating equation (1) in the complete case (see I-O elasticities in Table 9.4).
ICA method: TFP measure with inputs and output replaced by *ICA method* and input-output elasticities from Table 8.4.
Random ICA: TFP measure with inputs and output replaced by *random ICA method* and input output elasticities from Table 9.4.
ICA on inputs: TFP measure when only inputs are imputed by the ICA method (not sales). The I-O elasticities and semi-elasticities used are in Table 9.4.
Em alg. [1]: TFP measure obtained under imputation of inputs and output by the EM algorithm described in section 5.1.1. Likewise, the I-O elasticities are in Table 9.4.
Em alg. [2]: In this case the EM algorithm used is that described in section 5.1.2. The I-O elasticities are in Table 9.4.
Em alg. [3]: The description of the EM algorithm is in section 5.1.3 and the I-O elasticities in Table 9.4.
Source: Authors' estimations with ICSs.

In Turkey and Tanzania the results are somewhat different. The larger proportion of missing values in these two countries results in two different blocks of TFP measures. The first block comprises the TFP measures from the complete case, the ICA method on inputs, and the EM algorithms [2] and [3]. The second block includes the remaining measures, that is, those from the ICA method, the EM algorithm [1] and the Random ICA method. TFP measures are similar within each group, however between blocks there are evident differences in all the shapes of the distribution, the skewness, the kurtosis, as well as in the estimated means and standard errors, as panel B shows. In spite of all these differences, panel C shows that the correlations of the TFP measure from the ICA method with the remaining cases are between .8 and .99. Likewise, the correlation among the remaining measures is considerably high.

6.4 Summary and main conclusions

The ICA method performs reasonably well. Even under very different patterns of missing data and assumptions we are able to get robust results from different methods of handling missing data after controlling for IC variables in the estimation. When we assume that the MDM is MAR then there are two main issues we should consider: uncertainty and amount of information used in the imputation. On the other hand, if a non-ignorable pattern of missing data is assumed, then we are forced to test the robustness of the results of the ICA method with the Heckman models.

We find that, overall, the ICA method is a good alternative even when the proportion of missing values is relatively high and the underlying variables are manifestly non-normal., leading to rather more homogenous results than other more sophisticated methods. We also observe that uncertainty, amount of information and non-ignorability of the MDM are not big issues in the context of ICSs; or at least they are not so serious as to invalidate the results of the ICA method. Lastly, we find that in order to get robust results under different imputation mechanisms, it is essential to control for the same set of IC variables, as they contain a good deal of information on the MDM.[1]

The main conclusions of this section can be summarized as follows:

- Overall, there are small differences in the estimated distribution of the imputed PF variables. Nonetheless, these differences become more marked as the number of missing values imputed increases and when the variables are not normally distributed. In particular, the Random ICA method, is the mechanism with the worst performance under a large proportion of missing values, followed by the EM algorithms. The ICA method preserves with reasonable precision the main moments of the distribution of the variables in the complete case.[2]

[1] Obviously, this assertion is conditioned by the objectives one may have.

[2] This would imply that the ICA method performs well when the MDM is MCAR or MAR, since in that case, under regularity conditions, the distribution in the complete case shares the same characteristics as the population distribution. Nonetheless, at this point if the MDM is non-ignorable we cannot say anything about the goodness-of- fit of the ICA method, since it could be replicating any distribution different from the population distribution.

