Regional Wage Structure: A Panel Data Approach*

David Henriques†

Faculdade de Economia
Univ. Nova de Lisboa

February, 2009

Abstract

Spatial wage disparities can result from spatial differences in the skill composition of the workforce, in non-human endowments, and in local interactions, namely the distance to the economic centre. In this paper, the main equation to be estimated is the wage of a region relatively to the economic centre. Central to this equation is the idea that wages will be higher in those regions that have easy access to economic centers because for those regions demand linkages are relatively strong. I test this hypothesis using data on regional economic activities from the portuguese database *Quadros de Pessoal* (QP).

JEL Classification: C33, J31, R10.

Keywords: New Economic Geography (NEG), panel-data models and methods, spatial wage disparities.

* I thank Pedro Portugal for helpful comments and Ernesto Freitas for Stata technical support. I am grateful to DGEEP, Ministério do Trabalho e da Solidariedade Social in Portugal, for allowing the access and use of the Quadros de Pessoal database. The analyses, opinions and findings in this paper represent the views of the author, they are not necessarily those of the DGEEP. All errors are my own.

† E-mail: dthenriques@fe.unl.pt; Site: `docentes.fe.unl.pt/~dthenriques`.

1 Introduction

Agglomeration results from pecuniary externalities associated with increasing returns and transport costs: firms that locate in densely populated regions economize on fixed costs, by concentrating production in a single plant, and on transport costs, by locating near a large market. To the extent agglomeration creates congestion costs, firms in agglomerated regions must compensate workers by paying them high wages relative to outlying areas (Krugman and Livas, 1992). The theory has two predictions: *(i)* industry concentrates geographically, and *(ii)* relative wages decrease with transport costs from industrial centres. Using *QP*, a portuguese database, I will test this second prediction from the theory.

Regional variation in resource endowments and exogenous amenities are two sources of wage differentials that have been studied extensively in the literature (Roback, 1982, 1988; Beeson, 1991). Exogenous site-specific characteristics surely matter for wages, but only with low probability will they cause wages to decrease with distance from industry centres. Also, a large fraction of regional differences in labor efficiency doesn't stem from the presence of local externalities, but from the fact that some workers have a higher level of skill than others in other regions. See for instance, Combes, Duranton and Gobillon (2003), where the main findings suggest that individual skills account for a large fraction of existing spacial wage disparities with strong evidence of spatial sorting by skills.[1]

In this paper it is proposed that wages decrease monotonically as one moves away from industry centre. I examine the link between proximity to industry centres and regional wages by estimating Portugal's regional wage structure using data from 1996 to 2005. The variables which I deal with are: nominal wage distribution; the distance in minutes (by car) that separate each region in Portugal from the centre (Lisbon) and from Oporto (economic reference in north of Portugal) as proxy for the transportation costs. For the controls, I expect to use average skills (education), tenure, experience, gender, density (n^o workers/km^2) and industry dummy variables, which are fundamental vectors in the wage determination. Since there is empirical evidence for the small role of natural endowments and amenities on wage determination and, hardly all types of natural endowments can be quantified, I consider them as unobservable.

In section 3, I present two sets of estimation results. In the first set I apply cross-sectional methods to 1996 and 2005 data separately; in the second, I use panel data estimators in order to infer if there has been a drastic reduction of the effect of transport costs to the economic centre on regional wages from 1996 to 2005 in Portugal. Section 4 concludes.

[1] They estimate a model of wage determination across local labour markets using a very large panel of French workers. They find out that endowments only appear to play small role in the wage determination.

2 Theory and brief literature review

This section starts by presenting models used in the literature to estimate regional wage gaps. After a brief literature review it is important to mention other potential sources of wage differentials besides the distance to the economic centre, such as the *natural amenities* and *local public goods*. Also, in subsection 2.2, I will discuss the sort of problems that these other sources of wage differentials might raise with regard to econometric bias, if ignored.

2.1 Motivation and models to measure inter-regional wage differentials

In a simple model proposed in Hanson's (1997) paper, and here adapted to the portuguese case, the predictions for regional relative wages are summarized in the following equation:

$$\frac{w_r}{w_c} = F\left(x_{r,L}; x_{r,P}\right), \tag{1}$$

which holds,

$$w_c \geq w_i,\ i \neq c,\ \text{and}\ \frac{\partial\left(\frac{w_r}{w_c}\right)}{\partial x_{r,L}} < 0; \frac{\partial\left(\frac{w_r}{w_c}\right)}{\partial x_{r,P}} < 0, \tag{2}$$

where w_r is the nominal wage in region r, w_c is the nominal wage in the centre, $x_{r,L}$ is the unit transport from region r to Lisbon, $x_{r,P}$ is the unit transport from region r to Oporto. Since, $w_c \geq w_r$, $r \neq c$ and $w_r \geq 0$, $\forall r$ by definition wages are non-negative, thus $0 \leq \frac{w_r}{w_c} \leq 1$, which implies $F\left(x_{r,L}; x_{r,P}\right)$ to be a bounded function. The function $F(.)$ embodies preferences and technology.

By equations (1) and (2), I specify the following log-linear regression equation:

$$\ln\left(\frac{w_{r,t}}{w_{c,t}}\right) = \beta_0 + \beta_1 \ln\left(x_{(r,L),t}\right) + \beta_2 \ln\left(x_{(r,P),t}\right) + \epsilon_{r,t},$$

where r indexes geographic region, t indexes time, c is the industry centre, $\epsilon_{r,t}$, is an error term.

According to the New Economic Geography (NEG) theory one should expect the coefficients β_1 and β_2 to be negative, *i.e.* the higher the distance from region r to the centre, Lisbon, or Oporto, the lower will be the relative wage. Another prediction of NEG theory is that if there is a structural break in this relationship, such as a drastic reduction of the effect of transport costs to the economic centre, than the coefficients β_1 and β_2 will loose much of their importance in determining the wage ratio.

In Hanson, G. (1997) the estimation was performed for the mexican economy, where β_1 was the coefficient associated to the distance (in kilometers) from region r to Mexico city at time t, and β_2 was the coefficient associated to the distance (in kilometers) from region r to the US border nearest point. The final estimations were consistent with NEG theory since both betas were negative and statistically significant at the 1% level, $\hat{\beta}_1 = -0,143$ and $\hat{\beta}_2 = -0,151$.

In another paper, according to Vieira (2006), we can write the following equation to capture the inter-regional wage differential:

$$\ln w_i = \boldsymbol{\beta}' X_i + \boldsymbol{\theta}' Z_i + \epsilon_i,\ i = 1, 2, ...N. \tag{3}$$

where $\ln w_i$ denotes the natural logarithm of wage for worker i and X_i is a vector of explanatory variables which include a unit vector and controls for gender and human capital accumulation indicators such as years of education, years of labor marked experience and its square and years of tenure with the firm. It also includes controls for the logarithm of firm size, and eight industry dummies. The inclusion of these variables are justified by the fact that several authors have shown that firm size and industry affiliation play a role in explaining wage differences for apparently equally-skilled workers (see Krueger and Summers, 1988, Edin and Zatterberg, Arai 1994, Lausten, 1995, Idson and Feaster, 1990 and Oosterbeek and van Praag, 1995). In such a case, and to extent that industry location and firm size differ among regions, the effect of regions on wages would be biased in the absence of the inclusion of those variables. Finally, Z_i is a set of regional dummies; each of these dummies takes the value 1 if the individual i works in an establishment located in that specific region and 0 otherwise. For this purpose, were used the Portuguese districts (18 districts), each district with one dummy in the regression. After estimating the model we can get a map of region wage differentials just by looking at the dummy coefficients associated to each one of the regions. In order to evaluate the importance of regions in shaping the wage structure, the authors used conventional F-tests. The null hypothesis that regions play no role in explaining the wage structure (i.e., $\boldsymbol{\theta}' = \mathbf{0}$ in Vieira's approach) was rejected in all cases (for different years) at 1% level of significance.

