Article

Equal opportunities to access the university in Chile? An application with a spatial Heckman probit model

Juan Luis Quiroz [1], Ludo Peeters [2], Coro Chasco [3,4]*, and Patricio Aroca [5]

[1] Faculty of Economics and Business Administration, Universidad Autónoma de Madrid. C/ Francisco Tomás y Valiente, 5, 28049 Madrid, Spain; juan.quiroz@predoc.uam.es
[2] Faculty of Business Economics (BEW), Hasselt University. Campus Diepenbeek | Edificio Agoralaan D BE3590 Diepenbeek, Belgium; ludo.peeters@uhasselt.be
[3] Department of Applied Economics, Universidad Autónoma de Madrid. C/ Francisco Tomás y Valiente, 5, 28049 Madrid, Spain; coro.chasco@uam.es
[4] Nebrija University. C/ Sta. Cruz de Marcenado, 27, 28015 Madrid, Spain
[5] Adolfo Ibáñez University. Padre Hurtado 750, Viña del Mar, Chile; patricio.aroca@uai.cl
* Correspondence: coro.chasco@uam.es

Abstract: This study contributes to the debate on accessibility of higher education in Chile, focusing on both socioeconomic and geospatial dimensions of access to university study. The central question we address in this paper is the following: Does geography (physical distance and neighborhood effects) play a significant role in determining accessibility of higher education in Chile? We use Heckman probit-type (*Heckit*) models to adjust for selection in the process of completing the trajectory towards higher education – that is, pre-selection, application to study at university, and ultimately admission (or denial) to a higher education institution. The results shows that the geospatial elements have a significant local effect on the student's application and access to Chilean universities.

Keywords: Heckit models; spatial effects; local spatial autocorrelation; SLX model; education accessibility; Chile

1. Introduction

Access to university education is a real concern for policymakers all over the world. This is particularly true in the case of Chile, where the government has invested intense effort in providing equal access for students from families of different socioeconomic levels. Due to stratification of education in Chile, admission to universities reflects an inequitable educational system ([1,2]). At the same time, economic and demographic concentration in the central part of the country plays a large role in students' performance, as is reflected in the results of the University Selection Test (PSU); the best results are clustered in the central area around the city of Santiago de Chile.

Unfortunately, an OECD report ([3]) concluded that differences in university access and student performance persist in Chile. Specifically, the report found indications that access to higher education depends on students' socioeconomic status, secondary schooling, and region of origin. It is thus important to understand how personal characteristics, reflected in socioeconomic and geographic items, influence students' enrollment in university education to generate empirical evidence that contributes to the design of public policies in the area of university education. To our knowledge, this conclusion has never been tested with real microdata in Chile. This paper's goal is thus to examine access to university education in Chile using cross-sectional data for about 300,000 students who finished high school in 2016. The database we use contains information on students' high-school grades, selection-test scores, and applications to the university, as well as personal and family characteristics.

Our analysis focuses on the two stages each student must complete before he/she can begin university study in Chile. In the first stage, each student must pass a selection test; if they pass the test, they may apply to university. The university application implies the student's deliberate decision or willingness to participate in university education, which is clearly conditional on passing the test. In the second stage, after the university application, each student must wait for the admissions decision (admission/denial) to access university education. This decision is taken by the Department of Evaluation, Measurement and Educational Registration (DEMRE) of the University of Chile in Santiago de Chile.

While sociology and psychology have highlighted the main role of the environment where students live and their social networks, the economic literature focuses mainly on the socioeconomic characteristics of the family as a key factor defining students' probability of attending university. Our database on the PSU overcomes the difficulty of finding fine-scale georeferenced secondary statistics from which to build spatial models. We focus specifically on two types of potential socio-interaction effects on the student's decisions: neighborhood and network (social capital).

