A Wage Based Measure of Regional Aggregate Human Capital

João Carlos Cerejeira da Silva*

European University Institute and NIPE - University of Minho

E-mail address: jccsilva@eeg.uminho.pt

April, 2004

JEL Classification: E24, J24, I21

Keywords: Human capital, wages, income based measures

Abstract

The role of the accumulation of human capital to per capita income growth has been sharply debated among economists and policy makers. One open question of this debate is how to measure human capital. The standard approach is to use the average years of education of the labour force or the school enrolment rates as proxies for the stock of human capital. However, formal schooling achievement does not fully capture all the human capital stock. In fact, other forms

*Financial support by the Portuguese Foundation for Science and Technology (grant POCTI/47624/2002) is gratefully acknowledge.

of human capital accumulation are unmeasured. Also, it is assumed that the productivity differentials among workers with different levels of schooling are proportional to their years of education. In order to solve these problems, we develop the Mulligan and Sala-i-Martin's measure of human capital based, on labour income. This measure has some nice properties: is consistent with variable elasticities of substitution across types of workers, and does not impose all workers with the same amount of education to have the same amount of skill. It is also allowed for changes in the relative productivities over time and across different economies.We compute the index at the firm level and, finally, and we compare the evolution of our index with the evolution of average years of education for the Portuguese regions, highlighting the shortcomings of the latter measure of human capital.

1 Introduction

The role of the accumulation of human capital to per capita income growth has been sharply debated among economists and policy makers. Recently, this debate has reemerged mainly because, while several theories of endogenous growth would point towards a positive effect of human capital on economic growth, empirical evidence on this issue has been mixed. Early empirical contributions (e.g. Mankiw, Romer and Weil, 1992) established a robust link between enrollment rates and per capita GDP. More recently, other authors questioned this conclusion (e.g. Benhabib and Spiegel, 1994; Pritchett, 1996) and argued that the role of human capital on GDP growth has been vastly over-stated: human capital explains a much smaller proportion of the variation in the income per capita, and this relationship is far to be simply linear or positive (Kalaitzidakis, Mamuneas, Savvides and Stengos, 2001).

One possible reason why this debate has not reached a clear conclusion is due to the measurement of human capital. Until now, it has not been very clear how to proxy human capital. The standard approach is to use the average years of education of the labor force or the school enrollment rates as proxies for the stock of human capital. However, if we define human capital as the embedding of productive resources in people, formal schooling achievement does not fully capture all the human capital stock. In fact, as pointed out by Topel (1999, p.2954) "... schooling is only one form

of human capital. Other forms of human capital accumulation - like on the job training, acquisition of knowledge outside of formal schooling, and learning by doing - are unmeasured." The measurement of these other forms of human capital has been a central issue in the micro-literature of labor economics, but surprisingly only recently the macro-literature has turned to the micro-literature for help. As an example of this, Temple (1999), in a survey of the growth literature discusses the success of micro studies in finding a positive effect of schooling on wages and points out that "... the failure to discern this effect at the macro level is worrying."

There is a variety of reasons that explain why the average years of schooling is not necessarily a good measure of human capital. The first is related with the fact that there is no reason to believe that individuals with the same educational level must have the same productivity, even if the physical capital available is the same for everybody. Differences in accumulated experience related with the on-the-job training is one possible explanation, for differences in productivity across workers with the same educational level. Other possible reason is related with the fact that the economic relevance of what is taught in school may be not constant across different subjects. On the other hand, the individual's stock of human capital can decrease as his knowledge becomes irrelevant, out of dated or forgotten. The second reason is the assumption that the productivity differentials among workers with different levels of schooling are proportional to their years of educa-

tion. For example, it is assumed that a worker with 12 years of schooling is 12 times as productive as a worker with one year of schooling , regardless their wage rate differentials. Another reason is the assumption that workers of each education category are perfect substitutes for workers of all other categories.

Mulligan and Sala-i-Martin (1997) attempted to solve these problems by constructing a measure of human capital based on labor income. They argued that the individual's human capital is related with his wage rate, and therefore we can expect that people with more productive skills will earn more than the ones with low (productive) human capital. The main problem of an approach of this type is how to eliminate the effect of the other aggregate inputs (e.g. physical capital) from the worker's wage. To solve this problem, they divided each person's wage rate by the wage rate of the zero-skill worker. They called this measure as *labor-income-based* (LIB) measure of human capital. This measure has some nice properties: is consistent with variable elasticities of substitution across types of workers, and does not impose all workers with the same amount of education to have the same amount of skill. It is also allowed for changes in the relative productivities over time and across different economies.