2.2 Other sources of regional wage differentials

More variables that must be controlled for are known under the general heading of *natural amenities* and *local public goods*. Natural amenities are benefits ranging from a favorable climate, a coast-line location, the presence of lakes and mountains, to any natural endowments in raw materials. Amenities may also be the outcome of public policies, however, as for leisure facilities (theaters, swimming pool, etc.) or public services (schools, hospitals, etc.). Public goods are said to be local when their benefits are only reaped by local consumers, while the access cost of using these goods by more distant consumers are prohibitively high. Local public goods can also apply to firms. Transport infrastructures, research laboratories and job training centers are just a few examples. What happens when these amenities and local public goods are not included in the regressions? Omitting local public goods inflate the productivity of other production factors, such as labor and intermediate goods. If these local inputs were randomly distributed across space, their omission would be taken into account by the error term. Unfortunately, the supply of local public goods is the outcome of selected policies and often greater in areas characterized by concentrated economic activity, namely the economic centre. In this case, the

effect of distance to the economic centre will be overestimated, as the distance variable also captures the positive effect of these (omitted) local inputs. As shown by Roback (1982), dealing with natural amenities is slightly more involved. To see it, assume that a region is endowed with such amenities that attract migrants, all else being equal. The inflow of this new population exerts an upward pressure on the demand for housing, thereby pushing up rents. Such higher rents induce firms to substitute other production factors such as labor for land. As the marginal productivity of labor decreases, land-labor substitution leads to a drop in wages. When amenities are more abundant in heavily populated regions (as is the case for leisure facilities), the effect of distance is then underestimated. The key point is that omitted variables such as those ones can bias estimates in both directions, thus leaving us in the dark as to the magnitude and direction of each individual source of bias. In the spatial context, there is still another group of omitted variables. All the explanatory variables considered so far have been restricted to the geographic area r under consideration; none have taken into account effects, such as inter- or intra-industry externalities, that could emanate from neighboring areas. In other words, the implicit assumption so far has been a complete absence of interactions between neighboring areas. Everything is estimated under the presumption that no spillover effects exist between regions or that those are randomly distributed across regions. If distance has an impact on interregional interactions (via trade flows or knowledge transfers), such an assumption seems untenable. First of all, a market-potential variable could be introduced. Another approach consists of using techniques borrowed from *spatial econometrics*, by adding spatially lagged variables (which consists on a sum of spatial weights multiplied with values for observations at neighboring locations)[2], but bearing in mind the potential autocorrelation of the residuals. In both cases, the objective of such variables is to correct for an econometric bias, but they are often introduced in an *ad-hoc* manner (for instance, functional forms for distance-decay effects are chosen arbitrarily) and might be difficult to interpret. In a way, we find ourselves with the familiar quest of adding to our regressions a seemingly endless string of control variables. Namely, a panel of industries in different regions can be used, allowing for the introduction of both region and industry fixed effects. However, if such data are not available for all industries, will be impossible to study the magnitude of intra-industry externalities. In the same vein, we can evaluate the extent of inter-industry externalities. Given that data is available for a number of industries, industry fixed effects should be introduced. Indeed, they are necessary to capture differences in labor-shares across different industries, implying in turn that the intercept is now industry-specific. Moreover, we should even consider industry-time fixed effects to purge the model of business cycle effects that are common to all regions.

[2] There exists software that deals with this kind of issues such as GeoDa, where spacial weights and spatial lagged variables can be constructed.

3 Estimations and empirical results

3.1 Wage disparities across Portuguese employment areas

In this sub-section I document the extent and persistence over the 90's and till 2005 of wage disparities between employment regions in Portugal. I consider the Portuguese employment areas (districts) mean wage in 1991, 1996, 2000 and 2005 and compare it to the mean wage in Lisbon.

Table 1

Average nominal wages in euros in the Portuguese employment areas

Area (districts)	1991	$\frac{w_{district,91}}{w_{Lisbon,91}}$	1996	$\frac{w_{district,96}}{w_{Lisbon,96}}$	2000	$\frac{w_{district,00}}{w_{Lisbon,00}}$	2005	$\frac{w_{district,05}}{w_{Lisbon,05}}$
Aveiro	282,63	0,67	420,19	0,66	502,73	0,68	631,51	0,73
Beja	266,08	0,63	398,33	0,62	454,33	0,61	546,63	0,63
Braga	258,93	0,62	378,39	0,59	437,72	0,59	540,38	0,63
Bragança	238,70	0,57	342,72	0,54(Min)	413,14	0,56(Min)	506,42	0,59(Min)
Castelo Branco	254,03	0,61	368,16	0,58	422,47	0,57	538,83	0,62
Coimbra	279,72	0,67	421,62	0,66	479,25	0,65	591,37	0,69
Évora	262,88	0,63	389,15	0,61	454,9	0,62	577,52	0,67
Faro	298,64	0,71	425,42	0,67	493,95	0,67	594,15	0,69
Guarda	252,22	0,60	356,35	0,56	420,96	0,57	527,81	0,61
Leiria	294,12	0,70	436,28	0,68	502,78	0,68	608,58	0,71
Lisboa (centre)	419,29	1	638,99	1	739,78	1	863,29	1
Portalegre	271,75	0,65	417,11	0,65	480,32	0,65	582,70	0,68
Oporto	303,30	0,72	467,66	0,73	525,19	0,71	638,60	0,74
Santarém	275,355	0,66	419,08	0,66	495,96	0,67	598,39	0,69
Setúbal	304,71	0,73	479,33	0,75	540,66	0,73	665,96	0,77
Viana do Castelo	238,34	0,57(Min)	366,63	0,57	425,84	0,58	524,48	0,61
Vila Real	246,38	0,59	358,83	0,56	416,73	0,56	536,73	0,62
Viseu	252,32	0,60	371,41	0,58	434,61	0,59	540,65	0,63
Portugal	334,60	0,80	501,23	0,78	577,94	0,78	690,02	0,80

To be more systematic, the ratio of the lowest average to the highest wage average across the Portuguese employment areas remains between 0,54 and 0,59 over the considered years. On another way, the ratio of the highest average to the lowest across the Portuguese districts is between 1,71 and 1,87 during the analyzed years. Typically, the average wage in Lisbon is around 25 to 28% higher than the national average over the analyzed time period. It follows the descriptive statistics of the sample used in the research.