- These differences in the estimated distributions become even clearer if we focus on the TFP. However, the conclusions are the same whether we focus on inputs and output or TFP.
- We found reasonably robust elasticities in equation (1) under all the imputation methods proposed. However, there are important differences in the I-O elasticities and in the significance of the IC variables.
- The ICA method, EM algorithm [1], Random ICA method and Bootstrap ICA method lead to homogeneous results among them. That is, introducing uncertainty into the ICA method, regardless of whether, in order to get it, we use the EM algorithm [1], Random ICA method or Bootstrap ICA method, does not change significantly either the estimated effects or the level of significance of IC variables This suggests that uncertainty is not a big issue. Obviously, there are slight differences in the standard errors, but we argue that they are not so serious as to invalidate the results of the ICA method.
- In all cases, EM algorithms [2] and [3] lead to differences in the I-O estimates, although the IC parameters are again quite robust and do not vary much, the level of significance is affected in a higher proportion of cases than in the EM algorithm [1].
- More importantly, EM algorithms [2] and [3] are not homogeneous among themselves, suggesting that the amount of information embodied in the imputation algorithm does not consequently improve the results.
- Another interesting observation is that the performance of the EM algorithms [2] and [3] greatly depends on the structure of the MDM. When the pattern of missing data is very unbalanced, meaning that it is common to observe only one or two PF variables in each cross-sectional observation, these two EM algorithms lead to rather different results from the ICA method and EM algorithm [1]. Intuitively, this is probably due to the unbalanced amount of information included in each cross-sectional observation.
- Only in Tanzania and Turkey, when the proportion of missing values is larger than in the other two countries, do we observe significant changes in the estimated I-O elasticities under the Heckman models with respect to the ICA method.
- As a general rule, there are more significant IC variables under the Heckman models than under the ICA method.
- Heckman's Lambda is never significant, which does not support the story of non-ignorable MDM and confirms that correcting for endogenous selection does not change considerably the results.
- It is also important to note that it does not matter whether we replace only the independent variables, the dependent variable or both of them. In all the cases, the results are similar. More importantly, the Heckman model with the inputs replaced by the ICA method and the case of the ICA method on the inputs are similar in both cases.
- Finally, we find it essential to control for IC variables in the estimation in all the cases. We believe that this is what allows us to get such robust results under very

different assumptions and patterns of missing data. This is supported by section 4.4, where we saw that IC variables are able to explain a rather large proportion of the variability of the MDM in all the countries.

7. Conclusions

When the missing data mechanism (MDM) is ignorable, the objective of the imputation methods is not to augment the sample size, but to preserve the sample representativity, to gain efficiency in the estimation and to retrieve for the analysis a large number of very expensive interviews. The alternative to these methods is the complete case or *listwise deletion*, which is not a panacea even when the MDM is ignorable. Operating with the complete case is only acceptable if incomplete cases attributable to missing data comprise a small percentage, say 5% or less, of the number of total cases (Schafer, 1997), and when the complete case preserves the representativeness of the original sampling frame. In addition, in models with a large number of regressors, the problem of missing data may encourage analysts to leave out of the regression some explanatory variables with a high proportion of missing values. As Cameron and Trivedi (2005) point out, this practice may be misleading as it leads to an omitted variables problem, which could be more serious than the missing data problem *per se*. The first question we raise in this paper is, hence, whether the researcher should do something about the missing values when dealing with investment climate surveys (ICSs).

In the context of ICSs, a large proportion of the sample size is lost in the complete case and the representativeness of the original sample frame is, to some extent, modified. Given these results, the MDM can in no way be considered as missing completely at random (MCAR), and consequently a complete case could lead to inconsistent and inefficient results. In order to overcome this problem, we propose a imputation mechanism that fits well with the characteristics of ICSs—with unbalanced patterns of missing data and a low proportion of available observations in the complete case—likely to be used to construct structural models composed of single, or even systems of, equations with a large number of explanatory variables, all of them containing missing data.

The imputation method proposed, which we call the *ICA method*, departs from the class of EM type algorithms and relies on the expectation of the imputed variables conditional to the sector, region and size they belong to. The performance of the ICA method depends on several characteristics of the MDM, such as the number of variables replaced or the proportion of missing values in the complete case; but especially, it depends on the nature of the MDM: *missing at random (MAR) or non-ignorable*. Taking this into account, we analyze the MDM of four countries with very different patterns of missing data (India, Turkey, South Africa and Tanzania) to find out to what extent the MDM can be treated as MAR or not. Although not conclusive on the nature of the MDM, the descriptive analysis shows that this has to do with a variety of IC determinants, such as informality and corruption and also with the capacity of the firms. More dynamic firms engaged in R&D, quality, innovation of new products, technologies and operating in more exigent and competitive

export markets tend to report fewer missing values. Accountability and size can by themselves explain a large share of missing data too. On the other hand, the analysis does not allow us to reject the non-ignorability assumption on the MDM in any case.