Our paper applies this index to the Portuguese economy over four years (1989, 1992, 1995 and 1998). Our dataset enables us to use information concerning workers and firms and their location. Since physical capital is

combined with human capital at the firm level, the computation of the wage of the hypothetical individual without any skill, might be done at the firm level. We show that the index that use firm characteristics variables is preferred relative to the one used by Mulligan and Sala-i-Martin. The structure of the paper is the following: Section 2 presents the wage based measure of aggregate human capital, Section 3 describes the dataset and presents the main results, and finally, Section 4 concludes.

2 The Wage-Based Measure of the Aggregate Human Capital: Methodology and Discussion

2.1 The Wage-Based Measure of the Aggregate Human Capital

Both individual or aggregate human capital stocks are unobservable, because human capital includes all productive aspects embodied in people in a certain economy, for example: education and its productive relevance, on-the-job-training and the quality of the match between workers and firms. Since the labor force is heterogeneous, different workers contribute to production in different degrees. Hence, to aggregate different workers we need to give a larger weight to those workers that are more productive. In this line, the definition of aggregate human capital in a an economy should be equal to the quality-adjusted sum of the human capital of all its workers:

$$H_i(t) = \int_0^{\infty} \theta_i(t,s) N_i(t,s) ds \tag{1}$$

where $H_i(t)$ is the aggregate stock of human capital in i at time t, $N_i(t,s)$ denotes the number of people in economy i at time t with the skill s. The contribution of each worker to the aggregate human capital is $\theta_i(t,s)$. Dividing H_i by the stock of workers, we get the average stock of human capital in the economy i at time t:

$$h_i(t) = \int_0^{\infty} \theta_i(t,s) \eta_i(t,s) ds \tag{2}$$

where $\eta_i(t,s) = N_i(t,s)/N_i(t)$ is the share of economy i's labor force with skill s at time t and $h_i(t) = H_i(t)/N_i(t)$ is the stock of human capital per person.

Consider now that aggregate output of the economy i in time t,$Y_i(t)$, is determined by an aggregate production function that only depends on the total human capital $H_i(t)$ and total nonhuman capital $K_i(t)$:

$$Y_i(t) = F(K_i(t), H_i(t)). \tag{3}$$

Assuming that a worker's marginal product equals his wage, then the wage rate of a person in economy i with skill s in time t is given by:

$$w_i(t,s) = \partial Y_i(t)/\partial N_i(t,s) = [\partial F(K_i, H_i)/\partial H_i]\, \partial H_i/\partial N_i(t,s). \qquad (4)$$

Note that from (1), $\partial H_i/\partial N_i(t,s) = \theta_i(t,s)$, and denoting $\partial F(K_i, H_i)/\partial H_i = F_H$, then

$$w_i(t,s) = F_H * \theta_i(t,s). \qquad (5)$$

Normalizing the efficiency parameter $\theta_i(t,0) = 1$,then the wage rate of a worker with no skills is given by

$$w_i(t,0) = F_H * \theta_i(t,0) \equiv F_H. \qquad (6)$$

Dividing (5) by (6), we can infer the value of $\theta_i(t,s)$:

$$\theta_i(t,s) = w_i(t,s)/w_i(t,0). \tag{7}$$

By plugging (7) in (2) we get that the average stock of human capital can be measured as

$$h_i(t) = \int_0^\infty \left[w_i(t,s)/w_i(t,0)\right] \eta_i(t,s)ds = \int_0^\infty \left[w_i(t,s)\eta_i(t,s)ds\right] / w_i(t,0). \tag{8}$$

The expression inside the square brackets is the sum of all wages in the economy divided by the number of workers, or simply the average wage of economy i. Therefore this expression suggests a simple measure of the aggregate stock of human capital, which consists in the computation of the average labor income of each economy and then divide it by the wage of the zero-skill workers in that economy. While there is no particular problem on the computation of the average labor income (as far as the data is available) the computation and the meaning of the wage of the zero-skill worker requires some further discussion.