Table 2

Sample descriptive statistics

Part 1 - National variable Means

	1996	**2000**	**2005**
years of education	6,86	7,47	8,40
years of tenure	8,10	7,43	6,85
years of experience	23,12	23,17	23,49
% of male	0,61	0,59	0,57

Part 2 - % of workers in each industry

	1996	**2000**	**2005**
agriculture and fisheries	0,0174	0,0160	0,022
extractive and transforming industries	0,4016	0,3487	0,259
electricity and construction	0,1069	0,1147	0,132
wholesale and retail trade, lodging and restaurants	0,2460	0,2521	0,273
transport and communications	0,0672	0,0651	0,054
banking and insurance	0,0472	0,0383	0,027
real state and allied services	0,0540	0,0802	0,123
education, health and related services	0,0416	0,0602	0,076
other services	0,0180	0,0248	0,035

Part 3 - % of workers in each region

Regions	**1996**	**2000**	**2005**
Aveiro	0,081	0,077	0,0715
Beja	0,006	0,006	0,0087
Braga	0,099	0,094	0,0865
Bragança	0,003	0,004	0,0058
Castelo Branco	0,016	0,015	0,0013
Coimbra	0,030	0,030	0,03
Évora	0,009	0,011	0,0122
Faro	0,026	0,031	0,038
Guarda	0,009	0,009	0,01
Leiria	0,040	0,043	0,0466
Lisbon	0,355	0,350	0,3455
Portalegre	0,007	0,006	0,0076
Oporto	0,204	0,197	0,1837
Santarém	0,031	0,034	0,0362
Setúbal	0,045	0,046	0,05
Viana do Castelo	0,016	0,016	0,0179
Vila real	0,007	0,008	0,0115
Viseu	0,018	0,022	0,0251

Table 3

Some simple correlations

Mean local wage (in euros) of region r, w_r, in 1996 as a function of:

	(1) $\ln Emp_r$	(2) $\ln den_r$	(3) $schooling_r$	(4) $\ln d_r$	(5) d_r
Intercept	-40,23	348,25	-381,08	634,41	504,10
Coefficient	42,17 (t=4,68)	29,57 (t=3,82)	129,66 (t=6,88)	-46,43 (t=-8,67)	-0,54 (t=-4,19)
R^2	0,5780	0,4766	0,7472	0,8244	0,5227

Notation: The variable d_r measures the distance in minutes by car from the region r to the centre; Emp_r is the number of employees in region r; den_r is the population density in r, i.e., the population of region r over the total area of region r; $schooling_r$ is the average number of years of school in region r.

The log of mean local wage (in euros) of region r, $\ln w_r$, in 1996 as a function of the log of the distance (in minutes by car) to the centre: $\ln w_r = \beta_0 + \beta_1 \ln d_r + u_r$, where $\beta_0 = 6,47$ and $\beta_1 = -0,0961$, $t_{\beta_1} = -7,30$, $R^2 = 0,7692$.

According to table 3 (simple correlations), from column (3) the number of years of schooling has a big and significant effect over the mean local wages, in fact, one more year in school increases, on average, the wage in 130€. Summarizing the table from regression (1) to (3), I conclude that, denser, more populous and more educated employment areas seem to command on average a higher wage. From columns (4) and (5) and the $\log - \log$ regression I get very significant relations between wages in a region r and its distance to Lisbon (economic centre). To be more specific, on average, the workers located 1% far more from Lisbon earn less $46,43$€/month according to the estimation in (4) or, by (5) those who work 1 minute far more (by car) from the centre earn less 54 cents/month. According to the $\log - \log$ estimated equation, those who work 1% more far from Lisbon earn, on average, 9,6% less.

The goal for the rest of the paper is to assess the robustness of these basic results and uncover the determinants of spatial wage disparities.

3.2 Empirical specification

Wage differences across areas can reflect differences in individual skills or alternatively they can also reflect true productivity differences caused either by differences in non-human endowments or by local interactions. Hence, I propose the following micro-econometric specification,

$$w_{r,k,t} = B_{r,k,t}.s_{r,k,t} \tag{4}$$

where, $w_{r,k,t}$ is the the average wage in the region r, sector k, at time t. Average skills in region r, sector k, time t are captured by the last term, $s_{r,k,t}$, whereas the other two explanations (non-human endowments and local interactions) enter the term $B_{r,k,t}$ in equation (4).

Assume that the regional skills are given by

$$\ln s_{r,k,t} = \mu_0 + \mu_1 male_{r,k,t} + \mu_2 educ_{r,k,t} + \mu_3 \exp_{r,k,t} + \mu_4 \exp^2_{r,k,t} + \mu_5 tenure_{r,k,t} + \mu_6 den_{r,t} + u_{r,k,t}$$

and $B_{r,k,t}$, which reflects productivity differences from non-human endowments and local interactions in equation (4), I assume that is given by

$$\ln B_{r,t} = \alpha_0 + \alpha_1 \ln d_{(r,c),t} + \alpha_2 \ln d_{(r,Oporto),t} + v_{r,t}.$$

Putting all the parts together, the regression equation for the first specification (without fixed effects) is,

$$\ln (w_{r,k,t}) = \alpha_0 + \alpha_1 \ln d_{(r,c),t} + \alpha_2 \ln d_{(r,Oporto),t} + \ln s_{r,k,t} + \epsilon_{r,k,t}. \quad (5)$$

Intuitively we expect the coefficients of interest, α_1 and α_2, both to present negative values.

The measure of transport costs that has been used is the distance in minutes by car[3] (in 1996 and 2007 road structure) from a given *conselho* r to Lisbon (center) and to Oporto; *male* is the percentage of male workers, *educ* corresponds to the average number of years of school, exp is the average number of years of labor market experience which is computed as age minus the number year of education minus six and *tenure* is the average number of years of effective work. The variable $den_{r,t}$ corresponds to the regional density of employment in a given year t. The industry controls (presented in table 2, part 2) consist in 8 dummy variables, one for each industry except the reference group, "other services".

It is noteworthy that the choice of working with data at the regional level and, not at the city or individual level, is justified by computational reasons. With the least-squares dummy variable estimator, as N (number of observations) increases there are too many regressors to allow the inversion of the $(N + K) \times (N + K)$ regressor matrix, where K is the number of independent variables. While at the regional level there are only 17 dummies to compute, at the city (individual) level would be required to compute hundreds (millions) of dummy coefficients.

3.3 Estimation methods and issues

I allow idiosyncratic components, such as exogenous natural-resource supplies and amenities, to affect regional relative wages by taking a fixed-effects approach to estimating equation (5). Hence, I assume the error term $\epsilon_{r,k,t}$ has the following form

$$\epsilon_{r,k,t} = \beta_r + \lambda_k + v_t + \eta_{r,k,t}, \quad (6)$$

where, β_r is the area fixed effect; λ_k is the industry fixed effect, v_t is fixed effect for time t and $\eta_{r,k,t}$ is an i.d.d. term with zero mean and constant variance, σ. The fixed-effects assumption is

[3] Subsection "Data description: *Quadros de Pessoal* and Road network data" presents the methodology to compute the road travel time.

guaranteed by the fact that my sample includes all portuguese regions and economic activities.

If I had distance data (in minutes) only for one year or a time-invariant distance measure, then the fixed-effects estimation presents a problem. In that case in equation (5) the distance variable that measures transport costs would vary across regions but not across industries or years. First-differencing the data would eliminate the distance variables from the regression and using region dummy variables to capture fixed region effects would introduce perfect multicollinearity problems. For some estimators, notably the within and first differences estimators, only the coefficients of time-varying regressors are identified. Equations (5) and (6) constitutes my full econometric specification where the relevant parameters to estimate are α_1 and α_2, the ones related with the distance to Lisbon and Oporto.