In addition, given the results of the descriptive analysis and apart from the discussion concerning MAR and non-ignorable MDM, an interesting result is the need to control for those variables related with the MDM. Inconsistency would follow if we did not control for the large set of IC variables in the estimation.

In the next step of the analysis presented in the paper, we estimated an extended production function under imputation of missing values by the ICA method and we test the estimating results against other imputation mechanisms. We first considered imputation mechanisms requiring the MAR assumption like the ICA method, including the complete case, EM algorithms, extensions of the ICA method and multiple imputation. We then included in the analysis methods considering the non-ignorable assumption on the MDM; essentially we considered the Heckman model under different specifications.

Although caution is always a requisite when drawing conclusions from a model with imputed data, the *ICA method* leads the results to be more robust than even more sophisticated imputation methods also requiring the MAR assumption. We observe that more complex imputation mechanisms are rather sensitive to both the proportion of missing values and how these missing values are distributed among variables. When the MDM is very unbalanced, in the sense that we can observe only one or two PF variables for each cross-sectional observation, those EM algorithms including additional explanatory variables, such as inputs or IC variables, lead to changes in the results compared with the more linear, parsimonious and simpler ICA method and EM algorithm [1], both including only industry/region/size variables always available. This suggests that more complex imputation methods based on simulations, especially EM algorithms and multiple imputation based on Markov Chain Monte Carlo, require a deeper and more thorough knowledge of MDM that would allow us to handle proper assumptions on the unknown densities of data generating processes. The issue of the sensitivity of the results to the selection of a proper model for the MDM constitutes an interesting question to be handled in further research regarding ICSs.[3]

In this sense, we believe that incorporating systematically more information concerning the imputation mechanism does not constitute, per se, an improvement in the estimates. Rather, given the sensitivity of the results to the model choice for the MDM, extending the matrix of covariates used to impute missing values requires detailed, thorough knowledge of the determinants of the MDM, and this is likely to vary from country to country.

Regarding the lack of uncertainty inherent in the ICA method as a deterministic imputation method, we find that using other mechanisms allowing for additional uncertainty in the imputation mechanisms, such as the so-called Random ICA method, Bootstrap ICA method or EM algorithms, does not change the results significantly. Despite changes in the

[3] ICSs in particular and data collected from developing countries in general present the missingness issue as an additional challenge for applied researchers. We consider that a proper, systematic methodology to deal with this problem is required, especially if more sophisticated imputation mechanisms are applied.

level of significance of some coefficients, most of the variables remain significant when incorporating additional randomness. Nonetheless, we also observe that the randomness issue becomes more important as the proportion of missing values increases (in the cases of Turkey and Tanzania).

On the other hand, provided we control for the same set of IC variables in all the specifications, the results under the complete case and the ICA method are reasonably consistent between the two. Even in those cases in which the complete case represents less than half of the original sampling frame, the estimated parameters of production function (PF) and IC variables is within a reasonable range of values. This illustrates the importance of using the large set of IC variables, in order to control for the data generating process in the estimation.[4]

Likewise, the ICA method shows reasonable robustness to the endogenous sampling case. Heckman's lambda is non-significant in all cases, which does not support the endogenous sampling selection hypotheses. The results of the ICA method are similar to those of the Heckman regressions, indicating that even if there were an endogenous sampling selection problem, this would not be serious enough to bias the final results. In this sense, replacing only those RHS variables and not the dependent variable (sales in our case) does not change the results, provided the endogenous sample selection is not supported by the models and the robustness in the results.

As the use of Investment Climate Surveys becomes more and more important among policy makers, scholars and applied researchers, thorough research into the causes of the missingness problem in order to improve the quality of the data is becoming a requisite. The parsimonious methodology we propose here is intended to be a first step in helping prepare the way forward and delve further into this line of research.

[4] In order to pursue this issue more deeply, further research is needed. Nonetheless, once the relation between IC variables and the MDM is proved, using them to gain independency between our model and the MDM is a requisite. We believe that this procedure is what balances the results, in the sense that it is what allows us to get robust results in specifications.

References

Allison, P.D (2001): "Missing Data," *Quantitative Applications in the Social Sciences*. Sage University Paper.

Buuren v S., H.C. Boshuizen and D.L. Knook (1999): "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*; 18(6):681-94.