2.2 The wage of the zero-skill worker: discussion and estimation issues

We assume that any worker exposed to some economy-wide influences, such as schooling or labor experience, can have different productivities in different economies and in different time periods. However we need to define a numeraire in order to express the human capital index in a unit which is homogeneous across space and time. This numeraire will be the zero-skill worker, defined as the one with no schooling and no labor market experience or on the job training. This means that the zero-skill worker can only offer his physical effort combined with basic knowledge, which we assume that is equal for everybody, $\theta_i(t,0) = \theta(0) = 1$.

Nevertheless, the assumption of homogeneity of the zero-skill worker does not imply that this worker type will earn the same wage always and everywhere, because the available stocks of the other inputs as well as the level of technology will differ across economies. This is important because any productive shocks or differences in the schooling quality can be accommodated by this index. For example, an increase in the stock of capital will not change the ratio between the wages of the skilled worker and the zero-skilled worker, while differences on the economic relevance of schooling or on its quality will be reflected in the index.

Mulligan and Sala-i-Martin (1997) consider the case where productivity differences of the zero-skill worker occur only across geographical units or

time. Our data allows us to consider also sector and firm productivity differences. In fact the amount of capital available to each worker differs not only across regions, but mainly through differences in the firms endowments. Therefore the average stock of human capital in the economy i at time t, can be redefined as:

$$h_i(t) = \int_0^\infty \int_0^\infty \left[w_i(t,s,j)\eta_i(t,s,j)dsdj \right] / w_i(t,0,j). \tag{9}$$

where j is the index for firm. Now,

$$w_i(t,s,j) = F_H * \theta_i(t,s,j). \tag{10}$$

and

$$w_i(t,0,j) = F_H * \theta_i(t,0,j) \equiv F_H, \tag{11}$$

because $\theta_i(t,0,j) = 1$. Equation (11) expression implies that we need to quantify the wage of the zero-skilled worker for each firm, and use it as the numeraire for all co-workers within the same firm.

The first step in order to compute $h_i(t)$ is to estimate the wage of somebody with no skill $w_i(t,0,j)$. To do that we will use a Mincerian wage re-

gression, estimated from our wage dataset. This Mincerian regression has two main advantages over simply computing the sample mean of the wage of a zero-skilled worker. The first is that we can estimate that wage even if there are no workers with zero skill in a particular firm or region. The second advantage is that our estimate will be more precise, assuming that the Mincerian specification imposes the correct structure on the data, because we will use information from the full dataset, and therefore over the entire skill distribution.

3 Data Description and Results

The dataset used in this paper was constructed from the Quadros de Pessoal, of the Ministry of Labor and Solidarity (MTS). Beginning in 1982 and on a yearly basis, this Ministry has been collecting information on all companies operating in Portugal, except family businesses without wage-earning employees, through a mandatory questionnaire. Reported data match the firm, the establishment and each of the workers, and include the worker's gender, age, skill, occupation, schooling, tenure and earnings as well as the firm's location, industry employment level, sales volume and legal setting.

From the original dataset, we selected the observations on the following basis. First we dropped part-time workers as well as workers that did not work the normal period in the month of the survey (23% in 1989, 23% in 1992, 20% in 1995 and 22% in 1998). Recall that the information on social

security numbers is not validated because is not used for the production of official statistics and consequently there are some coding error and missing observations. Therefore, we dropped all observations without a valid identification number (7% in 1989, 4% in 1992, 3% in 1995 and 1998) and dropped individuals whose identification number appear twice or more, after keeping the full-time workers. This is a suspicion of a typo or a mistake when the data was introduced, but also could be the case that some individuals have more than one full time job. Note that if some workers have a full-time job and a part-time one, than the information related with the later job is deleted, while we maintained the former.

Then, we excluded all the observations for which one of the variables used in our analysis is missing or clearly wrong (examples of typos are changes in gender or changes in the date of birth). Then we retained only the workers in non agriculture or fishery firms, and located in the continental part of Portugal. Our final (unbalanced) panel has 4,768,187 observations over 2,616,233 different workers . Table 1 summarizes the average hourly wages as well as the (weighted) average county education.

Table 1: Information extracted from the original dataset in 1989, 1992, 1995 and 1998

Year	Final dataset		
	Nr. of obser.	Nom. hourly wages	Av. Education
1989	1,192,815	342.1	5.80
1992	1,302,952	572.0	6.22
1995	1,465,080	728.7	6.8
1998	1,593,149	881.8	7.38

The hourly wage was defined as the summation of all regular wage components divided by the normal labor time.