3.3.1 Estimation methods with time-separated data (cross-section only)

In section 3.5.1, I estimate equation (5) by the usual OLS with industry and year dummies in the regression. The distance variables, in addition to capturing transport costs, will pick up any other regional effects that are correlated with distance. The question of interest is what portion of fixed region effects are associated with distance to industry centres. I perform a second regression (equation *(II)*, see Tables of section 3.5.1) in which I replace distance variables with regional dummy variables. The estimated region fixed effects are the mean effect of region characteristics on relative wages, controlling for year and industry. To determine the relative importance of distance among other region attributes, I regress the estimated region fixed effects on the distances (measured in minutes by car), given the portuguese road infrastructures at the time t. To the extent that the distance variables explain a large portion of the variance in the estimated region fixed effects, I am led to believe that distance is an important region characteristic for relative wages. I run this estimation methods twice, first to 1996-2000 data and second to the 2000-2005 data. In the end of 3.5.1, I compare the results and test for a structural change of the impact of the distance to the industry centre in the regional wages from 1996 to 2005.

3.3.2 Panel Data methods

In 3.5.2 I start by applying the Pooled OLS model to the equation in (5) but at the individual-level. The pooled OLS estimator is obtained by stacking the data over the N individuals (i) and T time periods (t) into one long regression with $N \times T$ observations, and estimating by OLS. In the statistics literature the model is called a population-averaged (PA) model, as there is no explicit model of $\ln w_{i,t}$ conditional on individual effects. Instead, any individual effects have implicitly been averaged out. The random effects model is a special case where the error term is equicorrelated over t for given i. The main complication for statistical inference, assuming no fixed effects, is that the distribution of least-squares estimators of this model varies with the

assumed distribution of the error. All the statistical inference is based on panel-robust standard errors. Like at the region-industry level, also at the individual level I find out robust statistical evidence of the impact of the distance to the centre over the wages (table 7).

Finally, I construct one last model where the cluster variable is the region-industry. All the controls presented in equation (5) still in equation (7). In order to control for inflation and other time effects I introduce one time dummy variable[4]. The model was estimated by several panel-data estimators, I report the results about the coefficients of interest in table 8. One variant of the model (equation 7) treats $\alpha_{(r,k)}$ as an unobserved random variable that is potentially correlated with the observed regressors. This variant is the called the fixed effects (FE) model as early treatments modeled these effects as parameters $\alpha_{1,1}, ...\alpha_{1,K}, \alpha_{2,1}, ...\alpha_{N,K}$ to be estimated. If fixed effects are present and correlated with regressors then many estimators such as pooled OLS are inconsistent. Instead, alternative estimation methods that eliminate the $\alpha_{(r,k)}$ are needed to ensure consistent estimation of β_1 and β_2 (the distance coefficients).

The other variant of the model assumes that the unobservable individual effects are random variables that are distributed independently of the regressors. This model is called the random effects (RE) model, which usually makes the additional assumption that both the random effects and the error term are i.i.d.

In the end I perform Hausman tests in order to conclude which of the estimators is the most appropriate.

3.3.3 Missing variables and endogeneity issues

The last key estimation regards missing variables. I distinguish two types of missing variables: those related to local (productive) endowments and those related to consumption amenities. One can think about airports, high-speed train lines, a favorable climate, closeness to a navigable river or a deep-sea harbour, etc. Gathering data for a complete list of endowments is a task much beyond the scope of this paper. Moreover, using a complete set of endowments would raise serious endogeneity concerns. For instance train stations or airports are likely to be endogenous. In absence of a complete set of endowments variables, I expect much of the effects of endowments to be captured by the error term. We could treat missing endowments as a random effect entering the residual. However, as a source of higher wages, these endowments are also likely to attract more workers in the area. In this case, failure to control for endowments will bias upwards the estimate of the effect of the explanatory variables on wages.

To deal with this omitted variables problem, I consider the fixed effects approach. Some of these endowments such as climate have a structural (or permanent) nature. To control for any of them, we can estimate equation (5) with time-invariant local fixed-effects. Finally, for some endowments, if I had access to that kind of data I could be using some direct controls,

[4] The introduction of a time dummy variable is important since I'm using nominal data, (not deflated). In this part of the paper, I only consider the years, 1996 and 2005, those which I have information about the distance variables.

for instance, a set of endowment in each region with the following attributes: seashore, lakes, mountains and cultural or architectural heritage.

Turning to amenities, the price of non-labor inputs matter in the determination of local wages. Then it is worth noting that the price of some non-labor factor (such as land) may not be solely determined by variables playing a direct role in the production function. As highlighted first by Roback (1982), better consumption amenities (*i.e.*, amenities unrelated to production) increase the willingness of consumers to pay for land and thus imply higher local land rents. As a result, firms use less land. In turn, this lowers the marginal product of labor (and consequently the wages) when land and labor are imperfect substitutes in the production function. Put differently, wages may capitalize the effect of non-production variables. This is in itself not an issue for my purpose.

This missing variable problem would only imply more noisy estimates for the wage effects since observationally identical employment areas end up paying different wages. It becomes an issue when consumption amenities are positively correlated with a variable of interest because, as shown by Wheaton and Lewis (2002), this introduces a negative correlation between this variable and the residuals. Because of this, the estimated effect of the variable is potentially biased downwards. However, just as with productive endowments, we can use instrumental variables, time-invariant area fixed-effects and further controls to deal with this missing amenity problem.

To summarize, we face both endogeneity and omitted variables problems that may bias the coefficients. Introducing some time-invariant area fixed-effects in (5), as I suggest in (6), will take care of permanent unobserved characteristics. Before going to the results, note that my estimations (*II*) in section 3.5.1 allow me to estimate not only the effect of a particular area on wages but also what percentage of such area fixed-effects is determined by the distance to economic centres, thanks to the second stage where I regress the region dummy coefficients over the distances to the economic centres.

3.4 Data description: *Quadros de Pessoal* and Road network data

Part of the data used in this paper have been drawn from the Portuguese *Quadros de Pessoal* (Personnel Records). This is a standardized questionnaire that all firms with wage earners must complete every year for the Department of Labor. The data include information on individual workers such as wage, age, tenure with the current firm, the highest completed level of education, and gender. Information is also available on hours of work, firm size, industry affiliation, and regions.

For the final sample were considered workers above 16 years of age and excluded individuals such as unpaid family workers and apprentices and all those who have a base wage below 150€ in 1996, below 250€ in 2000 and below 300€ in 2005. Individuals working in the islands of Madeira and Açores were not considered in the sample. The final sample considers the years of 1996,

2000 and 2005.

Regarding the measure of the distance from region r to Lisbon and Oporto were considered the road travel times computed in Holl (2004a, b) by the following method. "For year 1996, road travel times have been calculated on the basis of the 1996 Portuguese road network. The road network data has been compiled from road maps (ACP 1998/9; Michelin 1999) and detailed information from the *Portuguese Instituto de Estradas* on the historic development of all IP's (*Itinerários Principais*) and IC's (*Itinerários Complementares*). The 1996 Portuguese network is made up of 792 arcs and 616 nodes. These comprise all major roads (IP´s and IC´s), as well as the most important regional roads linking to municipality capitals (*sedes de concelhos*).

