Central European Journal of Economic Modelling and Econometrics

Spatial Pseudo Panel Data Models with an Application to Mincer Wage Equations

Selahattin Güriş* and Gizem Kaya Aydın†

Submitted: 12.10.2020, Accepted: 6.08.2021

Abstract

The studies using Mincer equations are generally applied to cross-sectional data at the micro-level. There are however limited studies conducted with macro or panel data for wage equations. Pseudo panel data methods can be applied to empirical studies by creating cohorts from repeated cross-sectional data in the absence of genuine panel data. Difference in both the human and labour resources according to the spatial positions may also affect the prediction of the wage equations. We aim to introduce the application of spatial pseudo panel models by creating cohorts according to the birth years of employees and regions in which they live from the Turkish household labour survey for the period 2010–2015. As a result, we find that the spatial autocorrelation model is appropriate for wage equations of Turkey. We also find that return of education on wages is 11% while return of experience on wages is 4%.

Keywords: spatial econometrics, pseudo panel data, Mincer wage equations

JEL Classification: E24, C31, C33

*Marmara University, Department of Econometrics, Istanbul, Turkey;
e-mail: sguris@marmara.edu.tr; ORCID: 0000-0002-1017-1431

†Istanbul Technical University, Department of Management Engineering, Istanbul, Turkey;
e-mail: kayagizem@itu.edu.tr; ORCID: 0000-0002-6870-7219

1 Introduction

As it is known, the methods used in econometrics vary according to the data type. The data of different units in a single period is named cross-sectional data, the data of a single unit over different periods is named as time series, and the data which is combination of these two data types is called panel data. Finally, the data created by gathering different units together for different periods is called pooled data. Pooling data across cross-sections and time series improves the quality of data analysis; however, the model is limited in its ability to accurately predict variables of interest due to severe practical data limitations and the ability to properly capture varying market structures (Howie and Kleczyk, 2007). Since the panel data sets contain the same observation every year, it is more accurate to monitor the structural/characteristic changes in this data type. In the pooled panel version, the data is created by combining the cross-section data, there are different observations in each period. Therefore, it is difficult to track changes as the characteristics are not the same for each period.

Models with panel data, i.e. same observations on individual units (workers, households, firms, etc.) – also referred to as longitudinal data – show to have superior statistical properties compared with models of cross-sectional data, particularly when there is relevant but unobservable information on the units and when causal relationships are of interest. There has been huge interest in longitudinal data for several decades. However, panel data collection is often very expensive and data sources may open them to researchers only if a large fee is paid. There are also privacy concerns that may explain why longitudinal microdata is not always available. Provided that the terms of confidentiality are met, the researcher may see information about individuals through online access, but is barred from publicly releasing any information that can identify individuals.

Considering the related costs and the institutions required, large panel data sets are less commonly available in developing countries. In this case, Deaton (1985) advocated the pooling of a series of much more common, and cheaper, cross-sectional surveys that can be turned into a pseudo panel model by identifying personal characteristics that do no change over time. Averaging variables of interest across all individuals that share the same fixed personal characteristics year by year, i.e. by following cohorts, generates a pseudo panel for which there are efficient estimators. Similarly, in the absence of genuine panel data, it is possible to obtain pseudo panel data by creating cohorts from repeated cross-sectional data. These cohorts take replace of the cross-sections in genuine panel data and, these models are called "pseudo panel data models." In this wise, researchers have the opportunity to use pseudo panel models, rather than obtaining separate results for each year based on cross-sectional data. For this reason, pseudo panel data created from repeated cross-sectional data according to fixed characteristics provide more information than the data generated from the pooling of repetitive cross-sectional data in the absence of genuine panel

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

data. When the location is one of the variables for which groups are defined, a spatial pseudo panel model is created.
The contribution of this article to apply the spatial pseudo panel model estimation to the earnings function state pioneered by Jacob Mincer (1958). It is the first application of the spatial pseudo panel data models.
In the literature, the determinants of wages have been examined and improved by Mincer (1958 and 1974) and many other studies. However, these studies were usually conducted with micro-level cross-sectional data. On the other hand, we assume that the human and labour resources of the regions may differ according to the spatial conditions (see Ramos et al. 2015; Longhi et al. 2006). Thus, the inclusion of spatial effects in the regression model will give in more consistent or efficient (is depend on the most suitable spatial model) results in modelling the determinants of the wage.
As far as is known, the number of studies performed by using pseudo panel data (even in the genuine panel) and including spatial effects is quite few in the literature (except Baltagi et al. 2015). With this motivation, in this study, we mainly aimed to introduce spatial pseudo panel data models as an integrated model via Mincer wage equations. Thus, we created 108 cohorts that are based on 9 year-of-birth groups of employees and 12 regions by using the repeated cross-sectional microdata of household labour force surveys of the Turkish Statistical Institute (Turkstat) for the period 2010–2015. Spatial determinants of wages were also included by adding the spatial contiguity matrix into the equations. We then estimated the human capital Mincer equation for Turkey by using spatial pseudo panel models and we observed the differences in estimations by comparing it with the models with cross-sectional data for a specific year, pooled data, and pseudo panel data without spatial effects.
As a consequence, in the absence of macro-level or genuine panel data, adding the spatial relationships into the pseudo panel data model will lead to more consistent or efficient estimators and enable the implementation of more appropriate policies on a regional basis. This study aims to make a major contribution to the literature since it demonstrates the applicability of spatial pseudo panel data models empirically.