Earnings and labor time were measured in the month of March (in 1989 and 1992) and October (1995 and 1998).

Source: Portuguese Ministry of Labor and Solidarity, "Quadros de Pessoal" Dataset.

3.1 Wage Determination in the Portuguese Labor Market

Portugal is one of the OECD economies with the highest degrees of wage flexibility and responsiveness of wages to the macro unemployment rate (see OECD, 1992 or Modesto and Monteiro, 1993). The inequality pattern is close to that of the UK, and has been increasing over the last two decades (Cardoso, 1998). This increase of inequality is related mainly to a rise in the premium to higher education and in more complex jobs, while the premium related to tenure has been falling.

The intermediate nature of centralization in the Portuguese wage bargaining system does not allow any clear answer about wage adjustment at the micro level. In fact, some guidelines for wage increases are set at the central level by the government, unions and employers' associations. On

the other hand, it is possible to bargain at the firm or sectorial level due to the scattered nature of the union structure. This means that collective bargaining is extensively applied, setting minimum wage levels for different categories of workers. Therefore, the use of information about the firms' characteristics and worker's occupation is crucial in our subject.

Nevertheless, wage drift has been increasing in the Portuguese economy, especially for highly skilled and white-collar workers. According to Cardoso (2000), wage dispersion across firms is particularly pronounced for workers with high levels of schooling and for those with high tenure, while experience is valued in a more uniform way. This fact will give us an additional reason to estimate the wage of the zero-skill worker considering some firm's characteristics

In terms of the inequality observed at wage level, Portugal has an inequality pattern close to that of the UK, which has been increasing over the last two decades (Cardoso, 1998). This increase of inequality is related mainly to a rise in the premium to higher education and in more complex jobs, while the premium related to tenure has been falling.

The spatial wage dispersion has been less studied than the dispersion observed at sectorial or firm level. However, some authors (see e.g. Vieira, Hartog and Pereira (1997)) argue that earnings differ significantly across regions, even when other characteristics of the firms or workers are controlled for. Hence the wage of the zero-skilled worker must be also location specific,

in order to control to regional effects that can have impact on the worker's productivity.

3.2 The Variables of Interest

The Data Appendix gives us detailed information about all the variables. The wage variable that we used was the log of hourly earnings, where earnings were defined as the summation of all regular wage components. Earnings and labor time were measured in the months of March (in 1989 and 1992) and October (1995 and 1998). This variable is not deflated by the consumer price index because the constant of the wage regression can accommodate the effect of inflation on the wage of the non skilled workers.

The information about the education of the workers was given in levels, so we converted it to the correspondent years of schooling. From the workers file we extracted the variables gender, age, occupation and tenure. From the firms file we used sector (we set 23 different sectors), legal setting, equity capital share of foreign and private owners and employment level. The location of the worker was computed using the location of his establishment.

3.3 Computing the average stock of human capital

Our first procedure is to estimate the wage of the zero-skill worker. To do that we run an wage regression of the type:

$$w_{ijr} = \alpha_0 + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}_{ij}\boldsymbol{\delta} + \mathbf{Z}_j\boldsymbol{\lambda} + \sum_{r=1}^{18} \xi_r R_{jr} + \varepsilon_{ijr}, \tag{12}$$

where w_{ijr} is the log of the hourly wage of the individual i, who works in the firm j, in the region r. $\mathbf{X}_i$ is a vector of individual characteristics (gender, potential experience and years of schooling), $\mathbf{Q}_{ijxt}$ is a vector of variables related to job quality (such as tenure and occupation) and $\mathbf{Z}_j$ is a vector of the observable characteristics of the worker's firm j. $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are the corresponding vectors of the associated coefficients.R_{jr} are dummy variables that have a value of one if the establishment where the worker is employed is located in region r, and 0 otherwise and ε_{ijr}, is the error term, and α_0 the constant.

This regression was done for each year considered (1989, 1992, 1995 and 1998), and after retrieving the relevant coefficients, for each firm we computed the (log of the) wage of the zero skilled worker. This is equal to:

$$w_{ijr}(0) = \alpha_0 + \mathbf{Z}_j\boldsymbol{\lambda} + \sum_{r=1}^{18} \xi_r R_{jr},$$

which means that the (log of the) wage of the zero skill worker can be interpreted as the log wage of woman with no experience, no schooling and in the lowest occupation status (apprentice).