Each link or arc has associated tabular data on the length of the arc (in meters), the national road identifier, the link category and the year of inauguration in the case of IP´s and IC´s. There are four types of link categories: toll motorways, free motorways or dual carriageways, major trunk roads and other roads.

Travel times between locations are calculated using the computer programme ARC/INFO. Travel times between two locations consist of three parts:

1. Access time from each location to the nearest network node;
2. Minimum-path travel time on the network;
3. Egress time to the destination from the nearest network node.

Travel times were associates with links in relation to the link category. Travel times have been calculated for maximum speeds and for assumed average speeds. As maximum speeds 120 km/h for motorways and dual carriageways, 90 km/h for major trunk roads and 60 km/h for other roads have been taken. Assumed average speeds are 100 km/h for motorways and dual carriageways, 75 for major trunk roads and 60 km/h for other roads."[5]

As for year 2005, road travel times have been provided by Google - Map data 2008 Tele Atlas, according to the 2007 road network. By doing this it is implicitly assumed that the road network has not suffered significant changes from 2005 to 2007.[6]

Finally, all the computations and database manipulation have been done using Stata/MP 9.2 for Unix.

3.5 Estimation results

In this section I present two sets of estimation results. In the first set I apply cross-sectional methods to 1996 and 2005 data separately. Although I am restricting myself in the use of the data by not stacking all the observations together, I will be able to test the magnitude and significance of the coefficients of interest for 1996 and 2005 separately. Then, I infer if there has been a drastic reduction of the effect of transport costs to the economic centre on regional wages from 1996 to 2005.

[5] Road travel times in Portugal in 1996 and respective calculation method provided by Adelheid Holl (Institute of Public Goods and Policies at the Spanish National Research Council). See Table A in appendix.

[6] At the moment I was running regressions only data till 2005 was available in Quadros de Pessoal.

In the second block of results I use Panel Data estimators. I start by running Pooled OLS estimation over a large individual dataset, then I aggregate individual observations by region-industry and apply (besides POLS) the within, first-differences, between and the random effects estimators. I conclude this section running the Hausman test as a way of testing for the presence of fixed effects.[7]

3.5.1 Results with time-separated data (cross-section only)

Table 4 summarizes the results of estimation on equation (5), regressing $\ln(w_{r,k,t})$, where $w_{r,k,t}$ is the average wage of region r in industry k at time t, over different sets of independent variables.

Observations are by one-letter industry and region for the years 1996 and 2000. All regressions report t-statistics with robust standard errors. The results show strong support for the hypothesis that relative wages decline with distance from activity centres. Both distance variables are, as predicted, negative and statistically significant at the 1% level in all regressions. Moreover, the estimated quantitative effects of distance on relative wages are substantial. From column (*I a*), a 1% increase in distance from Lisbon leads to a 3,96% decrease in the region nominal wage, and a 1% increase in distance from Oporto leads to a 2,09% decrease in the region nominal wage.

Table 4

Regression results for region one-letter activities (A to Q), 1996-2000

(with robust standard errors)

Independent variables	*(I a)*	*(I b)*	*(I c)*	*(II)*
$\ln d_{r,c}$	-0.0396 (t=-7,21)	-0,0256	-0,0421 (t=-7,23)	-
$\ln d_{r,Oporto}$	-0.0209 (t=-3,40)	-0,0565	-0,0582 (t=-3,37)	-
$(\ln d_{r,c})^2$	-	-0,0025	-	-
$(\ln d_{r,Oporto})^2$	-	0,0063	0,0065 (t=2,21)	-
$Yr00$	0,0652 (t=5,32)	0,0793 (t=5,25)	0,0801 (t=5,24)	0,0846 (t=5,77)
R^2	0,8949	0,8984	0,8980	0,9175
N	324	324	324	324

$Yr00$ is a dummy variable indicating the year is 2000 (the base year is 1996). All regressions include dummy variables for the economic activity (activity sectors). Only column (*II*) includes region dummy variables, one for each region, except for Lisbon that is the reference group. For expositional ease, I do not report coefficient estimates on the constant term, the industry (activity), human skills/endowments or regional dummy variables. In regression (*I b*) the F-statistics over $\ln d_{r,c}$ and $(\ln d_{r,c})^2$ is 28,65 and over $\ln d_{r,Oporto}$ and $(\ln d_{r,Oporto})^2$ is 8,51.

[7] In the fixed effects model, only the within and first differences estimators are consistent. Estimators such as the pooled OLS, the between and the random effects will be all inconsistent.

To check the robustness of the regression results, I replace distance variables with region dummy variables and re-estimate equation (5). Column (*II*) of Table 4 shows the result. The R^2 increases from 89,49% to 91,75%, which suggests that there are other region-specific characteristics that matter for relative wages. To pursue the issue further, I regress the estimated region dummies on the distance variables (t-statistics below the estimated coefficients):

$$\begin{aligned} \beta_r^{OLS} &= \xi_0 + \xi_1 \ln d_{r,c} + \xi_2 \ln d_{r,Oporto} \\ &= 0,1497 - 0,0526 \ln d_{r,c} - 0,0197 \ln d_{r,Oporto} \\ &\qquad\qquad (t = \text{-}2{,}94) \qquad (t = \text{-}2{,}43) \\ R^2 &= 0,4613,\ N = 17 \end{aligned}$$

where β_r^{OLS} is the estimated region effect for region r. Distance explains **46,13%** of the variance in fixed region effects suggests that transport costs, as measured by distance in minutes, are an important characteristic of regions for relative wages.

As a further robustness check, I estimate the effects of distance on relative wages using observations on aggregate activity by region. The dependent variable is the average region wage relative to the Lisbon average wage. The time period is again 1996 and 2000. Table 5 reports the coefficient estimates. The results are generally consistent with those in Table 4. I again find strong evidence that relative wages are decreasing in distance to Lisbon. The associated coefficient with $\ln d_{r,c}$ is negative and statistically significant at the 1% level in all regressions. Results on distance to Oporto are somewhat weaker. Overall, I find strong evidence that regional nominal wages are positively correlated with proximity to industry centres. This accords with the descriptive analysis in section 3.1: the regions with the highest wages were those located near Lisbon or Oporto; the regions with the lowest wages were those proximate to neither the capital nor the Oporto market.

Table 5[8]

Regional regression results, 1996-2000

(with robust standard errors)

Independent variables	*(I)*	*(II)*
$\ln d_{r,c}$	-0,0811 (t=-3,29)	-
$\ln d_{r,Oporto}$	-0,0416 (t=-3.00)	-
$Yr00$	0,0294 (t=1,14)	0,0602 (t=2,12)
R^2	0,9432	0,999
N	36	36

[8] The distance coefficients in table 5 are higher (in absolute terms) and more significant than in table 4. Since in this regression I didn't consider the industry dummies the distance regressors are capturing a part of industry effects. This estimation suffers from an endogeneity problem which results on upward biased distance estimations.