2 Literature review

The estimation of the determinants of the wage equations is first examined in detail by Mincer (1958 and 1974). According to Mincer (1981), the quality of human capital increases the production and the income at the country level. In his study, Mincer indicates that gender, race, age, and experience affect the wages of the labour force; in fact, after a certain age or years of experience, the effect of age or experience on wages decreases. Moreover, he reaches the conclusion that as the education level increases, the average wage increases.
The application of spatial models to labour force studies has long been present in the literature. For instance, Ramos et al. (2015) finds the spatial effects as statistically significant in the estimation of the wage equation for Spain. Using the static and

dynamic spatial panel data methods for the period 2000–2010, they find that the wage equation is highly autoregressive, and regional dissemination is appropriate to explain the relationship between unemployment and wages in Spanish states. Similarly, Longhi et al. (2006) examine the wage equations by utilizing spatial panel data methods for 327 regions in West Germany between 1990–1997. They do not reject the hypothesis indicating that "wages should be higher in regions that strongly interact with other regions." Elhorst et al. (2007) investigate the wage equations for 114 regions of Germany for the 1993–1999 period and they find the wage equation includes a spatial error autocorrelation model as statistically significant.

The number of studies conducted in the labour force literature by using the pseudo panel models has been increasing since the study of Deaton (1985). For instance, Russel and Fraas (2005) use pseudo panel models for the income data of the United States creating 10 years' interval cohorts for the period 1940–2000 based on gender, race, generation, and age. According to the results of fixed and random effect models, they stated that as the number of children increases, the probability of both spouses having an income decreases. On the other hand, Warunsiri and McNown (2010), investigate returns of education in Thailand by using national labour surveys. They choose who born between 1946 and 1967 in cohort generating process. At the end of the analysis, they compare the results of the least squares (OLS), instrumented variable (IV), pseudo panel model, and pseudo-panel instrumented variable methods and find that there are some biases in the coefficients of least squares regressions with individual data.

To the best of our knowledge, the number of studies in which pseudo panel data models are combined with spatial panel data models is few. Baltagi et al. (2015) examine the hedonic housing prices for Paris in 20 boroughs, four districts per borough, and 15–169 islands per district between 1990 and 2003. They add spatial autocorrelation into equations by describing the structure of the data as a nested spatial unbalanced pseudo panel. They use the square meter of the apartment, the number of rooms, bathrooms, balconies, and servants, distance to the district, and distance to the borough as the explanatory variables. As a result of the study, they find the spatial autoregressive effect as statistically significant. However, unlike familiar pseudo panel studies, this study adopts the cohort approach based on the nested structure of the data, not taking averages of cohorts.

As the reviewed studies show, there is no study in the literature which examines Mincer equations by using a spatial pseudo panel data model. To the best of our knowledge, even methodologically, the empirical application of this hybrid method has not been implemented. Therefore, the main contribution of this study is the demonstration of the use of this hybrid method and its application to wage equations.

3 Methodology

3.1 Pseudo panel data models

Accessing the data at the macro or panel level for least developed and developing countries is a big limitation for research about these countries. However, researchers may find the data on repeated annual household surveys, which include large samples. In such repeated cross-sectional data, it would be difficult to follow the same observation over the years. Therefore, in 1985, Deaton proposed creating cohorts from repeated cross-sectional data and examining economic relationships based on cohort-averaged data set. Creating cohorts from such repeated cross-sectional datasets allow us to obtain the panel data set, and these data sets are called "pseudo panel data." These panels are not affected by the attrition problem as in the genuine panel data. When creating cohorts, grouping should be implemented according to variables that are constant over time. In general, birth year, gender, region, or race variables are used for the creations of cohorts (Baltagi, 2008).

The literature of pseudo panel data generally uses fixed-effects (FE) models. Unlike fixed-effects models, the random effect model assumes that the individual effect is not correlated with explanatory variables. There is no point in using pseudo panel data models with making such an assumption. For independent cross-sectional data, there is no relationship between observations since each individual is observed only once. Therefore, the models can be estimated based on individual pooled data, and there is no need to convert data into pseudo panel data (Guillerm, 2017). As a result, it is assumed that pseudo panel data created by calculating cohort averages comply with the fixed-effects model.

To recall a standard panel data model:

$$Y_{it} = X_{\text{it}}^{'} \beta + \mu_i + u_{it}, \tag{1}$$

where t, i, Y, X, β, u and μ are the time, individuals/units, vector of dependant variable, matrix of independent variables, the coefficient vector, the error term and the individual effects, respectively. In the genuine panel datasets, this model is solved by calculating the difference of each variable from its mean by time with the help of a within estimator in the fixed-effects model. However, this method cannot be used since the same cross-sections are not included in the repeated cross-sectional data. To deal with this, Deaton (1985) proposes a cohort approach to obtain consistent estimators in repeated cross-sectional data. In this approach, the cohorts created according to specific common characteristics contains similar individual effects. When all observations are combined based on cohorts, the model can be obtained as follows:

$$\overline{Y}_{ct} = \overline{X}_{ct}^{'} \beta + \mu_{ct} + u_{ct}. \tag{2}$$

Here t indicates the time, while c indicates cohort number. As seen in Equation (2), variables are defined in terms of averages ($\overline{Y}_{ct}$ and $\overline{X}_{ct}^{'}$). Here, $\overline{Y}_{ct}$ shows the average

of all individuals in cohort c of time t compared to the Y variable. Cohorts created in pseudo panel data models take the place of individuals in genuine panel data. The time-varying unit effects (μ_{ct}) may be correlated with explanatory variables and random-effect models, which may lead to inconsistent estimators. On the other hand, when the cross-sectional changes are only valid for the cohorts, the time is fixed, and the number of observations in each cohort is large enough, the model is written as in Equation (3), and it can be estimated as a model of fixed-effects (Deaton, 1985):

$$\overline{Y}_{ct} = \overline{X}_{ct}^{'} \beta + \mu_c + u_{ct}. \tag{3}$$

The estimation used in the fixed-effects model is within the estimator ($\widehat{B}_W$) method and it is calculated as in Equation (4) (Verbeek, 2008):

$$\widehat{B}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\overline{x}_{ct} - \overline{x}_c)(\overline{x}_{\mathrm{ct}} - \overline{x}_c)' \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\overline{x}_{ct} - \overline{x}_c)(\overline{y}_{ct} - \overline{y}_c) \right). \tag{4}$$