Then, we calculated the difference of the effective log hourly wage to the log wage of the zero-skilled worker, and we get the log of the parameter $\theta_i(t,s)$ (note that $\log\theta_i(t,s) = \log w_i(t,s) - \log w_i(t,0)$), which we denominated as the skill of the worker i. The average of the skills of the labor force gives us the average skill of the economy. Table 2 reports our estimates of the main variables for each year considered:

Table 2: Main Results (national averages)

	1989	1992	1995	1998
Log of Nominal Hourly Wage	5.66	6.13	6.38	6.57
Years of Education	5.82	6.22	6.80	7.38
Zero skill wage (log)	4.90	5.40	5.60	5.80
Skill (log)	0.75	0.73	0.77	0.78

This table shows some interesting results: while the average schooling of the worforce increased almost 27% from 1989 to 1998, the average skill of the labor force increased only 3% (from 0.75 to 0.78). This result can be explained by the fact that younger cohorts of workers are more likely to have more schooling that the older ones. However, young workers have less experience and specific human capital, than older workers. Therefore, as old workers retire, the economy loose their experience skills, but this lost will be compensated by the higher level of education of the newcomers. Note that formal education is mainly acquired before labor market entrance, then if the average education increases it is because the newcomers have more schooling

than more experienced workers. The coefficients of the regressions confirm this statement: a worker with 30 years of experience earn (on average) 35% more than a worker without experience, keeping all other characteristics constant. This coefficient is similar to the 34,4% of premium that a worker with 12 years of education earns, relative to the one without schooling.

The (average) wage of the zero schooling worker is closely correlated with the evolution of the nominal wages, since it is capturing inflation and the evolution of the capital stock, as well as reallocations of workers between firms, sectors and regions.

In order to check the relevance of the variables that capture firm characteristics, we did the same experiment, but excluding the variables in the vector **Z**. Denoting the resulting indexes by Skill(2) and Zero(2), and comparing with the former indexes, we have:

Table 3: Comparing results (national averages)

	1989	1992	1995	1998
Zero skill wage (log)	4.90	5.40	5.60	5.80
Zero(2) skill wage (log)	4.58	5.08	5.24	5.47
Skill (log)	0.75	0.73	0.77	0.78
Skill(2) (log)	1.08	1.05	1.13	1.11

This table shows that the index Skill(2) has much more variation than our preferred index Skill, which consider firm characteristics, which means that it is more sensible to business cycles effects, and have higher absolute

values. On the other hand, a simple statistic test on the joint significance of the coefficients related with firm characteristics, concludes that this variables are relevant and must be included in the wage regression. Therefore, the availability of data on firms, enables us to provide a more precise measure of the average human capital of the Portuguese economy, than the one used by Mulligan and Sala-i-Martin (1997).

4 Conclusion and Summary

There is a variety of reasons that explain why the average years of schooling is not necessarily a good measure of human capital. The first is related with the fact that there is no reason to believe that individuals with the same educational level must have the same productivity, even if the physical capital available is the same for everybody. Differences in accumulated experience related with the on-the-job training is one possible explanation, for differences in productivity across workers with the same educational level. Other possible reason is related with the fact that the economic relevance of what is taught in school may be not constant across different subjects.

Mulligan and Sala-i-Martin (1997) attempted to solve these problems by constructing a measure of human capital based on labor income, and on the assumption that people with more productive skills will earn more than the ones with low (productive) human capital. To solve the problem of how to eliminate the effect of the other aggregate inputs (e.g. physical capital)

from the worker's wage, they divided each person's wage rate by the wage rate of the zero-skill worker. They called this measure as *labor-income-based* (LIB) measure of human capital. This measure has some nice properties: is consistent with variable elasticities of substitution across types of workers, and does not impose all workers with the same amount of education to have the same amount of skill. It is also allowed for changes in the relative productivities over time and across different economies.

Our paper applies this index to the Portuguese economy over four years (1989, 1992, 1995 and 1998). Our dataset enables us to use information concerning workers and firms and their location. Since physical capital is combined with human capital at the firm level, the computation of the wage of the hypothetical individual without any skill, might be done at the firm level. We show that the index that use firm characteristics variables is preferred relative to the one used by Mulligan and Sala-i-Martin.