$Yr00$ is a dummy variable indicating the year is 2000 (the base year is 1996). Only column (II) includes region dummy variables, one for each region, except for Lisbon that is the reference group. For expositional ease, I do not report coefficient estimates on the constant term, human skills/endowments and regional dummy variables.

$$\begin{aligned} \beta_r^{OLS} &= 0,2158 - 0,0703 \ln d_{r,c} - 0,0363 \ln d_{r,Oporto} \\ &\qquad\qquad (t = \text{-}3{,}41) \qquad (t = \text{-}3{,}87) \\ \bar{R}^2 &= 0,5547,\ N = 17 \end{aligned}$$

I run again the same regressions at the region-industry level substituting the data from 1996 by 2005 data. This allow me to test the other prediction of NEG theory, *i.e.*, if there is a structural break in this relationship, such as a drastic reduction of the effect of transport costs to the economic centre, than the coefficients associated with the distances to Lisbon and Oporto will loose much of their importance in determining the regional wage.

The estimation results with data from years 2000 and 2005 are the following,

Table 6

Regression results for region one-letter activities (A to Q), 2000-2005

(with robust standard errors)

Independent variables	*(I)*	*(II)*
$\ln d_{r,c}$	-0,034 ($t = -5,35$)	-
$\ln d_{r,Oporto}$	-0,018 ($t = -3$)	-
$Yr00$	-0,0712 ($t = -5,11$)	-0,0999 ($t = -5,57$)
R^2	0,9184	0,9381
N	324	324

$Yr00$ is a dummy variable indicating the year is 2000 (the base year is 2005). All regressions include dummy variables for the economic activity (activity sectors). Only column (II) includes region dummy variables, one for each region, except for Lisbon that is the reference group. For expositional ease, I do not report coefficient estimates on the constant term, on the industry (activity), on human skills/endowments or regional dummy variables.

$$\begin{aligned} \beta_r^{OLS} &= \xi_0 + \xi_1 \ln d_{r,c} + \xi_2 \ln d_{r,Oporto} \\ &= 0,2573 - 0,067 \ln d_{r,c} - 0,017 \ln d_{r,Oporto} \\ &\qquad\qquad (t = -3,51) \qquad (t = -2,25) \\ R^2 &= 0,5025,\ N = 17 \end{aligned}$$

In table 6 I still get negative and significant coefficients for the distance regressors. Furthermore, the distance explains **50,25%** of the variance in fixed region effects suggests that transport costs are an important characteristic of regions for relative wages. Hence, my first conclusion is that the negative relation among wages and distance from the industry centres is kept over the time interval considered (1996 - 2005). Comparing estimations (I) in table 4 and 6, both distance coefficients become closer to zero[9] and the t statistics less significant[10] as we go further in time. Performing the usual $t-$test for distance to Lisbon,

$$\begin{aligned} H_0 &: \alpha_{Lisbon,2005} = -0,0396 = \alpha_{Lisbon,1996} \\ H_1 &: \alpha_{Lisbon,2005} > \alpha_{Lisbon,1996} \end{aligned}$$

$t = \frac{-0.034-(-0.0396)}{0.0063551} = 0.881\,18$ and doing the same for Oporto I get $t = \frac{-0.018-(-0.0209)}{0.006} = 0.483\,33$. The nulls are rejected at significance levels above 19% and 32%, respectively. Thus, my second conclusion is that there's no clear evidence of a structural break in the relationship between distances and wages. The distance to the centre (on average) is not loosing power over time in the determination of regional wages.

This result can be interpreted in two ways. The first, assuming that the NEG theory is right then I conclude that there was not a drastic change in the national road structure from 1996 to 2005. The second possible interpretation, assuming that there was a drastic change in the portuguese road structure then the NEG theory does not apply to the portuguese case. Possibly the NEG theory would need a wider time range of study (several decades or a century) in order to be applied to the portuguese road network.

3.5.2 Panel Data results with 1996 and 2005 distance data

The Pooled OLS model for individual-level data. Among the linear panel models, the most restrictive model is a pooled model that specifies constant coefficients. Here I will be able to estimate the impact of the distance to the most important industry centres in Portugal over the individual wages. To complete the task I adapt equation (5) to individual data. The equation to estimate,

$$\begin{aligned} \ln w_{i,t} &= \alpha + \beta_1 \ln d_{(i,c),t} + \beta_2 \ln d_{(i,Oporto),t} + \kappa_1 male_i + \kappa_2 educ_{i,t} + \kappa_3 \exp_{i,t} + \kappa_4 \exp_{i,t}^2 + \\ &\quad + \kappa_5 tenure_{i,t} + \text{industry dummies} + \delta_{05} Yr05 + \eta_{(r,k),t} \end{aligned} \tag{7}$$

where $w_{i,t}$ denotes the wage of worker i at time t, all the independent variables take the usual meaning already explained, here adapted to individual observations.

If this model is correctly specified and regressors are uncorrelated with the error then it can be consistently estimated using the pooled OLS estimator.

[9] Lisbon coefficient goes from -0,0396 to -0,0344 while Oporto coef. passes from -0,0209 to -0,018.
[10] Lisbon t-stat goes from -7,21 to -5,35 while Oporto t-ratio passes from -3,4 to -3.

Since the error term is likely to be correlated over time for a given individual i I use robust standard errors to compute the t statistics, otherwise the standard errors would be downward biased inducing to too much statistical significance of the estimators. The usual OLS output treats each of the two years observations as independent pieces of information, but the information content is less than this given the positive error correlation. This leads to overstatement of estimator precision that can be very large. Moreover, the pooled OLS estimator is inconsistent if the fixed effects model is the true model.

Table 7

Individual regression results

(with robust standard errors)

Independent variables	
$\ln d_{i,c}$	-0,0266 (t=-135,30)
$\ln d_{i,Oporto}$	-0,0004 (t=-2,29)
$Yr05$	0,0294 (t=598,22)
R^2	0,5148
N	4.053.703

For expositional ease, only coefficient estimates of interest are reported.

The POLS, Within, First-difference, Between and the Random effects models. Here, I consider the following regression model

$$\begin{aligned}\ln w_{(r,k),t} &= \alpha_{(r,k)} + \kappa_1 male_{(r,k),t} + \kappa_2 educ_{(r,k),t} + \kappa_3 \exp_{(r,k),t} + \kappa_4 \exp^2_{(r,k),t} + \kappa_5 tenure_{(r,k),t} + \\ &+ \kappa_6 den_{(r,k),t} + \beta_1 \ln d_{(r,c),t} + \beta_2 \ln d_{(r,Oporto),t} + \delta_{05} Yr05 + \eta_{(r,k),t}\end{aligned}$$

where the region-industry specific effect is denoted by $\alpha_{(r,k)}$ that captures all unobserved, time-constant factors that affect $\ln w_{(r,k),t}$. All the variables have the usual meaning already explained in previous sections. The index (r,k) refers to the unit region r of sector k, and t denotes the time period. The panel was constructed with observations by region-industry for years 1996 and 2005. Considering 18 regions and 9 economic activities I got 162 observations per year, i.e., a panel with 324 observations.

Table 8 summarizes results from application of the standard panel estimators along with default and corrected estimates of the standard errors. Statistical inference should use either the panel-robust or the panel bootstrap standard error.