Nevertheless, Deaton (1985) states that the averages are the incorrect mean estimator of the population, and thus the measurement error in the intra-group estimation method needs to be corrected. In this case, the estimator ($\widehat{B}_D$) is written as in Equation (5):

$$\widehat{B}_D = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\overline{x}_{\mathrm{ct}} - \overline{x}_c)(\overline{x}_{\mathrm{ct}} - \overline{x}_c)) \right)^{'} - \mathcal{T}\widehat{\Sigma} \Bigg)^{-1} \times \\ \times \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\overline{x}_{\mathrm{ct}} - \overline{x}_c) (\overline{y}_{\mathrm{ct}} - \overline{y}_c) - \mathcal{T}\widehat{\sigma} \right). \tag{5}$$

In this equation, $\widehat{\sigma}$ stands for the estimates of the variance of measurement errors and $\widehat{\Sigma}$ stands for estimates of covariance of them. The coefficient $\mathcal{T}$ is generally taken as 1.

According to Verbeek (1992), there are two extra dimensions in pseudo panel data, alongside with two dimensions (N and T) in genuine panel data, which are the number of cohorts C and the number of observations for each cohort n_c. Following this, Table 1 summarizes whether the estimator of Deaton or the within-group estimator is appropriate according to the N, C, n_c, and T dimensions required to obtain a consistent β coefficient. For a given n_c, the bias becomes smaller if the cohorts are chosen such that the relative magnitude of the measurement errors is smaller compared to the within cohort variance of x_{ct}. However, it may not be easy to construct cohorts in such a way. Finally, letting $n_c \to \infty$ using type 1 asymptotic is a convenient choice to arrive at a consistent estimator for β.

Table 1: Appropriate estimators according to values of N, C and n_c

	N fixed C fixed n_c fixed	N large C fixed n_c large	N large C large n_c fixed	N large C large n_c large
T fixed		*	B_D	B_D, B_W
T large	B_D	B_D, B_W	B_D	B_D, B_W

3.2 Spatial pseudo panel data models

As it is the goal of this study, we extend the existing pseudo panel data models by integrating spatial econometrics methodology into them. In this study, the spatial pseudo panel models were estimated by the application of the spatial weight matrix to the pseudo panel data models.

Based on the model in Equation (3), the fixed-effects general nested spatial pseudo panel data model (GNS FE) can be demonstrated as in Equations (6) and (7):

$$\overline{Y}_{\mathrm{ct}} = \rho W\overline{Y}_{ct} + \alpha I_C + \overline{X}_{ct}\beta + W\overline{X}_{ct}\theta + \mu_c + u_{ct}, \tag{6}$$

$$u_{\mathrm{ct}} = \lambda W u_{ct} + \varepsilon_{ct}. \tag{7}$$

Here, c is the cohort number, t is the time dimension, W is the spatial weights matrix, I is the identity matrix, α is constant, ρ is the coefficient of spatial autoregressive, θ is the coefficient of spatial effect in explanatory variables, and λ is the coefficient of the spatial impact on error terms.

Spatial pseudo panel data models can be applied to models of spatial autoregressive (SAR), spatial error (SEM), spatial Durbin (SDM), spatial autocorrelation (SAC), spatial Durbin error (SDEM), and spatial lag of X models (SLX).

The fixed-effects spatial autoregressive pseudo panel model (SAR FE) can be formed by combining the representation of the fixed-effects spatial panel autoregressive model and the pseudo panel model. It can be written as in Equation (8):

$$\overline{Y}_{ct} = \rho W\overline{Y}_{ct} + \alpha I_C + \overline{X}_{ct}\beta + \mu_c + u_{ct}. \tag{8}$$

Fixed-effects spatial error pseudo panel model (SEM FE) can be formed by combining the representation of the fixed-effects spatial error panel model and the pseudo panel model. It is shown as in Equations (9) and (10):

$$\overline{Y}_{ct} = \alpha I_C + \overline{X}_{ct}\beta + \mu_c + u_{ct}, \tag{9}$$

$$u_{ct} = \lambda W u_{ct} + \varepsilon_{ct}. \tag{10}$$

The logarithmic likelihood function of the fixed-effects spatial error model and spatial autoregressive model specified by Elhorst (2014) can be rewritten by adapting the pseudo panel to the logarithmic likelihood function.

For the rest of the models; the fixed-effects spatial autocorrelation pseudo panel model (SAC FE) is defined as in equation Equations (11) and (12):

$$\overline{Y}_{\mathrm{ct}} = \rho W \overline{Y}_{ct} + \alpha I_C + \overline{X}_{ct}\beta + \mu_c + u_{ct}, \tag{11}$$

$$u_{ct} = \lambda W u_{ct} + \varepsilon_{ct}. \tag{12}$$

The fixed-effects spatial Durbin pseudo panel model (SDM FE) can be written as in Equation (13):

$$\overline{Y}_{ct} = \rho W \overline{Y}_{ct} + \alpha I_C + \overline{X}_{ct}\beta + W\overline{X}_{ct}\theta + \mu_c + u_{ct}. \tag{13}$$

The spatial pseudo panel data estimators described here are explained by using within estimator (B_W). For SLX and SDEM models, these equations also can be adapted. All those models are estimated by maximum likelihood.