In spite off the advantages of this measure of the aggregate human capital some critics may be done. We had to assume that the unskilled worker was a perfect substitute for all others, although we allowed for any degree of substitutability among all the other types. If this assumption does not hold for some economies, than our measure will be biased. Further research, on this subject is needed for the Portuguese case, but Mulligan and Sala-i-Martin reported that this assumption is not too strong for the US case. However, it is important to remember that the use of average years of schooling is

more restrictive than this measure because assumes perfect substitutability between every different types. Empirical research comparing the performance of different human capital measures in growth equations can be an interesting way to expand this work.

5 References

References

[1] Benhabib, J. and M., Spiegel (1994), "The role of human capital in economic development: evidence from aggregate cross-country data", *Journal of Monetary Economics,* 34, 143-174.

[2] Cardoso, A.R. (1998), "Earnings inequality in Portugal: high and rising", *Review of Income and Wealth*, 44, 325-343.

[3] Cardoso, A.R. (2000), "Wage differentials across firms: an application of multilevel

[4] Lucas, R. (1988), "On the mechanics of economic development", *Journal of Monetary Economics*, 22, 3-42.

[5] Modesto, L. and Monteiro, M.L. (1993), "Wages productivity and efficiency: an empirical study for the Portuguese manufacturing sector", *Economia*, 17, 1-25.

[6] OECD (1992), *OECD Economic Surveys, Portugal, 1991/92*, Paris: OECD.

[7] OECD (1994), *The OECD Jobs Study: Evidence and Explanations*, Paris: OECD.

[8] Kalaitzidakis, P., T. Mamuneas, A. Savvides and T. Stengos (2001), "Measure of Human Capital and Nonlinearities in Economic Growth", *Journal of Economic Growth*, 6, 229-254.

[9] Mankiw, N., D. Romer and D. Weil (1992), "A Contribution to the empirics of economic growth", *Quarterly Journal of Economics* 108, 407-437.

[10] Mulligan, C. and Sala-i-Martin (1995), "A labor-income-based measure of the value of human capital: an application to the states of the United States", Working Paper No. 5018, National Bureau of Economic Research.

[11] Mulligan, C. and Sala-i-Martin (1997), "A labor income based measure of human capital", *Japan and the World Economy* 9, 159-191.

[12] Mulligan, C. and Sala-i-Martin (2000), "Measuring aggregate human capital", *Journal of Economic Growth* 5, 253-275.

[13] Pritchett, L. (1996), "Where has all the education gone?" Working Paper No. 1581, World Bank.

[14] Romer, P. (1986), "Increasing returns and long run growth", *Journal of Political Economy*, 94, 1002-1037.

[15] Romer, P. (1990), "Endogenous technological change", *Journal of Political Economy*, 98, S71-S102.

[16] Temple, J. (1999), "The new growth evidence", *Journal of Economic Literature* 37, 112-156.

[17] Topel, R. (1999), "Labor Markets and Economic Growth", *Handbook of Labor Economics*, Vol. 3, Ch. 44, 2943-2984.

[18] Vieira, J., Hartog, J. and Pereira, P. (1997), "A look at changes in the Portuguese wage structure and job level allocation during the 1980s and early 1990s", *Tinbergen Institute Working Paper* No. 97-008/3.

Data

The empirical work presented in this paper is based on the dataset "Quadros de Pessoal", of the Ministry of Labor and Solidarity (MTS). Beginning in 1982 and on a yearly basis, this Ministry has been collecting information on all companies operating in Portugal, except family businesses without wage-earning employees, through a mandatory questionnaire. This dataset covers, roughly, one half of all the active population. Table A1 reports the number of records for the years under consideration.

Table A1: Number of records in 1989, 1992, 1995 and 1998

Year	Workers	Firms	Establishments
1989	2 169 835	137 155	161 994
1992	2 268 151	159 192	185 777
1995	2 232 548	192 270	223 393
1998	2 430 691	213 589	248 664

The access to this dataset is conditional on the rules presented in the agreement between the University of Minho and the Department of Statistics of the MTS, and is possible under request.

The dataset is made up of three files:

(i) the workers' file, with data from 1985 to 1989 and from 1991 to 1998. This includes the worker's identification number (social security number), gender, age, skill, occupation, schooling, tenure, date of the last promotion, profession, earnings and number of working hours. These information is

relative to the month of March (from 1989 to 1993) or October (from 1994 to now).

