"Immigration and productivity: a Spanish tale"

Complementary material:

A) **Data organization**

B) **Labor productivity: computation of the representative value**

C) **Total factor productivity: preliminary test for the statistical robustness of the database**

D) **Estimations for industry and service sectors in the subsection 4.2 of the main text**

A) Data organization

The original database material exploited in this study was obtained by the (partial) matching of the information provided by the SABI database and MCVL (Social security records data, CDF version). The SABI database provides information on firm activity, while the MCVL provides data about individual workers. Our sample covers the period 2004-2010 for the SABI and 2005-2010 for the MCVL.

The SABI database compiles information at a firm level by gathering data extracted from the individual firms' balance sheets (according to the availability of information disclosed by Registro Mercantil). It includes all types of firms, excluding self-employed persons. This sample is not random or stratified but its size makes it a reliable reference for economic studies at a national level. In particular, despite other available databases (the EESE,[1] for instance), SABI does not present any structural distortion with regard to large firms. Therefore, it is more representative of the Spanish entrepreneurial environment.

In selecting information from the SABI database for any one year, we can include in our sample all firms with positive sales and with at least two employees.[2] The selection was made from more than 340,000 firms per-year, between 2004 to 2010. The Spanish Social Security records (Muestra Continua de Vidas Laborales, or MCVL) provided all the relevant information about persons with an active position in the Spanish social security system in any given year. We focused on the period spanning from 2005 to 2010. This dataset is a representative sample of the whole population. The sample from each year was randomly chosen and referred to more than one million persons (the MCVL accounts for 4% of the whole population). This database is composed by different blocks of data referring to different groups of variables (social, demographic or economics) that can be easily matched on the base of the identification code of each individual (that is a numerical string). In this way it is possible to get a quite complete overview of the status of the characteristics of an individual and his or her environment in a specific moment in time. The advantage of this database is that it threws light on some specific

[1] Acronym for *Encuestra sobre estrategías empresariales*

[2] These two selection constraints have been introduced to uniform the SABI content with that provided by the Social Security Records (MCVL).

information about the quality of the active workers as well as some information about the firms hiring them. It is not as complete as an employment-employee database, but it is a good proxy.

In this study, we made a preliminary selection of the individuals included in our database. Information included in the two databases does not overlap. There is no concrete element allowing direct association between these entries. Therefore, our strategy has been to create representative information at the sector level by province (for each year) in each individual database and, then, to merge that information.

In order to perform the fusion, we first elaborated an ad-hoc classification by sector to make the information in SABI (following the NACE-93 classification) comparable with the details included in the MCVL that follow the NACE-93 and NACE-2009 classification. In creating our own classification by sector we adopted a general strategy to create 23 sectors by identifying specific production activities. Each of these sectors was identified by collapsing the information from several sub-sectors, according to the criteria described in Table 1.A.

It is worth noting that sectors 100-800 refer to the industrial activity, while those from 1000 to 2300 refer to service activity. Once we had introduced this codification, we proceeded to adopt certain selection criteria in order to examine information in the same category as agents in the two databases (and therefore produce a proper matching situation).

The social security records represent a sample of the total population of working age, gathering together information about all employees that were hired (for the first time) in the period 2005-2010, as well as any relevant information about (active) persons who experienced an important change in their professional status during this period (retirement, firing etc..).

The firms included in the MCVL database are those with at least two employees (otherwise they would not have a social security account and we could not track them) and that were active during the period 2005-2010. In this respect, in the SABI database we only extracted information about firms with at least two employees (and reduced our sample considerably since most of the Spanish firms belong to the category of individual firms) and that had been active in each single year. This selection makes our final sample was different from a standard balanced panel when one experienced the entry-exit movement of firms. Furthermore, in the case of the SABI data, the firm distribution (by sector) was particularly skewed. In order to be able to provide

representatives measures of relevant indicators of firm activities (by sector) at province level,[3] we computed the median value of each indicator by following the same strategy we adopted for labor productivity as described in Section B. Once we had obtained these representative measures, we ranked them by province and, for each province, by sector.