Cameron, A.C. and P.K Trivedi (2005): "Microeconometrics: Theory and Applications," *Cambridge University Press.*

Dempster A.P., N.M. Laird and D.B. Rubin (1977): "Maximum Likelihood Estimation for Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Escribano, A., and J. L. Guasch (2005): "Assessing the Impact of Investment Climate on Productivity Using Firm Level Data: Methodology and the Cases of Guatemala, Honduras and Nicaragua," *World Bank Policy Research Working Paper* #3621, Washington, DC.

Escribano, A., and J. L. Guasch (2008): "Robust Methodology for Investment Climate Assessment on Productivity: Application to Investment Climate Surveys from Central America," *Working Paper* # 08-19 (11), *Universidad Carlos III de Madrid*.

Escribano, A., J. L. Guasch, and M. de Orte (2009): "INDIA: Investment Climate Assessment on Productivity, Allocative Efficiency and Other Economic Performance Measures of the Manufacturing Sector,"mimeo, *Universidad Carlos III de Madrid*.

Escribano, A., J.L. Guasch, M. de Orte and J. Pena (2008a): "Investment Climate and Firm's Performance: Econometric and Applications to Turkey's Investment Climate Survey," *Working Paper* # 08-20 (12), *Universidad Carlos III de Madrid*.

Escribano, A., J.L. Guasch, M. de Orte and J. Pena (2008b): "Investment Climate Assessment Based on Demean Olley and Pakes Decompositions: Methodology and Applications to Turkey's Investment Climate Survey," *Working Paper* # 08-20 (12), *Universidad Carlos III de Madrid*.

Escribano, A., J. L. Guasch and J. Pena (2009): "Assessing the Impact of Infrastructure Quality on Firm Productivity in Africa," *World Bank Policy Research Working Paper* #(forthcoming), Washington DC.

Escribano, A., M. de Orte, J. Pena and J. L. Guasch. (2009): "Investment Climate Assessment on Economic Performance Using Firm Level Data: Pooling Manufacturing Firms from Indonesia, Malaysia, Philippines and Thailand from 2001 to 2002," *Singapore Economic Review, Special Issue on the Econometric Analysis of Panel Data*, Vol. 54, #3, August 20092009.

Gelman, A., G. King and C. Liu (1998): "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys," *Journal of the American Statistical Association*, 93, 846-874.

Griliches, Z. (1986): "Economic Data Issues," *Handbook of Econometrics*, Vol, III. Ed. R.F. Engle and D. McFadden. Amsterdam: North Holland, 1464-1514.

Heckman, J.J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475-492.

Heitjan, D.F. (1994): "Ignorability in General Incomplete-Data Models," *Biometrika*, 81, 701-708.

_____(1999): "Causal Inference in Clinical Trials, A Comparative Example," *Controlled Clinical Trials*, 20, 309-318.

Heitjan D.F. and S. Basu (1999): "Distinguishing 'Missing at Random' and `Missing Completely at Random'," *The American Statistician*, 50, 207-213.

Ibrahim, J.G., S.R. Lipsitz and M-H Chen (1999): "Missing Covariates in Generalized Linear Models When the Missing Data Mechanism is Ignorable," *Journal of the Royal Statistical Society*, Ser. B, 61, 173-190.

Little R.J.A. and D.B. Rubin, (1987): "Statistical Analysis with Missing Data," *Wiley Series in Probability and Mathematical Statistics*. John Wiley and Sons Eds.

McLachlan G.J and T. Krishnan (1997): "The EM Algorithm and Extensions," New York: *Wiley*.

Meng, X.L. (2000): "Missing Data: Dial M for?" *Journal of the American Statistical Association*, Vol. 95, No. 452, (Dec., 2000), pp. 1325 -1330.

Molerberghs, G., E.J. Goetghebeur, S.R. Lipsitz, M.G. Kenward (1999): "Nonrandom Missingness in Categorical Data: Strengths and Limitations," *The American Statistician*, 53, 110-118.

Murphy, K. M. and R. H. Topel (1985): "Estimation and inference in two-step econometric models," *Journal of Business and Economic Statistics,* 3(4): 370–379.