4 Data

In this study, a micro-data set of household labour force surveys of Turkstat between 2010 and 2015 is used. The panel version of this dataset is not available. Since the data is a repeated microdata, panel data analysis can be conducted via "pseudo panel data models" which is created by cohorts. We aim to examine the human capital wage equation from Standard Mincer wage equations by using spatial pseudo panel data. Cohorts were formed for each year according to 12 regions (NUTS1) and 9 year-of-birth groups.

There are some views in the literature regarding the modelling process of Mincer equations. Heckman et al. (2003) report that older data sets support Mincer's view. Therefore, they state that the functional form should be stretched. This study aims to examine the cohorts by evaluating the changes in the average wages of people in the relevant year-of-birth groups over the years. Lemieux (2006) also mentions that the effect of years of education on wages may not yield significant results as in previous studies in the next years' studies. He also states that age groups should be followed according to years and averages should be examined based on cohorts. That's why we preferred to generate year-of-birth groups when conducting pseudo panel data.

In the micro data set of household labour force survey, people who are employed and whose monthly wages are greater than zero are included in the analysis. The natural logarithm of the real wage is used as the dependent variable, whereas years of education, years of experience, and the squared of years of experience, and the proportions of female employees are used as independent variables. We preferred to use potential and proxy years of work experience variable. It is defined as age minus years of education minus 6 (see Dougherty, 2011). Additionally, the nine year-of-birth groups were defined which are the year between 1950-1954, 1955-1959, 1960-1964, 1965-1969, 1970-1974, 1975-1979, 1980-1984 and 1985-1990. We identified the upper limit age as 65 years since it is the maximum retirement age in Turkey. Additionally,

as we mentioned before, it is necessary to use averaged variables in the data generating process of pseudo panel data. However, taking the average of categorical variables would be meaningless. That's why we calculated the percentage of female employees in each cohort to show the usage of categorical variables in pseudo panel data. Lastly, one might think that spatial mobility can be important for the modelling process. In Turkey, for the 2014–2015 period net internal migration rate is relatively small as 0.03 (population ratio). Hence, we assumed that the migration would not have contaminating effect on the results.

The human capital equation was firstly examined on the micro-data set. Following that, the cohorts were formed, the pseudo panel data were analysed consecutively by using pooled data, pseudo panel data, and spatial pseudo panel data models. For the pseudo panel level, we used the within estimator (B_W) due to having large $\overline{n}_c$ (see Table 1). It was observed that the average number of observations in all cohorts is 753. To conclude, when the time dimension t is fixed (equals 6), N (equals 488140), and C (equals 108) are large enough, and it is therefore appropriate to use B_W estimator.

Table 2: Descriptive statistics in years from cross-sectional data

	Mean	Standard Deviation	N
Real Monthly Wage			
2010	562.3	436.8	83703
2011	544.3	407.1	89985
2012	581.1	445.4	93312
2013	590.1	441.8	94457
2014	613.7	472.7	92560
2015	612.6	481.0	93776
Years of Experience			
2010	19.4	11.5	83703
2011	19.6	11.7	89985
2012	19.9	11.8	93312
2013	20.0	12.0	94457
2014	20.3	12.3	92560
2015	20.4	12.5	93776
Years of Education			
2010	9.2	4.2	83703
2011	9.3	4.2	89985
2012	9.5	4.3	93312
2013	9.5	4.2	94457
2014	9.5	4.3	92560
2015	9.6	4.3	93776

The descriptive statistics of cross-sectional data were calculated before the creation of the cohorts are shown in Table 2. According to table, real wages was increased in period from 2010 to 2015 except for 2011. In 2011, Turkey experienced the highest

Table 3: Frequencies of female employees from cross-sectional data

	2010		2011		2012		2013		2014		2015	
	N	%	N	%	N	%	N	%	N	%	N	%
Male	64,560	77.13	68,732	76.38	69,832	74.84	69,708	73.8	68,443	73.94	68,406	72.95
Female	19,143	22.87	21,253	23.62	23,480	25.16	24,749	26.2	24,117	26.06	25,370	27.05

growth rate (8.5%) in recent years. Accordingly, the increases in the minimum wage for the relevant year were kept smaller (4.7%) compared to the previous year and inflation rate was 10.5% in the corresponding year. As result of these, real wages were decreased in 2011. On the other hand, in Table 3, the number of female employees is less compared to males but it tends to increase.

Table 4: Descriptive statistics in years from pseudo panel data

	Mean	Standard Deviation	N
Real Monthly Wage			
2010	601.7	122.3	108
2011	573.5	102.0	108
2012	611.6	115.6	108
2013	616.6	113.3	108
2014	640.2	107.7	108
2015	638.2	142.0	108
Experience			
2010	28.1	13.7	108
2011	29.0	13.9	108
2012	30.0	14.1	108
2013	31.1	14.3	108
2014	32.3	14.5	108
2015	33.3	14.6	108
Years of Education			
2010	8.8	1.3	108
2011	8.8	1.4	108
2012	8.9	1.5	108
2013	8.7	1.7	108
2014	8.6	1.9	108
2015	8.5	2.1	108

Descriptive statistics obtained from pseudo panel data after the creation of cohorts are given in Table 4 above. The year of experience ranges from 28 to 33, while the average education year changes between 8 and 9. Although the averages in Table 4 differ from the averages of pooled data, it should be noted that same characteristic

features were considered as cohorts, in which they were not considered in the pooled data.