Instead, in the case of the MCVL database, first we target to select a representative sample of all people that had entered and were active in the labor market in each single year. The distinguishing feature of the MCVL data is that it allows for the disclosure of information about the personal, economic and social environment of each new hiring as well as some general information about the position the person will be expected to fill, with some hints about the correspondent economic treatment.

In this respect, after recoding the ID of the sector of the activity of the hiring firm (according to the criteria presented in Table 1.A), we can cluster all new hirings into three principal groups according to the education degree associated with the filled positions (Higher degree, Middle degree and Elementary degree). After applying these two selection criteria, we are able to compute the representative indicators of the principal features of the new hirings according to our (general) province-sector-education group classification.

[3] We are referring to profits, sales, gross capital formation and wages among others.

TABLE 1.A: Structure of codification by sector

SABI: / MCVL:	CNAE-93 or CNAE2009 CNAE-93 (2005-2008; 2010)	CNAE2009 (2009)	
	SABI Y MCVL (2005-2008; 2010)	SABI y MCVL 2009	Description
OUR CODE	CNAE-93	CNAE 2009	
100	15+16	10+11+12	Food, beverages, and tobacco
200	17+18+19+20+36	13+14+15+16+31+32	Textiles, leather, and wood
300	21+22	17+18+58	Paper and edition
400	23+24+25+26	19+20+21+22+23	Chemical, plastic, and petroleum transformation
500	27+28+29	24+25+28	Metallurgy and mechanical equipment manufacture
600	30+31+32+33	26+27	Electrical machinery, computer system, and medical instruments
700	34+35	29+30	Automotive
800	40+41	35+36+37+38+39	Energy
900	45	41+42+43	Construction
1000	50+51+52	45+46+47	Car retail services
1100	55	55+56	Hotel
1200	60+61+62+63	49+50+51+52	Transport
1300	64	53+61	Telecommunication
1400	65+66+67	64+65+66	Financial activities
1500	70	68	Real estate activities
1600	71	77	Renting
1700	72	62+63+33	Information technology and computer services
1800	73	72	R+D
1900	74	69+70+71+74+78+82	Administrative and support services
2000	75	84	Public Administration
2100	80	85	Education
2200	85+90+91+92+93	86+87+88+90+91+92+93+94+95+96	Leisure activities
2300	95	97.98	Housekeeping services

Once we had completed the organization of the information included in the previous database according to the previously mentioned criteria, we could pass on to the last step of our exercise before the merger by focusing on the time dimension.

The information we organized in the SABI database is information at the year level. For each year, for instance, we controlled for a representative measure of the firm profits or the firm investment. The information disclosed by the MCVL database unveils the changes in the work force composition from one year to another. Then, before matching the two databases we needed to guarantee consistency in the type of information we were combining. To this end, we computed the variations of the selected variables extracted from the SABI database (namely the difference in value of the levels of our variables) in order to be able to associate them with the type of variables extracted from the MCVL. For instance, the difference of values of the indicators taken from the SABI database between 2004 and 2005 (as for labor productivity) are matched with the value of (representative) values of our selected variables in the MCVL in 2005 that already has

been defined not in level but as a change with respect 2004. The same applied for the remaining years

After this manipulation, our final database included 1150 yearly entries and we could expect to dispose of about 6900 entries for the overall period. Unfortunately, the dataset suffers from several missing values (for the different selected indicators) and finally we work on average with 3600 observation as full sample and something more than 3070 observations when we run estimations with lagged covariates.

B) Labor productivity: determination of the representative value.

Our data source is the SABI database. We compute labor productivity by calculating the ratio between the value of sales (at 2011 constant prices) and the total employment, and then we need to define a representative measure for the labor productivity of all firms belonging to a sector, in a province in each year. To this end (and similarly for all statistics associated with the variables we are extracting from the SABI database) we need to control for outliers.

In order to perform this exercise, we applied the box-plot transformation criterion. Using our information by sector at the province level, we ranked all the firm-data for each province and sector for each year. We then computed the interquintile ranges –IQR- (i.e. the difference between the third and first quartile). We then multiplied the value of the IQR by 1.5 and we excluded (for the median computation) all values outside this interval.