Newey, W.K (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economic Letters,* 14, 201-206.

Newey, W.K and D. McFadden (1994); "Large Sample Estimation and Hypotheses Testing," *Handbook of Econometrics*, Vol, 4. Ed. R.F. Engle and D. McFadden. Amsterdam: North Holland, 2111-2245.

Pagan, A.R (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review,* 25, 221-247.

Rubin, D.B. (1976): "Inference and Missing Data," *Biometrika*, 63, 581-592.

Rubin, D.B. (1987): "Multiple Imputation for Nonresponse in Surveys," *New York: Wiley*.

Schafer, J.L (1997): "Analysis of Incomplete Multivariate Data," *London: Chapman and Hall*.

Schafer, J.L (1999): "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3-15.

Wooldridge, J.M (2007): "Econometric Analysis of Cross Section and Panel Data," *The MIT Press*. Cambridge, Massachusetts.

Appendix I: definition of IC variables

I Infrastructures			
IC variables	**Country**	**Measurement units**	**Definition**
Longest # of days to clear customs for exports	(IND)	Log	Longest number of days that it took to clear customs when exporting
Days to clear customs for imports	(TUR, SA)	Log	Average number of days that it takes to clear customs when importing
Dummy for own generator	(IND)	0 or 1	Dummy variable taking value 1 if the firm has own generator
Electricity from own generator	(TUR, TZA)	Percentage	Percentage of total electricity used that came from own generators
Losses due to power outages	(IND, TUR, SA, TZA)	Perc	Percentage of total annual sales lost as a result of power outages
Wait for electric supply	(SA)	Log	Average number of days that it takes to obtain a power supply
Water supply from public sources	(IND)	Perc.	Percentage of the water used by the establishment that came from public sources
Water from own well or water infrastructure	(SA)	Perc.	Percentage of the water used by the establishment that came from own well or water infrastructures
Losses due to water outages	(TUR, TZA)	Perc.	Percentage of total annual sales lost as a result of water outages
Water outages	(SA)	Log	Total number of water outages experienced per year
Wait for a water supply	(TUR, TZA)	Log	Average number of days that it takes to obtain a water supply
Shipment losses in the domestic market	(IND, TUR)	Perc.	Percentage of products shipped that were lost as a consequence of theft, breakage, or spoilage
Dummy for own transport	(IND)	0 or 1	Dummy variable taking value 1 if uses own transport services
Average duration of transport failures	(SA)	Log	Average duration in hours of transport failures
Transport outages	(TZA)	Log	Total number of transport failures per year
Losses due to transport delay	(IND, TZA)	Perc.	Percentage of total annual sales lost as a consequence of transport delays
Losses due to phone outages	(TZA)	Perc.	Percentage of total annual sales lost as a consequence of phone interruptions
Dummy for web page	(IND, SA, TZA)	0 or 1	Dummy variable taking value 1 if the firm uses web page to communicate with clients or suppliers
Dummy for e-mail	(IND, TUR, SA)	0 or 1	Dummy variable taking value 1 if the firm uses e-mail to communicate with clients or suppliers
Sales lost due to delivery delays	(SA)	Perc.	Percentage of total annual sales lost as a consequence of delivery delays
Dummy for own roads	(TZA)	0 or 1	Dummy variable taking value 1 if the firm has own roads.
Low quality supplies	(TZA)	Perc.	Percentage of total supplies that were of lower than agreed upon quality per year
Days of inventory of main supply	(TZA)	Log	Days of inventory that the establishment kept its main supply in storage on average during the last year