5 Results

Firstly, we analysed the Standard Mincer wage equation from the pooled data and cross-sectional data of 2015 and the results are presented in Table 5. The variables used in the creation of cohorts such as NUTS1 regions were added to the models as explanatory variables.
Appropriate equation representations are shown for both equations (as in (14) and (15)) in the following way:

$$\begin{aligned}\ln(\text{Wage})_{it} = {} & \beta_0 + \beta_1 \text{Experience}_{it} + \beta_2 \text{Experience}_{it}^2 + \beta_3 \text{Education}_{it} + \beta_4 \text{Sex}_{it} + \\ & + \sum_{i=5}^{9} \beta_i \text{YearDummies} + \sum_{i=10}^{20} \beta_i \text{RegionDummies} + \\ & + \sum_{i=21}^{30} \beta_i \text{YearofBirthDummies} + e_{it} \end{aligned} \tag{14}$$

$$\begin{aligned}\ln(\text{Wage})_{i} = {} & \beta_0 + \beta_1 \text{Experience}_{i} + \beta_2 \text{Experience}_{i}^2 + \beta_3 \text{Education}_{i} + \beta_4 \text{Sex}_{i} + \\ & + \sum_{i=5}^{15} \beta_i \text{RegionDummies} + \sum_{i=16}^{25} \beta_i \text{YearofBirthDummies} + e_{i}. \end{aligned} \tag{15}$$

Here, $\ln(\text{Wage})_{it}$ shows the natural logarithm of real monthly wage of an i (person) in t (time), Experience is the years of potential experience, Education is the years of education, Sex is a dummy variable that is 1 for females, and remains show the year dummies, region dummies, and year-of-birth dummies, respectively. We also used robust standard errors to prevent the heteroscedasticity problem.
In Table 5, we find that there is no significant difference in the magnitude of the coefficients, and we also find that the effects of education and experience on wages are positive. There is also a decrease in wages after a certain level of experience. The threshold year of experience for the decrease in wages were found as 40 and 38 years for pooled and cross-sectional data, respectively. The return is around 6.8% for education, and 3% for the experience. On the other hand, average wage for females is less than for males. Increases in average wages compared to 2010 are found to be significant. As expected, in the most developed region, Istanbul, average wages are significantly higher than average wages in other regions.
Pseudo panel data were obtained from cohorts which are formed by year-of-birth groups and regions. The regression was estimated by using pooled ordinary least squares (pooled OLS) and within estimator methods. Also, we used robust standard errors for both models. As it can be seen in Table 5, the fixed-effects model is preferable because fixed effects (FE) are statistically significant ($F(107, 531) = 9.77$,

Table 5: Regressions on pooled data and cross-sectional micro data of 2015

Dependent Variable Ln(Wage)	Pooled	Cross-Sectional
Experience	0.0323*** (0.0003)	0.0307*** (0.0006)
Experience^2	-0.0004*** (0.0000)	-0.0004*** (0.0000)
Years of Education	0.0689*** (0.0002)	0.0681*** (0.0005)
Sex - Female	-0.1998*** (0.0017)	-0.1991*** (0.0039)
Years		
2011	-0.0186*** (0.0023)	
2012	0.0361*** (0.0022)	
2013	0.0631*** (0.0023)	
2014	0.1113*** (0.0023)	
2015	0.1193*** (0.0023)	
NUTS1		
West Marmara	-0.2414*** (0.0029)	-0.2191*** (0.0070)
Aegean	-0.2108*** (0.0024)	-0.1932*** (0.0060)
East Marmara	-0.1810*** (0.0025)	-0.1502*** (0.0065)
West Anatolia	-0.1519*** (0.0025)	-0.1467*** (0.0065)
Mediterranean	-0.2354*** (0.0026)	-0.2089*** (0.0066)
Central Anatolia	-0.2025*** (0.003)	-0.1774*** (0.0072)
West Black Sea	-0.2251*** (0.0031)	-0.1964*** (0.0072)
East Black Sea	-0.2259*** (0.0033)	-0.2113*** (0.0084)
Northeast Anatolia	-0.1414*** (0.0039)	-0.1246*** (0.0097)
Central East Anatolia	-0.1711*** (0.0037)	-0.1666*** (0.0083)
Southeast Anatolia	-0.2241*** (0.0031)	-0.1929*** (0.0076)
Year-of-birth Groups		
1950-1954	0.0529*** (0.0149)	0.0950** (0.0463)
1955-1959	0.1021*** (0.0139)	0.1067** (0.0434)
1960-1964	0.1482*** (0.0136)	0.1603*** (0.0427)
1965-1969	0.1723*** (0.0136)	0.1961***(0.0425)
1970-1974	0.1829*** (0.0136)	0.2150*** (0.0425)
1975-1979	0.1540*** (0.0136)	0.2076*** (0.0425)
1980-1984	0.1209*** (0.0136)	0.1953*** (0.0426)
1985-1990	0.0079(0.0136)	0.1145*** (0.0426)
Constant	5.4597*** (0.0136)	5.5176*** (0.0424)
N	488140	78348
R-Sq	0.45	0.46
AIC	602764.68	93153.07
BIC	603086.53	93375.53

Note: * p<0.1; ** p<0.05; *** p<0.01. Standard errors are in parenthesis.