[*First quartile* – 1.5 IQR; *Third quartile* + 1.5 IQR]

Our final representative value of labor productivity (by sector, province and year) is the median of each series. The same applied for all the other variables we extracted from SABI and we compute as "representative value" of the firm level data.

C) Total factor productivity: preliminary test for the statistical robustness of the database.

Before start performing econometrics with our database, our concern has been to run some preliminary tests in order to be able to assess the goodness of the quality of the data we obtained by our statistical process.

The most natural test we were able to run has been the computation of the total factor productivity TFP with the purpose to test our results with the ones produced by the paper by Dolado et al. (2012) that exploit quite similar database to compute TFP [4] for a sample of manufacturing firms. As Dolado and al. (2012) we compute the TFP by following the empirical strategy discussed in Syverson (2011). According to this strategy the TFP is computed as a residual estimation of a firm level production function by following Levinshon and Petrin's (2003) suggestions for including materials to take into account part of the unobserved productivity (at the firm level). Thus, we did not dispose precise information about capital stock at the firm level. We overcame this limitation (as in Dolado et al. (2012)) by estimating TFP by including firm fixed effects. In order to control for the heterogeneity at sector level, we split the firms belonging to our sample into three main sectors: industry, services and construction. We tracked the evolution of the TFP (level) across time and we plotted our result in the following figure.

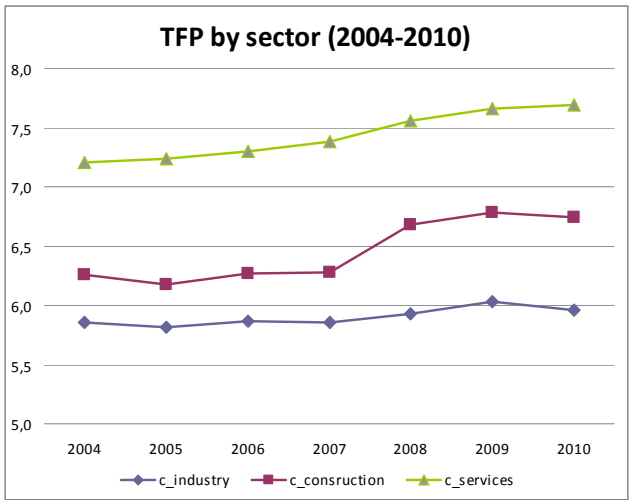

Figure 1.A: TPF estimation

(Data source: SABI; Calculus: authors)

These results were in line with the TFP values obtained by Dolado et al. (2012). They showed an average TFP growth equal to 0.4 for the selected sample of the manufacturing Spanish firm up to 2005. In our case, the TFP growth in industry (for the same period) was around 0.32. Therefore,

[4] Dolado, J. J, Ortiguera, S. and Stucchi, R. (2012): " *Does Dual Employment Protection Affect TFP? Evidence from Spanish manufacturing firms*", CEPR Discussion paper n. 8763.

the content of our database is reliable and our statistical strategy does not entailed at least visible and relevant distortions in the sample.

D) Estimations for industry and service sectors in the subsection 4.2 of the main text

Estimations results for industry and service sectors (Table 2.A and 3.A) are less informative than those for the overall sample, but they confirm the comments discussed for the overall sample. However, it is worth noting that the positive discrimination effect towards EU15-high-skill workers is extremely important for the top productive service sectors.