(ii) the firms' file, with data since 1985. The main variables present in this file are: the firm's identification number, location (at county level), the establishment and firm's identification number, sector, legal setting, type of agreement between firm and unions, equity capital, share of national owners in the equity capital, share of foreign owners in the equity capital, share of public owner in the equity capital, yearly sales, number of establishments (since 1994), employment level (observed in March, between 1985 and 1993, and observed in the last week of October, since 1994) and date of the constitution (since 1995).

(iii) the establishments' file, with the firm's identification number and that of the one of the establishment (generated inside each firm), location, sector and number of employees.

5.1 Variables extracted and / or generated from the dataset

From the dataset, and after merging the three files, we extracted the following variables:

(i) Information about workers (subscript i denotes worker i):

- Log of the hourly wages: $\log hour_i = log \frac{\text{regular monthly earnings before taxes}}{\text{regular working hours}}_i$.

- Potential experience:

$$Potexp_i = \begin{cases} \text{(age - years of education - 5.75), if years of education >= 9} \\ \text{(age-14) if years of education <9} \end{cases}.$$

- Gender: variable $male_i = \begin{cases} \text{1 if male} \\ \text{0 if not} \end{cases}$.

- Education, dummies for 8 classes of different education levels and the respective correspondence with years of schooling:

Education Level of i	Competence	Correspondence with years of education
Educ_0	No reading or writing	0
Educ_2	Basic reading or writing	2
Educ_4	Primary school complete	4
Educ_6	Intermediate school	6
Educ_9	Lower high school	9
Educ_12	High school	12
Educ_15	College degree (3 years)	15
Educ_17	College degree (5 years)	17

- Tenure: $tenure_i$ = (date of the questionnaire - date of admission), converted to years.

- Generated the dummy variable $new_i = \begin{cases} \text{1 if } tenure < 1 \\ \text{0 otherwise} \end{cases}$.

- Occupation : 8 different levels (converted to dummies):

Occupation Level of i	Description
Quali_1	Executive and managerial
Quali_2	Intermediate managerial and executive
Quali_3	Low managerial
Quali_4	Technicians highly specialized
Quali_5	Sales, administrative and precision production
Quali_6	Administrative support, and production
Quali_7	Unskilled
Quali_8	Apprentice

(ii) Information about firms:

- Firm's legal setting:

Var.	Legal setting
$Legal\_1$	firm owned by the state
$Legal\_2$	private firm - individual owner
$Legal\_3$	private firm - collective owner
$Legal\_4$	cooperative
$Legal\_5$	non profit organization

- Sector (one dummy for each sector):

Sector	Description	Sector	Description
1	Agriculture and fishery (dropped)	13	Water, electricity and gas
2	Mining	14	Construction
3	Food, beverages and tobacco	15	Services concerning vehicles
4	Textiles	16	Wholesale
5	Leather	17	Retail
6	Wood products and cork (without furniture)	18	Hotels and restaurants
7	Paper and printing	19	Transportation services and communications
8	Petroleum refining, rubber, plastics and chemicals	20	Banking and insurance services
9	Other non-metallic mineral products	21	Other business and professional services
10	Iron and steel	22	Real estate
11	Metal products and machinery	23	Other services
12	Furniture and other manufacturing		

- Level of employment: *npessm* : employment level (observed in March, between 1985 and 1993, and observed in the last week of October, since 1994).

- *pkestr* share of foreign equity capital

- *pkstate* share of state equity capital

5.2 Observations extracted from the original dataset

From the original dataset, we selected the observations on the following basis. First we dropped part-time workers as well as workers that did not

work the normal period in the month of the survey (23% in 1989, 23% in 1992, 20% in 1995 and 22% in 1998). Recall that the information on social security numbers is not validated because is not used for the production of official statistics and consequently there are some coding error and missing observations. Therefore, we dropped all observations without a valid identification number (7% in 1989, 4% in 1992, 3% in 1995 and 1998) and dropped individuals whose identification number appear twice or more, after keeping the full-time workers. This is a suspicion of a typo or a mistake when the data was introduced, but also could be the case that some individuals have more than one full time job. Note that if some workers have a full-time job and a part-time one, than the information related with the later job is deleted, while we maintain the former.

Then, we excluded all the observations for which one of the variables used in our analysis is missing or clearly wrong (examples of typos are changes in gender or changes in the date of birth). Then we retained only the workers in non agriculture or fishery firms, and located in the continental part of Portugal. Our final (unbalanced) panel has 4,768,187 observations over 2,616,233 different workers.