To improve our understanding of the determinants of accessibility to higher education in Chile, we estimate a sequence of two Heckman probit (Heckit) models, one for each stage mentioned above. We estimate both non-spatial and spatial versions of the Heckit model, where the spatial application accounts for the potential impact of social interactions—in the form of local spatial autocorrelation—on probability of passing the initial selection test and likelihood of applying to the university. We also used confirmatory factor analysis to create a latent variable used in our model to avoid collinearity and handle the problem of choosing between two well-known proxy variables for students' social capital: 1) parents' education, measured by household income; and 2) type of student's secondary school. Estimation of the Heckit models provides evidence of significant differences in university access in Chile, depending on gender, social network/capital, and geographic location of the student's home province.

This paper is organized as follows. After the introduction, we present the main characteristics of the university admission system in Chile. The third section presents data sources and variable statistics. Sections 4 and 5 then develop the estimation strategy and results, respectively. The conclusions and references close the paper.

2. The Chilean higher education admissions system

In Chile, the university admissions system has historically been based on two indicators ([1]). The first is the score the student obtains on a standardized test (PSU) that measures skills in the areas of mathematics, language and communication, history, social science and geography, and sciences. The second is the grades the student obtained in secondary education.

Until January 2021, the application process to access Chilean universities consisted of three steps:

1. The students had to pass the PSU, organized by the DEMRE ("Departamento de Evaluación, Medición y Registro Educacional" (DEMRE) of the Universidad de Santiago. To pass, they have to obtain a minimum score of 475 points out of 850.
2. Once they pass the PSU, prospective students must decide whether to apply to the universities that belong to the Unified Admission System (SUA). Historically, only the traditional universities used this system. In 2011, non-traditional universities were allowed to participate after evaluation by the Council of Chilean University Vice-Chancellors (CRUCH) to determine whether they met the necessary quality standards.
3. After submitting their application, students received an admission decision, based on their PSU score.

[3,4] analyzed the standard tests for the application process to universities in Chile. The authors noted the need to consider the geographical location of the students' home, as its impact was not clear. More specifically, the OECD's report observed that the PSU

score might well be explained by family income level, secondary school performance, and urbanization level. In fact, many rural areas have smaller numbers of schools and fewer resources.

Students who decide to apply to the SUA universities may choose up to 10 options, applying to different degree programs at different universities. They must rank their selections to prioritize the programs in order of preference; once their score earns them acceptance to a university program, the other options are disqualified. Since acceptance/denial is based on the student's PSU score, the students with the best scores are more likely to be accepted into the program of their choice.

The debate over higher education frequently mentions "equity." Improving "equity access" is vital because students from low-income families are least likely to access post-secondary studies ([1]). The problem of equity access stems from a multiplicity of related phenomena ([5,6]), but the most significant in the case of Chile is socioeconomic status. Secondary education quality varies significantly, as children from low-income families who cannot pay for a private school achieve less academic success and thus fall into the least advantaged student group. This situation has repercussions in the university application process. There is an urgent need to reflect family background in the process ([7,8]).

3. Data and variables

3.1. Data source and descriptive statistics

Our data were provided by the Universidad de Antofagasta from the DEMRE. This entity holds the official PSU test score records with some information about all high-school graduates who took this selection test in the year 2016. The database also contains information about students' secondary school type and grades, PSU selection-test scores and applications to the university, and basic personal characteristics. Initially, this database covered about 300,000 students, although only 267,233 students finally took the exam. After eliminating the records that contained missing values, the database retained a total of 260,775 "useable" observations, that is, the entire population sample of students potentially eligible for participation in higher education.

The database was georeferenced using an R script, to construct the spatial weights matrices and calculate the geographical distances to the centroid of Santiago de Chile city, which is also the socioeconomic center of the country. Rural-urban classification was taken from the Ministry of Education (MINEDU), which assigns this qualification to the schools according to students' origin.

Given that our main goal was to examine access to higher education in Chile and potential disparities in the process of university enrollment across different groups of high-school graduates, we focused on a set of core variables representing basic student characteristics, as well as geographical distance from students to universities. We also examined the role of the variable "social capital" in students' decisions to apply to university.