Table 2.A: Estimation results: lagged covariates (Industry)

Dependent variable: Variation of productivity (per-employee); Time period: 2005-2010; Stand. Errors in brackets

	Fixed Effects (province)	**QR (0.25)**	**QR (0.5)**	**QR (0.75)**	**QR (0.95)**
Constant	**-9.88 (1.24)*** **	**-3.24 (1.13)*** **	1.53 (1.07)	**8.08 (1.63)*** **	**14.88 (2.50)*** **
Spainsjt	**0.0029 (0.001)** **	0.002 (0.003)	0.004 (0.002)	0.002 (0.003)	-0.001 (0.006)
Eusjt	-0.498 (0.047)	0.021 (0.06)	-0.004 (0.053)	-0.083 (0.07)	-0.144 (0.092)
R-Eusjt	-0.007 (0.009)	-0.0176 (0.012)	**-0.023 (0.010)** **	-0.016 (0.011)	**-0.031 (0.018)** *
Asiasjt	-0.006 (0.013)	-0.009 (0.02)	-0.001 (0.02)	0.03 (0.022)	0.045 (0.044)
Africasjt	-0.008 (0.007)	0.0006 (0.0107)	-0.003 (0.009)	**-0.018 (0.009)** *	**-0.037 (0.018)** **
Latinsjt	0.015 (0.010)	-0.0064 (0.119)	-0.002 (0.011)	0.013 (0.013)	0.029 (0.019)
N_Americsjt	-0.014 (0.065)	0.063 (0.090)	0.0469 (0.072)	-0.038 (0.07)	-0.077 (0.13)
High sjt	-0.009 (0.023)	**0.058 (0.017)*** **	0.010 (0.021)	-0.012 (0.03)	**-0.09 (0.052)** *
Medium sjt	0.0027 (0.008)	0.0024 (0.010)	-0.0056 (0.0082)	-0.006 (0.010)	0.010 (0.023)
Low sjt	-0.0007 (0.001)	-0.0013 (0.0015)	-0.0007 (0.001)	-0.0005 (0.002)	0.0004 (0.003)
High_Spainsjt	-3.08e-07(0.00001)	**-0.00004 (0.00002)** *	-7.18e-06 (0.00002)	9.68e-06 (0.00003)	0.00005 (0.00005)
High_EUsjt	0.0003 (0.0004)	-0.0007 (0.0008)	0.0002 (0.0008)	0.0002 (0.001)	0.0012 (0.002)
High_Latinsjt	-0.00005 (0.00006)	0.0002 (0.0001)	-2.07e-06 (0.0001)	-0.00004 (0.0001)	-0.0002 (0.0003)
Low_Spainsjt	-3.84e-08(3.77e-07)	7.60e-07 (6.09e-07)	-1.29e-08 (4.9e-07)	-2.10e-07 (6.06e-07)	-7.8e-07 (1.12e-06)
Low_EUsjt	-3.47e-06 (0.000015)	8.22 e-06 (0.00002)	-6.05 e-06 (0.00002)	0.00001 (0.00002)	1.21e-06 (0.00005)
Low_Latinsjt	8.51e-07 (2.05e-06)	-1.55e-06 (3.77e-06)	1.53e-06(3.09e-06)	-3.55e-08 (3.39e-06)	2.12e-06 (7.10e-06)
Ass_emplsjt	**3.41e-06(4.12e-07)*** **	2.4e-06(0.00001)	3.20e-06 (0.00001)	4.06e-06 (0.00002)	6.46e-06 (0.00003)
Cost_emplsjt	0.496 (0.371)	0.14 (0.185)	**0.3846 (0.159)** **	0.28 (0.21)	0.65 (0.50)
Time dummies	Yes	Yes	Yes	Yes	Yes
Sect. dummies	Yes	Yes	Yes	Yes	Yes
Errors	Clustered (by province)	Bootstrap	Bootstrap	Bootstrap	Bootstrap
σ_μ	2.53				
ρ	0.02				
R-squared	0.19				
Pseudo R-squared		0.21	0.21	0.20	0.29
Observ	1790	1790	1790	1790	1790

Significance level: [a]*** 1%; ** 5% ; * 10 %.