II Red tape, corruption and crime			
IC variables	**Country**	**Measurement units**	**Definition**
Crime losses	(TUR, SA)	Perc.	Percentage of total annual sales lost as a consequence of crime, vandalism or arson
Dummy for security	(IND)	0 or 1	Dummy variable taking value 1 if the firm has security expenses
Security expenses	(TUR, SA, TZA)	Perc.	Security expenses as a percentage of total annual sales
Illegal payments for protection	(SA, TUR)	Perc.	Illegal payments for protection (e.g. to organized crime) to prevent violence as a percentage of total annual sales per year
Manager's time spent on bur. Issues	(TUR, SA)	Perc.	Percentage of manager's time spent in dealing with bureaucratic issues
Payments to deal with bureaucratic issues	(TUR, SA, TZA)	Perc.	Payments to deal with bureaucratic issues as a percentage of total annual sales
Payments to obtain a contract with the government	(TUR, SA, TZA)	Perc.	Payments to obtain a contract with the government as a percentage of total annual sales
Dummy for payments to speed up bureaucracy	(IND)	0 or 1	Dummy variable taking value 1 if the establishment declared making payments to 'speed up' bureaucratic issues
Dummy for payments to deal with bur. issues	(IND)	0 or 1	Dummy taking value 1 if the firm declared making 'irregular' payments to deal with bureaucratic issues
Dummy for interventionist labor regulation	(IND)	0 or 1	Dummy taking value 1 if the firm considers that regulation affected its decisions to hire or fire employees
Gift to obtain a operating license	(TZA)	Perc.	Gifts as a percentage of total annual sales paid to get an operating license
Number of inspections	(TUR)	Log	Total number of inspections received by the firm per year
Days in inspections	(TZA)	Log	Total number of days that the firm received inspections from public officials during the last year
Sales reported for taxes	(IND, TUR, SA)	Perc.	Percentage of total annual sales reported to IRS tax authorities
Workforce reported for taxes	(IND)	Perc.	Percentage of total workforce reported to IRS tax authorities
Production lost due to absenteeism	(IND, TUR)	Log	Days production lost as a consequence of employees' absenteeism
Dummy for informal competition	(TUR)	0 or 1	Dummy variable taking value 1 if the firm declared competing against informal competition
Dummy for lawsuit	(TUR)	0 or 1	Dummy variable taking value 1 if the firm had any lawsuit during the last year

III Finance			
IC variables	**Country**	**Measurement units**	**Definition**
Dummy for external audit	(IND, TUR, SA)	0 or 1	Dummy taking value 1 if the firm has its annual statements reviewed by an external auditor
Dummy for trade association	(IND)	0 or 1	Dummy variable taking value 1 if the firm belongs to a trade association
Dummy for loan	(IND, SA)	0 or 1	Dummy taking value 1 if the firm has access to a loan from any financial institution
Largest shareholder	(IND, SA)	Perc.	Percentage of firm's equity that belongs to the largest shareholder
Dummy for credit line	(TUR, SA, TZA)	0 or 1	Dummy taking value 1 if the firm has access to a credit line from any financial institution
Percentage of credit unused	(SA)	Perc.	Percentage of the credit line that is currently unused
Dummy for loan with collateral	(IND)	0 or 1	Dummy taking value 1 if the firm has a loan with associated collateral
Value of the collateral	(SA)	Perc.	Value of the collateral as a percentage of the total value of the loan
Loans denominated in foreign currency	(IND, TUR, SA, TZA)	Perc.	Percentage of total firm's loans that were denominated in foreign currency
Dummy for loan denominated in Turkish Lira	(TUR)	0 or 1	Dummy taking value 1 if the firm has access to a loan denominated in Turkish Lira
Dummy for loan denominated in foreign currency	(TUR)	0 or 1	Dummy taking value 1 if the firm has access to a loan denominated in foreign currency
Dummy for long-term loan	(TUR)	0 or 1	Dummy taking value 1 if the firm has access to a loan for more than 1 year
Interest rate of the loan	(TZA)	Perc.	Interest rate of the last loan obtained by the firm
Dummy for new land purchased	(TUR)	0 or 1	Dummy taking value 1 if the firm obtained new land in the last year
Charge to clear a check	(SA)	Perc.	Charges to clear a check as a percentage of the value of the check
Delay in clearing a domestic currency wire	(TZA)	Log	Average number of days that it takes to clear a domestic currency wire
Working capital financed by domestic private banks	(IND)	Perc.	Percentage of working capital financed by funds from domestic private banks
Working capital financed by commercial banks	(TZA)	Perc.	Percentage of working capital financed by funds from commercial banks
Working capital financed by foreign commercial banks	(SA)	Perc.	Percentage of working capital financed by funds from foreign commercial banks
Working capital financed by informal sources	(SA)	Perc.	Percentage of working capital financed by funds from informal sources
Working capital financed by leasing	(TZA)	Perc.	Percentage of working capital financed by funds from leasing arrangement
Dummy for current or saving account	(TZA)	0 or 1	Dummy taking value 1 if the firm has access to a current or saving account
Inputs bought on credit	(TZA)	Perc.	Percentage of inputs bought on credit per year
Sales bought on credit	(TZA)	Perc.	Percentage of sales bought on credit per year