3.1.1. Students' characteristics

Key variables in our analysis are individual characteristics of the high-school graduates, such as gender and age. Two sets of characteristics of great interest in our analysis were used as proxy variables for latent ability and motivation/opportunities, respectively. As shown in **Table 1**, the first set comprises secondary school grades ("GRADE_PTS") and PSU selection-test scores ("LIT_SCORE" and "MATH_SCORE"), which evaluate cognitive skills related to literature and mathematics, respectively). The second set of characteristics includes the students' employment status ("WORKING") and their siblings' education ("SIBL_UNIV") ([9,10]).

Table 1. Descriptive statistics for the sample of university candidates.

Variables	Description	# Obs.	Mean	Std. Dev.	Min.	Max.
A: Indicator (dummy) variables – Model dependent variables:						
PRE-SELECTION	Successful pre-selection test	260,775	0.603	0.489	0	1
APPLICATION	Application for university place	260,775	0.539	0.498	0	1
ADMISSION	Admission accepted	260,775	0.379	0.485	0	1
B: Indicator (dummy) variables – Model independent variables:						
FEMALE	Female	260,775	0.530	0.499	0	1
WORKING	Working	260,775	0.098	0.298	0	1
RURAL	Rural origin area	260,775	0.020	0.140	0	1
SIBL_UNIV	Siblings in university	260,775	0.300	0.458	0	1
C: Continuous variables – Model independent variables						
DISTANCE	Distance (km)	260,775	312.8	426.1	0.040	3,772
LIT_SCORE	Score literature test	260,775	506.0	109.8	150	850
MATH_SCORE	Score mathematics test	260,775	505.7	109.4	150	850
GRADE_PTS	High-school grade points	260,775	544.5	98.9	238	826
D: Continuous latent variable – Model independent variable						
SOCIAL_CAP	Social capital	260,775	0.000	2.505	-4.237	6.467

3.1.2. Location factors

First, we distinguish between two types of students' place of origin using information about the type of college (public, subsidized, private). Students who graduated are represented as a dummy variable, "rural" if the students graduated from a rural college or "urban" if they graduated from an "urban college."

Second, we considered the role of geographical distance as affecting access to higher education, through its relationship to success in passing the selection test and propensity to apply to study at university. To this end, we calculated the distances from each student's home location to Santiago de Chile city centroid ("DISTANCE"). The average distance to the Santiago city centroid is about 310 km, ranging from 40 meters to about 3,800 km. Distance is an important variable because the best PSU selection test scores correspond to residents living in the center of the country, close to Santiago, perhaps due to the extreme socioeconomic concentration of Chile around the Metropolitan Region of Santiago. We thus expect that the larger this distance the smaller the probability of successfully passing the selection test and the propensity to apply to study at university.

3.1.3. Localized social capital

The social capital variable is a complex variegated social mechanism. Parents garner social capital to give their children the best chance of success in personal and professional life. [11]'s notion of social capital is attractive because it provides a conceptual link between the attributes of individual actors and their immediate social contexts, most notably family, school, and neighborhood [12]. These authors provide a simple way to compute this variable, defining social capital as a mere combination of three variables: tangible economic aspect, intellectual aspect, and social networks.

A strong correlation exists between parents' (father's and mother's) education, family income, and students' school type. This correlation causes a multicollinearity problem when these variables are used in a regression model. To handle this situation, we decided to build a latent variable, "social capital," with these four variables, using the Furstenberg and Hughes' definition. In Chile, the students' ordinary school is a good social capital "proxy," since access to elitist private secondary schools is conditioned by household income and parents' high education level ([13]).

Table 1 also presents basic sample descriptive statistics for the 267,233 high-school graduates participating in the 2016 PSU selection test. The table also includes descriptive statistics for the model control variables. The variable "APPLICATION" includes the group of 18,885 students who did not pass the PSU. On average, only 2% of high-school graduates come from rural, and 30% have siblings in the university. Additionally, only 9.8% of the PSU candidates are working, probably because most people who work value their present incomes more than future earnings from a university degree.