Table 3.A: Estimation results: lagged covariates (Services)

Dependent variable: Variation of productivity (per-employee); Time period: 2005-2010; Stand. Errors in brackets

	Fixed Effects (province)	**QR (0.25)**	**QR (0.5)**	**QR (0.75)**	**QR (0.95)**
Constant	**5.35 (1.47)*** **	0.674 (1.01)	5.342 (0.769)	**9.11 (088)*** **	**18.61 (2.29)*** **
Spainsjt	-0.004 (0.003)	0.0011 (0.0012)	-0.0003 (0.001)	-0.00038 (0.00135)	-0.0029 (0.002)
Eusjt	**0.102 (0.037)*** **	**0.046 (0.0217) ** **	0.0048 (0.018)	-0.015 (0.028)	-0.0026 (0.049)
R-Eusjt	-0.0089 (0.023)	-0.005 (0.013)	0.011 (0.017)	0.0136 (0.021)	**0.104 (0.038)*** **
Asiasjt	**0.020 (0.009)** **	-0.00036 (0.0116)	0.0161 (0.0098)	0.0125 (0.016)	0.043 (0.0267)
Africasjt	**0.035 (0.016)** **	-0.0011 (0.0095)	-0.013 (0.013)	-0.0115 (0.0129)	**-0.051 (0.019)** **
Latinsjt	**-0.019 (0.009)** **	-0.003 (0.006)	0.0008 (0.006)	0.0017 (0.0077)	-0.010 (0.012)
N_Americsjt	0.052 (0.097)	0.077 (0.074)	-0.067 (0.073)	-0.143 (0.089)	-0.239 (0.172)
High sjt	**-0.010 (0.0045)** **	**0.0087 (0.003)*** **	-0.001 (0.002)	**-0.0086 (0.003)*** **	**-0.036 (0.006)*** **
Medium sjt	0.003 (0.0018)	-0.0092 (0.0009)	-0.0002 (0.0007)	-0.00003 (0.0011)	0.0011 (0.0017)
Low sjt	0.00064 (0.0006)	**-0.00106 (0.0005)** **	-0.0005 (0.0004)	-0.00012 (0.0005)	0.00105 (0.0001)
High_Spainsjt	**2.53e-06 (8.47e-07)*** **	-1.32e-06 (1.25e-06)	1.05e-06 (1.0e-06)	2.02e-06 (1.32e-06)	**0.00001 (3.15e-06)*** **
High_EUsjt	0.00002 (0.00004)	-0.00007 (0.00005)	0.00002 (0.00003)	0.00007 (0.00005)	**0.00025 (0.0001)** **
High_Latinsjt	**-8.35e-06 (4.35e-06)**	6.29e-06 (7.19e-06)	-7.90e-06 (5.23e-06)	-0.00001 (7.40e-06)	**-0.00006 (0.00002)*** **
Low_Spainsjt	-4.35e-08 (1.73e-07)	8.49e-08 (1.5e-07)	1.03e-08 (1.22e-07)	-1.0e-08 (1.54e-07)	-1.33e-07 (3.36e-07)
Low_EUsjt	**-9.46e-06 (4.11e-06)** **	-2.47e-06 (5.29e-06)	1.20e-06 (3.55e-06)	1.81e-06 (5.69e-06)	7.28e-06 (0.00001)
Low_Latinsjt	8.59e-07 (9.37e-07)	6.78e-07 (7.97e-07)	1.67e-07 (7.72e-07)	-2.03e-07 (9.34e-07)	-8.01e-07 (1.97e-06)
Ass_emplsjt	-1.50e-06 (2.32e-06)	-8.15e-06 (0.00001)	-4.41e-06 (0.00001)	2.61e-07 (0.00002)	-0.00002 (0.00004)
Cost_emplsjt	**0.307 (0.146)** **	**0.381 (0.0889)*** **	**0.353 (0.118)*** **	**0.292 (0.157)** **	0.202 (0.300)
Time dummies	Yes	Yes	Yes	Yes	Yes
Sect. dummies	Yes	Yes	Yes	Yes	Yes
Errors	Clustered (by province)	Bootstrap	Bootstrap	Bootstrap	Bootstrap
σ_μ	3.007				
ρ	0.013				
R-squared	0.06				
Pseudo R-squared		0.19	0.12	0.10	0.28

Observ	2521	2521	2521	2521	2521

Significance level: ª** 1%; ** 5% ; * 10 %.