$p-value = 0.00$). Appropriate equation representations are shown in Equations (16) and (17) in averages. Differently, Female shows the percentage of female employees in c cohort in t time.

$$\ln\left(\overline{\text{Wage}}\right)_{ct} = \beta_0 + \beta_1\overline{\text{Experience}}_{ct} + \beta_2\overline{\text{Experience}}^2_{ct} + \beta_3\overline{\text{Education}}_{ct} + \\ + \ \beta_4\text{Female}_{ct} + \sum_{i=5}^{9}\beta_i\text{YearDummies} + e_{ct} \tag{16}$$

$$\ln\left(\overline{\text{Wage}}\right)_{ct} = \beta_0 + \beta_1\overline{\text{Experience}}_{ct} + \beta_2\overline{\text{Experience}}^2_{ct} + \beta_3\overline{\text{Education}}_{ct} + \\ + \ \beta_4\text{Female}_{ct} + \sum_{i=5}^{9}\beta_i\text{YearDummies} + \mu_c + e_{ct} \tag{17}$$

Table 6: Regressions on pseudo panel data

Dependent Variable Ln(Wage)	Pooled OLS	FE
Experience	0.0851*** (0.0045)	0.0461*** (0.0101)
Experience^2	-0.0028*** (0.0002)	-0.0015*** (0.0004)
Education	0.0824*** (0.0052)	0.1109*** (0.0135)
Sex: Female	-0.0010 (0.0008)	0.0008 (0.0011)
Year		
2011	-0.0443*** (0.0150)	-0.0439*** (0.0068)
2012	0.0096 (0.0164)	0.0060 (0.0081)
2013	0.0334** (0.0162)	0.0324*** (0.0103)
2014	0.0876*** (0.0167)	0.0881*** (0.0103)
2015	0.0806*** (0.0192)	0.0821*** (0.0129)
Constant	5.1967*** (0.0628)	5.1222*** (0.1420)
R-Sq	0.65	0.61
AIC	-894.22	-1601.25
BIC	-849.48	-1560.98
c	108	108
t	6	6
N	648	648
σ_u		0.14
σ_e		0.08
ρ		0.77
F test $u_i = 0$:		9.77 (prob:0.00)

Note: * p<0.1; ** p<0.05; *** p<0.01. Standard errors are in parenthesis.

For the fixed-effects model, the effects of education and experience on wages are to be positive. There is also a decrease in wages after a certain level of experience. That

is found as about 15 years. The returns are around 11% for education and 4.6% for the experience. On the other hand, the coefficient of Female is not significant.
To perform sensitivity analysis, cohorts were changed and the regression with the fixed-effects is re-estimated. Firstly, wage equations obtained from cohorts which were formed according to 46 year-of-birth group (one-year interval age groups) and 12 regions were estimated, the results are given in Table 10 in Appendix A. The coefficients of years of experience and years of education slightly differ from values which are obtained in Table 6. The disadvantage of the model in the appendix is that there is a fewer number of observations in the cohorts due to the large number of cohorts. As it can be seen from data, it is an unbalanced data. Since there is a trade-off between the number of cohorts and the number of observations in each cohort, forming a cohort according to 9 year-of-birth groups and 12 regions gave more appropriate results.
After this step, existence of spatial relationship between the regions and wages was investigated with the binary contiguity matrix based on the neighbourhood relationship to the pseudo panel data model (see the results in Table 7). According to the results of Moran's I test, there is a statistically significant spatial relationship. However, LM tests should be considered if this relationship is in error terms or dependent variable. The results of LM tests reveal that there is a statistically significant spatial relation in both autoregressive and error terms. Beside these tests, the best model is determined by testing the nested structures between models.

Table 7: Tests of spatial correlation

Test	Statistic	df	p-value
Moran's I	11.1	1	0.000
Spatial Error Lagrange Multiplier	38.5	1	0.000
Spatial Lag/Autoregressive Lagrange Multiplier	21.1	1	0.000

Standard Mincer wage equations were then estimated by using fixed-effects spatial pseudo panel data models (SAR, SEM, SAC and SDM) with robust standard errors. In the model selection process, for SDM model, we obtained the coefficients of $W\overline{X}_{ct}$ to be insignificant ($[Wx]x1 = [Wx]x2 = [Wx]x3 = [Wx]x4 = 0$). Therefore, we did not estimate other models such as SLX and SDEM that investigate spatial effects in explanatory variables. We also found that the ρ coefficient is insignificant in the SDM model. Afterwards, we concluded that the λ and ρ coefficients together in the SAC model were different from zero ($\rho = \lambda \neq 0$). Same coefficients were tested for SEM and SAR models, the coefficients showing spatial relationships were found to be insignificant ($\rho = 0$ and $\lambda = 0$). Besides this results, AIC and BIC criteria found to be very close to each other. Therefore, there was enough evidence that SAC model is the best fit for our analysis.

Table 8: Fixed-effects spatial pseudo panel data models

Dependent Variable Ln(Wage)	SAR FE	SEM FE	SAC FE	SDM FE
Experience	0.0459*** (0.0099)	0.0459*** (0.0101)	0.0453*** (0.0097)	0.0430*** (0.0100)
Experience^2	-0.0015*** (0.0004)	-0.0014*** (0.0004)	-0.0014*** (0.0004)	-0.0014*** (0.0004)
Education	0.1109*** (0.0134)	0.1110*** (0.0135)	0.1103*** (0.0133)	0.1130*** (0.0134)
Sex: Female	0.0008 (0.0011)	0.0008 (0.0011)	0.0007 (0.0011)	0.0006 (0.0010)
Years				
2011	-0.0492*** (0.0074)	-0.0438*** (0.0073)	-0.0561*** (0.0103)	-0.0506*** (0.0093)
2012	0.0080 (0.0084)	0.0063 (0.0086)	0.0123 (0.0123)	0.0015 (0.0171)
2013	0.0359*** (0.0114)	0.0323*** (0.0110)	0.0405** (0.0162)	0.0137 (0.0224)
2014	0.0958*** (0.0128)	0.0881*** (0.0108)	0.1067*** (0.0173)	0.0555** (0.0267)
2015	0.0886*** (0.0130)	0.0826*** (0.0134)	0.0998*** (0.0173)	0.0428 (0.0372)
Wx				
Experience				-0.0295 (0.0558)
Experience^2				-0.0002 (0.0023)
Education				0.0234 (0.0378)
Female				0.0041 (0.0042)
Spatial				
ρ	-0.1182 (0.1210)		-0.2815* (0.1687)	-0.0099 (0.0859)
λ		0.0678 (0.0798)	0.2687** (0.1245)	
Variance				
σ_e^2	0.0048*** (0.0012)	0.0048*** (0.0012)	0.0057*** (0.0011)	0.0047*** (0.0011)
Statistics				
R-Sq	0.42	0.42	0.43	0.39
N	648	648	648	648
AIC	-1598.03	-1597.44	-1598.00	-1598.04
BIC	-1548.82	-1548.23	-1544.31	-1530.93

Note: * p<0.1; ** p<0.05; *** p<0.01. Standard errors are in parenthesis.