Table 8 - Part 1

Region-industry regression results: Standard Linear Panel Model Estimators (no time dummy)

	POLS-PA	Between	Within[d)]	RE-GLS	RE-MLE
β_1	**-0,0192**	**-0,0312**	**-0,0663**	**-0,01899**	**-0,0192**
Default se[a)]	0,0062	0,006	0,05771	0,00646	0,0062
Corrected[b)] se[a)]	0,0066	0,0073	0,0581	0,00596	0,0064
Corrected Reported statistics	$t = -2,9$	$t = -4,27$	$t = -1,14$	$t = -3,19$	$t = -3$
β_2	**-0,0151**	**-0,0153**	**-0,1046**	**-0,0152**	**-0,0151**
Default se[a)]	0,0055	0,0052	0,06852	0,0058	0,0056
Corrected[b)] se[a)]	0,007	0,0072	0,05619	0,0061	0,0062
Corrected Reported statistics	$t = -2,16$	$t = -2,125$	$t = -1,86$	$t = -2,49$	$t = -2,44$
R^2	NR[c)]	0,8774	0,7857	0,8922	NR[c)]
N	324	162	324	324	324

Notes: a) se, standard error; b) Corrected standard errors are panel robust se in Pooled OLS - Population-Averaged (POLS-PA), within, Random effects (RE) GLS and Population averaged estimation. The corrected se for between and RE-MLE estimation are bootstrap se. c) Stata did not report (NR).

Table 8 - Part 2

Region-industry regression results: Standard Linear Panel Model Estimators (with time dummy)

	POLS-PA	Between[d)]	Within	RE-GLS	RE-MLE
β_1	**-0,03836**	**-0,0312**	**-0,0076**	**-0,03854**	**-0,03836**
Default se[a)]	0,00574	0,006	0,04695	0,0059	0,00578
Corrected[b)] se[a)]	0,00641	0,0073	0,04451	0,00563	0,0061
Corrected Reported statistic	$t = -5,98$	$t = -4,27$	$t = -0,17$	$t = -6,85$	$t = -6,29$
β_2	**-0,0165**	**-0,0153**	**0,0188**	**-0,0165**	**-0,0165**
Default se[a)]	0,0051	0,0052	0,56831	0,00518	0,0050516
Corrected[b)] se[a)]	0,00611	0,0072	0,0551	0,0053	0,0086132
Corrected Reported statistics	$t = 2,7$	$t = -2,125$	$t = 0,34$	$t = -3,11$	$t = -1,92$
δ_{05}	0,1742	dropped	0,2253	0,175	0,1742
Corrected Reported statistics	$t = 10,30$	-	$t = 6,75$	$t = 10,81$	$t = 10,81$
R^2	NR[c)]	0,8774	0,8196	0,9213	NR[c)]
σ_α	0	-	0,1182	0,05935	0,05911
σ_η	-	-	0,05476	0,068	0,05633
λ	0	-	1	0,371	0,441
N	324	162	324	324	324

Notes: a) se, standard error; b) Corrected standard errors are panel robust se in Pooled OLS - Population-Averaged (POLS-PA), within, Random effects (RE) GLS and Population averaged estimation. The corrected se for between and RE-MLE estimation are bootstrap se. c) Stata did not report (NR). d) Time dummy variable was dropped.

Hausman tests - Within vs RE-GLS estimators

	χ^2 (8) (no time dummy)	χ^2 (9) (with time dummy)
H	-25,73[(*)]	30,98
H_{robust}	57,57	19,33

[(*)]Stata message: $\chi^2 < 0 \Rightarrow$ model fitted on these data fails to meet the asymptotic assumptions of the Hausman test.

The estimate of the distance parameters, β's, differ across the different estimation methods[11]. In part 1 of table 8, when there's no time dummy variable, the between estimate that uses only cross-section variation is lower than the Population Averaged estimate, however, when I introduce a time dummy (part 2 of table 8) in the regression this relation is inverted. In regression with no time dummy the within or fixed effects estimates of $(-0.0663; -0.1046)$ are much higher, in absolute terms, than the pooled OLS estimate of $(-0,0312; -0,0153)$, however, the FE estimates present small t statistics. When I introduce the time dummy the within estimation for the

[11]In this case and always when T=2, the within (fixed effects) and first difference (FD) estimates and all test statistics are identical, and so it does not matter which I use. However, when it is considered $T \geq 3$, the FE and FD estimators are not the same.

distance coefficients become close to zero, while the RE estimations become more negative and statistically significant.

The introduction of a time dummy is justified since the data is nominal and is important to purge the business cycle effects that are common to all regions. Hence, I belief that part 2 of table 8 gives the best estimation results.

The two RE estimates are very close to each other as here the estimates of the variances σ_α^2 and σ_η^2 are similar, leading to relatively similar values of λ.

Which estimates are preferred? The within and first-difference estimators are consistent under all models (pooled, RE and FE) whereas the other estimators are inconsistent under the fixed effects model. The most robust estimates are therefore the within or first-differences estimates. There is, however, an efficiency loss in using these more robust estimators, with standard errors that are much larger than those from pooled OLS and RE estimates.

If we can assume the $\alpha_{(r,k)}$ are uncorrelated will all regressors, then the RE-GLS method is appropriate, since it's the most efficient estimator. But if the $\alpha_{(r,k)}$ are correlated with some explanatory variables, the fixed effects method (or first differencing) is needed; if RE is used, then the estimators are generally inconsistent.

Comparing the FE and the RE estimates can be a test for whether there is correlation between the $\alpha_{(r,k)}$ and the regressors assuming that the idiosyncratic errors and explanatory variables are uncorrelated across both time periods. A formal **Hausman test** can be used to test whether or not the region-industry specific effects are fixed. A large value of the Hausman test statistic leads to rejection of the null hypothesis that the region–industry-specific effects are uncorrelated with regressors and to the conclusion that fixed effects are present.[12]

In general, to use the Hausman test, one has to perform the following steps: (1) obtain an estimator that is consistent whether or not the hypothesis is true; (2) obtain an estimator that is efficient (and consistent) under the hypothesis that I am testing, but inconsistent otherwise.

According to Cameron & Triverdi (2005) when $\tilde{\beta}_{RE}$ is fully efficient the Hausman test statistic simplifies to

$$H = \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right)' \left[\hat{V}\left[\hat{\beta}_{1,W}\right] - \hat{V}\left[\tilde{\beta}_{1,RE}\right]\right]^{-1} \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right),$$

where β_1 denotes the subcomponent of β corresponding to time-varying regressors since only that component can be estimated by the within estimator, $\hat{\beta}_{1,W}$. This test statistic is asymptotically $\chi^2\left(\dim\left[\beta_1\right]\right)$ distributed under the null hypothesis.

However, the simple form of the Hausman test is invalid if $\alpha_{(r,k)}$ or $\eta_{(r,k),t}$ are not i.i.d., which happens in the presence of heteroskedasticity in the data. Then the RE estimator is not fully

[12] Although, it may still be possible to avoid using a fixed effects model. If regressors are correlated with region-industry specific effects caused by omitted variables, then one can add further regressors, either time varying or time-invariant, and again perform a Hausman test in this larger model to see whether fixed effects are still necessary. Even if such correlation persists it may be possible to estimate a random effects model using instrumental variables methods.

efficient under the null hypothesis so the expression $\hat{V}\left[\hat{\beta}_{1,W}\right] - \hat{V}\left[\tilde{\beta}_{1,RE}\right]$ in the formula for H needs to be replaced by the more general $\hat{V}_{Boot}\left[\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right]$. Then a panel-robust Hausman test statistic is

$$H_{Robust} = \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right)' \left[\hat{V}_{Boot}\left[\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right]\right]^{-1} \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right).$$

This test statistic can be applied to subcomponents of β_1 and can use alternative estimators such as $\tilde{\beta}_{1,POLS}$ in place of $\tilde{\beta}_{1,RE}$ and $\hat{\beta}_{1,FD}$ in place of $\hat{\beta}_{1,W}$.[13]

Considering the robust H statistic with no time dummy, $H = 57,57 > \chi^2(8)_{0,01} = 20,09$ there's clear evidence of fixed effects. Hence, the within estimations are the appropriate.