IV Quality innovation and labor skills			
IC variables	**Country**	**Measurement units**	**Definition**
Dummy for R&D	(IND)	0 or 1	Dummy taking value 1 if the firm invests in R&D
Dummy for new technology	(TUR, TZA)	0 or 1	Dummy taking value 1 if the firm introduced new technology inherent to the production process during the last year
Dummy for new product	(SA, TZA)	0 or 1	Dummy taking value 1 if the firm introduced a new product of product line during the last year
Dummy for product innovation	(IND, TZA)	0 or 1	Dummy taking value 1 if the firm introduced a product innovation during the last year
Dummy for discontinued product line	(SA)	0 or 1	Dummy taking value 1 if the firm discontinued the production of any product during the last year
Dummy for foreign license	(IND, TUR, TZA)	0 or 1	Dummy taking value 1 if the firm has a technology licensed from a foreign company
Dummy for internal training	(IND, SA, TZA)	0 or 1	Dummy taking value 1 if the firm provides training to its employees
Training for unskilled workers	(SA)	Perc.	Percentage of unskilled workers that received training during the last year
Workforce with computer	(IND, TZA)	Perc.	Percentage of workers on the staff that regularly uses computer at job
Dummy for ISO quality certification	(IND, TUR, SA)	0 or 1	Dummy taking value 1 if the firm has an ISO quality certification
Dummy for outsourcing	(IND, SA, TZA)	0 or 1	Dummy taking value 1 if the firm outsourced any part of production in the last year
Dummy for brought in house	(TZA)	0 or 1	Dummy taking value 1 if the firm brought in house any part of the production process previously outsourced
Dummy for external training	(IND)	0 or 1	Dummy taking value 1 if the firm provided external training for its employees
Staff - skilled workers	(TZA)	Perc.	Percentage of skilled workers on staff
Staff - professional workers	(TZA)	Perc.	Percentage of professional workers on staff
Unskilled workforce	(IND)	Perc.	Percentage of unskilled workforce on staff
Staff with university education	(TUR, SA)	Perc.	Percentage of staff with at least one year of university education
Staff-part time workers	(TUR)	Perc.	Percentage of part time workers on staff
Staff - management	(SA)	Perc.	Percentage of management on the staff
Staff - non-production workers	(SA)	Perc.	Percentage of non-production workers in staff
Manager's experience	(SA)	Log	Manager's experience in years
Dummy for closed plant	(SA)	0 or 1	Dummy taking value 1 if the firm closed a plant during the year previous to the survey
Dummy for joint venture	(TZA)	0 or 1	Dummy taking value 1 if the firm agreed to do a joint venture during the last year

V Other control variables			
IC variables	**Country**	**Measurement units**	**Definition**
Dummy for incorporated company	(IND, TUR, TZA)	0 or 1	Dummy taking value 1 if the firm is constituted as an incorporated company
Age	(IND, TUR, SA)	Log	Age of the firm in years
Share of exports	(IND)	Perc.	Percentage of total annual sales exported
Trade union	(IND)	Perc.	Percentage of workers that belong to a trade union
Strikes	(IND, TUR)	Log	Days of production lost due to strikes
Market share	(TUR, SA)	Perc.	Share of market share
Dummy for recently privatized firm	(TUR)	0 or 1	Dummy taking value 1 if the firm was privatized within the last five years
Dummy for competition against imported products	(TUR)	0 or 1	Dummy taking value 1 if the firm competes against imported products
Capacity utilization	(SA)	Perc.	Percentage of total capacity used by the firm the last year
Dummy for FDI	(TZA)	0 or 1	Dummy taking value 1 if the firm received FDI inflows
Dummy for industrial zone	(TZA)	0 or 1	Dummy taking value 1 if the firm is located in an industrial zone