Finally, fixed-effects spatial autocorrelation pseudo panel data model can be expressed as in Equation (18):

$$\begin{aligned} \ln\left(\overline{\text{Wage}}\right)_{ct} &= \beta_0 + \rho W \ln\left(\overline{\text{Wage}}\right)_{ct} + \beta_1 \overline{\text{Experience}}_{\text{ct}} + \beta_2 \overline{\text{Experience}}^2_{ct} + \\ &+ \beta_3 \overline{\text{Education}}_{\text{ct}} + \beta_4 \text{Female}_{\text{ct}} + \sum_{i=5}^{9} \beta_i \text{YearDummies} + \mu_c + u_{ct}, \\ u_{\text{ct}} &= \lambda W u_{ct} + \varepsilon_{ct}. \end{aligned} \tag{18}$$

Although the magnitudes of the coefficients in Table 8 and Table 6 seem close to each

other, the coefficients obtained from this equation are unbiased and efficient due to the presence of spatial effect in both the autoregressive and error term.

6 Discussion

Individual characteristics that cannot be observed in pooled datasets can vary from year to year since the sample differs each year. Also, OLS estimates are biased if the unobservable individual characteristics in these data sets are correlated with explanatory variables. To deal with this bias, Deaton (1985) suggested the cohort approach which is based on individual characteristics and does not change over the years. Additionally, if there is any spatial correlation in the data set based on cohorts, estimation of OLS without including spatial effects may cause the coefficients to be biased or inefficient according to the type of spatial interaction. For instance, LeSage and Pace (2009) indicate that if the data generation process is the SEM model; the SAR, SAC, and SDM models will produce unbiased but inefficient coefficients. With these motivations, we examined the human capital Standard Mincer wage equation, which has been commonly used in the literature by using spatial pseudo panel data models rather than individual cross-sectional data. In this study, we aimed to introduce a new usage of it by applying it to other studies. This methodological advancement achieved in our research makes it significant and distinguished one from other studies in the field.

According to our results, the effects of experience and education on wages are similar in both the fixed-effects pseudo panel (FE) and fixed-effects spatial autocorrelation pseudo panel (SAC FE) model. For example, the turning point of the function for the years of experience is 15 for the pseudo panel (FE) model and 16 years for the spatial autocorrelation pseudo panel (SAC FE) model. Moreover, the effect of education on the wage is obtained as 6.8% in the pooled OLS estimation of cross-sectional data whereas it is 11% for both the fixed-effects pseudo panel data model (FE) and fixed-effects spatial autocorrelation pseudo panel data (SAC FE) model. Therefore, it can be concluded that the result from the OLS estimation is downward bias. However, Himaz and Aturupane (2015) use pseudo panel data models for Sri Lanka to estimate the return on education between 1997 and 2008 by constructing 9 cohorts. Using the pseudo panel estimation, rather than 8% as in the OLS estimation, they estimate that one extra year of education increases monthly earnings by about 5% for males. However, without controlling unobservable characteristic bias in the OLS estimation of returns, it upwards by about 3% points on average. In another study, Warunsiri and McNown (2010) examine the returns of education for Thailand by using Mincer equations for urban residents. They found the coefficient of education as 0.11 in OLS estimation, while it was 0.18 from the pseudo panel approach that is based on two-year cohorts mean. On the other hand, in our study, the coefficient of the experience variable is estimated as 3.2% from the OLS model, while it is estimated at 4.5% from the fixed-effects spatial autocorrelation pseudo panel data model (SAC FE).

The coefficient of experience or age is also found downward bias from OLS estimation by the results of Warunsiri and McNown (2010).
There is a negative spatial relationship directly in wages. The positive sign of ρ shows the clustering of similar regions as well as common reactions. The negative relations demonstrate the dissimilarity – a kind of competition or the backwash effect (Kao and Bera 2016). Finally, we also calculated the direct and indirect effects on wages (see Table 9). For example, the 1-year increase in the average year of education of any cohort in any region leads to a 11.2% increase in the average wages of the relevant cohort in that region. But this decreases the wages of neighbour region by 2.4%. So, the total return of education equals 8.8%.

Table 9: Direct, indirect and total effects

	Coefficients	Robust Standard Errors	z	$P > \|z\|$
Direct				
Experience	0.046	0.01	4.57	0.00
Experience^2	-0.001	0.00	-3.61	0.00
Education	0.112	0.01	8.62	0.00
Female	0.001	0.00	0.66	0.51
Indirect				
Experience	-0.010	0.01	-1.69	0.09
Experience^2	0.0003	0.00	1.62	0.11
Education	-0.024	0.01	-1.81	0.07
Female	0.000	0.00	-0.60	0.55
Total				
Experience	0.036	0.01	4.31	0.00
Experience^2	-0.001	0.00	-3.44	0.00
Education	0.088	0.01	7.34	0.00
Female	0.001	0.00	0.64	0.52

7 Conclusions

In this study, we examine the impact of human capital variables on the wage through the spatial effects of wages over the years by using spatial pseudo panel data models. As a result of the study, the coefficients obtained from the pseudo panel with fixed-effects (FE) model are found to be different from the coefficients of the model which is estimated by OLS. The pseudo panel data set give unbiased results due to the modelling of unobservable cohort effects. Unbiased and efficient coefficients are obtained by including spatial effects in this model. It has been observed that there is a spatial relationship between regions of Turkey in terms of wages.