However, when I consider the time effect dummy, the Hausman test does not reject the null hypothesis of random effects[14], despite the large difference between FE and RE estimates. So the more efficient random effects estimates could be used here, *i.e.* the RE-GLS[15]. Another advantage of random effects estimation, in general, is that it permits estimation of the coefficients of time-invariant estimators.

Inference should be based on panel-robust standard errors that permit errors to be correlated over time for a given region-activity and to have variances and covariances that differ across region-activity.

For brevity these estimates are called panel robust, though they are additionally robust to heteroskedasticity. The default se that is based on the assumption of i.i.d. errors. In this regression the correctly estimated standard errors are, in many cases, 10% to 20% larger as the default standard errors; sometimes the corrected se are below the default se. The between estimator is an estimator with standard errors that need only correction for heteroskedasticity since it uses only cross-section variation.

Default standard errors assume independence of model errors over t for given (r, k) when in practice they are likely to be positively correlated. This erroneous assumption overestimates the benefit of additional time periods, leading to downward bias in standard errors.

Additionally, ignoring heteroskedasticity in errors also leads to bias, though this bias could be in either direction. For the within and between estimators inclusion of the term $\alpha_{(r,k)}$ should control for some of the correlation in the error across time for a given individual. Clearly panel-robust standard errors should be used.

[13] Alternatively, Hausman (1978) proposes also a convenient regression format for the test. This auxiliary OLS regression is explained, for instance, in Cameron & Triverdi (2005).

[14] $H_{robust} = 19,33 < 21,67 = \chi^2(9)_{0,01}$

[15] Thus, admiting the model with the time dummy is the correct one, we can say that the regional wages decrease 3,89% or 1,65% as we get far 1% more from Lisbon or Oporto, respectively.

4 Conclusions

In this paper I provided evidence that in Portugal the distance to main economic centres matters for the regional wage determination: nominal wages are highest near the economic centres, Lisbon and Oporto. The importance of market access on relative wages highlights the role of trade policy in regional development.

Using portuguese data from years 1996, 2000 and 2005 on education, tenure, experience, gender, density, industries and distance I use cross-sectional methods showing a strong negative relation between distance to industry centres (Lisbon and Oporto) and wages. I performed the estimations at the individual, region-activity and regional levels. I have shown that the negative relation (common at all levels) is persistent and did not lose significance over the decade of study. The NEG theory predicts compression of wage disparities as transportation costs decrease and market access becomes easier, nonetheless I could not show evidence of this. In order to show evidence on wage compression would be required more data on distances to economic centres of previous years.

I provided a second set of estimation results based on panel data tools. I have estimated a model with five standard linear panel estimators and then selected the most appropriate according to the (efficiency and consistency criteria) Hausman test. The chosen estimator (RE-GLS) provided an estimation for the coefficients of interest close to estimations obtained in the cross-sectional methods.

According to the analysis, I conclude that workers who live 1% far more from Lisbon suffer a wage discount around 3,9%, while living 1% more far from Oporto the wage discount is around 1,7% to 2%, on average.

5 References

Beeson, P. (1991), Amenities and regional differences in returns to worker characteristics. Journal of Urban Economics, vol. 30 (September), pp. 224-41.

Beeson, P. and Eberts, R. (1989), Identifying productivity and amenity effects in interurban wage differentials. Review of Economics and Statistics, vol. 71 (August), pp. 443-52.

Cameron, A. Collin and Pravin K. Triverdi (2005) Microeconometrics: Methods and Applications, Cambridge University Press.

Combes, P-P; Duranton, G. and Laurent Gobillon (2003), Spatial wage disparities: Sorting matters!

Edin, P-A. and Zatterberg, J. (1992) "Interindustry wage dispersion: evidence from Sweden and a comparison with the United States", American Economic Review, 82, 1341-1349.

Hanson, G. (1997), Increasing returns, trade and the regional structure of wages. Economic Journal Vol. 107, pp. 113-133.

Holl, A. (2004a), Transport infrastructure, agglomeration economies, and firm birth. Empirical evidence from Portugal', Journal of Regional Science, 44 (4): 693-712.

Holl, A. (2004b), Start-Ups and Relocations: Manufacturing Plant Location in Portugal', Papers in Regional Science, 83 (4): 649-668.

Idson, T. and Feaster, D. (1990), "A selectivity model of employer wage differentials", Journal of Labor Economics, 8, 99-122.

Krueger, A. and Summers, L. (1988), "Efficiency wages and the inter-industry wage structure", *Econometrica*, 56, 259-293.

Krugman, P. and Livas, R. (1992), Trade policy and the third world metropolis. NBER Working Paper No. 4238.

Lausten, M. (1995) "Inter-industry wage differentials in Denmark", Center for Labour Market and Social Research, University of Aarhus and Aahrus School Business, working paper n^{o}18.

Oosterbeek, H. and van Praag, M. (1995), "Firm-size wage differentials in the Netherlands", Small Business Economics, 8, 173-182.

Roback, J. (1982), Wages, rents, and the quality of life. Journal of Political Economy, vol. 90 (December), pp. 1257-78.

Stata Release 9 Longitudinal/Panel Data, Reference Manual, Stata Press Publication, 2005.

Vieira J., Couto J., and Tiago M. (2006) Inter-regional wage dispersion in Portugal, Regional and Sectoral Economic Studies, Euro-American Association of Economics Development 6, 1.

Wheaton, William C. and Mark J. Lewis (2002), Urban wages and labor market agglomeration, Journal of Urban Economics, 51(3):542–562.

Wooldridge, Jeffrey M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

Wooldridge, Jeffrey M. (2003), Introductory Econometrics: A Modern Approach, 2e, Thomson.

6 Appendix

Table A - Distances in minutes from area r to Lisbon and Oporto, year 1996 and 2007

Area (districts)	to Lisbon, 1996	to Lisbon, 2007	to Oporto, 1996	to Oporto, 2007
Aveiro	168	150	51	55
Beja	136	125	314	267
Braga	223	213	39	46
Bragança	356	331	172	175
Castelo Branco	180	136	201	186
Coimbra	126	125	78	77
Évora	90	87	253	229
Faro	218	158	403	300
Guarda	253	188	155	153
Leiria	90	95	115	110
Lisboa (centre)	1	1	194	182
Portalegre	162	152	239	202
Oporto	194	183	1	1
Santarém	50	58	151	140
Setúbal	33	41	224	197
Viana do Castelo	251	220	67	51
Vila Real	254	240	70	73
Viseu	198	190	95	87
Portugal (mean)	166	150	157	141

Note: Road travel times in Portugal in 1996 provided by Adelheid Holl and in 2007 taken from Google - Map data 2008 Tele Atlas.