The main contribution of this study is to introduce the use of spatial pseudo panel data models empirically via human capital Mincer equation. The application of this hybrid method to different fields of study, where panel data cannot be observed directly, based on regions provides valuable and more accurate information for policy-makers.

References

[1] Baltagi B., (2008), Econometric Analysis of Panel Data, West Sussex, John Wiley & Sons.

[2] Baltagi B. H., Bresson G., Etienne J. M., (2015), Hedonic Housing Prices in Paris. An Unbalanced Spatial Lag Pseudo-Panel Model with Nested Random Effects, *Journal of Applied Econometrics* 30(3), 509–528, DOI: 10.1002/jae.2377.

[3] Deaton A., (1985), Panel Data from Time Series of Cross Sections, *Journal of Econometrics* 30, 109–12, DOI: 10.1002/jae.2377.

[4] Dougherty C., (2011), Introduction to Econometrics, Oxford University Press, UK.

[5] Elhorst J. P., (2014), Spatial Econometrics. From Cross-Sectional Data to Spatial Panels, Heidelberg, Springer.

[6] Elhorst J. P., Blien U., Wolf K., (2007), New Evidence on The Wage Curve. A Spatial Panel Approach, *International Regional Science Review* 30(2), 173–191, DOI: 10.1177/0160017606298426.

[7] Guillerm M., (2017), Pseudo-Panel Methods and An Example of Application to Household Wealth Data, *Economie et Statistique Année* 491–492, 109–130.

[8] Heckman J. J., Lochner L. J., Todd P. E., (2003), Fifty Years of Mincer Earnings Regressions, National Bureau of Economic Research Working Paper No 9732, DOI: 10.3386/w9732.

[9] Himaz R., Aturupane H., (2015), Returns to Education in Sri Lanka. A Pseudo-Panel Approach, *Education Economics* 24(3), 300–311, DOI: 09645292.2015.1005575.

[10] Howie P. J., Kleczyk E. J., (2007), New Developments in Panel Data Estimation: Full-Factorial Panel Data Model, American Agricultural Economics Association, Portland, Oregon, July 29-August 1, 2007.

[11] Kao S. Y. H., Bera A. K., (2016), Spatial Regression: The Curious Case of Negative Spatial Dependence, Urbana-Champaign, Mimeo: University of Illinois.

[12] Lemieux T., (2006), The "Mincer Equation" Thirty Years After Schooling, Experience, And Earnings, [in:] Jacob Mincer a Pioneer of Modern Labor Economics, [ed.:] S. Grossbard, 127–145, Springer, Boston, MA.

[13] LeSage J. P., Pace R. K., (2009), Introduction to Spatial Econometrics, Boca Raton, CRC Press Taylor & Francis Group.

[14] Longhi S., Nijkamp P., Poot J., (2006), Spatial Heterogeneity and The Wage Curve Revisited, *Journal of Regional Science* 46(4), 707–731, DOI: 10.1111/j.1467-9787.2006.00474.x.

[15] Mincer J., (1958), Investment in Human Capital and Personal Income Distribution, *Journal of Political Economy* 66(4), 281–302, DOI: 10.1086/258055.

[16] Mincer J., (1974), Schooling, Experience, and Earnings, Human Behavior & Social Institutions No 2.

[17] Mincer J., (1981), Human Capital and Economic Growth, NBER Working Paper No 0803, DOI: doi.org/10.3386/w0803.

[18] Ramos R., Nicodemo C., Sanromá E., (2015), A Spatial Panel Wage Curve for Spain, *Letters in Spatial and Resource Sciences* 8(2), 125–139, DOI: 10.1007/s12076-014-0118-y.

[19] Russell J. E., Fraas J. W., (2005), An Application of Panel Regression to Pseudo Panel Data, *Multiple Linear Regression Viewpoints* 31(1), 1–15.

[20] Verbeek M., (1992), Pseudo Panel Data, [in:] The Econometrics of Panel Data, [eds.:] Mátyás L., Sevestre P., Dordrecht, Springer, 303–315, DOI: 10.1007/978-94-009-0375-3_14.

[21] Verbeek M., (2008), Pseudo-Panels and repeated cross-sections, [in:] The Econometrics of Panel Data, [eds.:] Mátyás L., Sevestre P., Berlin, Springer, 369–383, DOI: 10.2139/ssrn.869445.

[22] Warunsiri S., McNown R., (2010), The Returns to Education in Thailand. A Pseudo-Panel Approach, *World Development* 38(11), 1616–1625, DOI: 10.1016/j.worlddev.2010.03.002.

Appendix A

Table 10: Pseudo panel FE model based on different cohorts

Dependent Variable Ln(Wage)	Pseudo Panel FE
Experience	0.0371*** (0.0066)
Experience^2	-0.0008*** (0.0002)
Education	0.0902*** (0.0055)
Sex: Female	-0.0006 (0.0007)
Year	
2011	-0.0265** (0.0104)
2012	0.0246** (0.0121)
2013	0.0504*** (0.0105)
2014	0.1091*** (0.0118)
2015	0.0993*** (0.0144)
Constant	5.3170*** (0.0528)
R-Sq	0.44
c	550
t	6
N	3300
σ_u	0.12
σ_e	0.14
rho	0.33
AIC	-2757.90
BIC	-2702.98
F test $u_i = 0$:	1.94 prob:0.00

Note: * p<0.1; ** p<0.05; *** p<0.01. Standard errors are in parenthesis.