3.2. *Exploratory Data Analysis*

3.2.1. Higher Education System Design: Selection – Application – Admission

Of all high-school graduates who took the PSU selection test in 2016, only 53.9% passed (**Figure 1**). Next, of the 60.3% students who passed the selection test, only in fact applied to university. Finally, only 37.9% of the initial high-school graduates who decided to take the PSU had access to higher education.

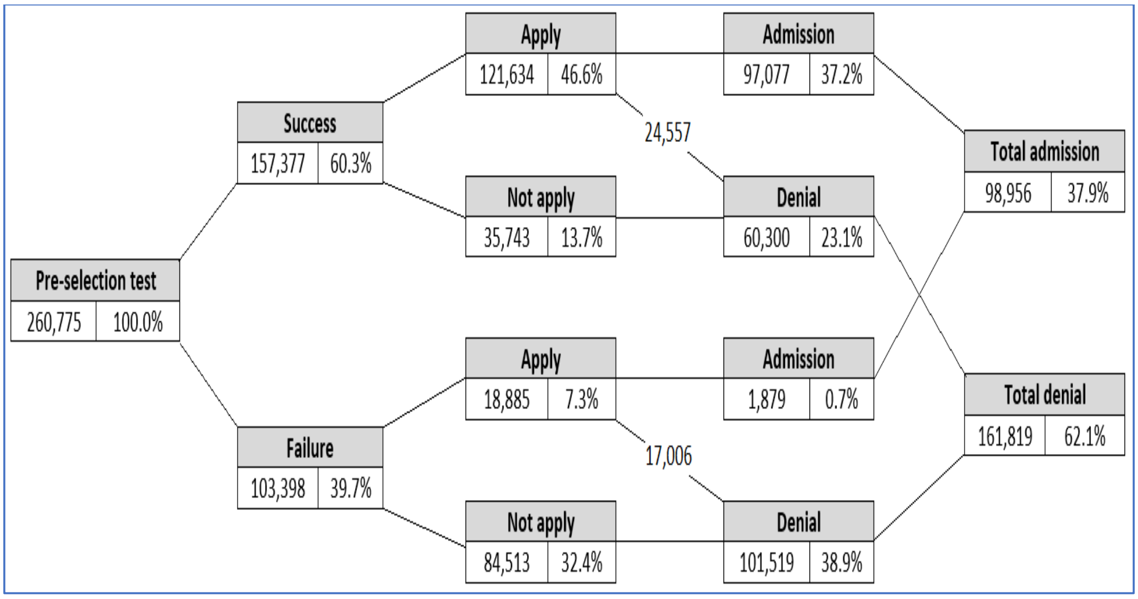


Figure 1. Trajectory data distribution

The minimum score for the pre-selection test is 475 points. A student must pass this pre-selection test before applying for university admission and accessing higher education. Yet some universities make exceptions, admitting candidates with a score of 450.

3.2.2. Geography of access to higher education: distances and neighborhoods

The detailed microlevel information provided by our database (i.e., each student's postal address) enables us to examine the spatial dimension of university access in Chile. Students' geospatial context is likely to determine their social class and identity, influencing their decision-making throughout the three stages of the process of accessing higher education in Chile.

Unfortunately, we encountered difficulties in georeferencing students' locations. Several postal addresses contained odd characters, and some addresses registered did not exist. To solve these problems, we applied an R-function to geocode the addresses based on Google's Application Programming Interface for the Geo-Coding Function. When we found erroneous addresses that we could not geocode exactly (25% of the total addresses), we assigned these locations the centroid coordinates of their corresponding communes.

We checked this variable to ensure that all addresses were within the expected distance radius. In **Figure 2**, boxplots are used to visualize the outcome of the georeferencing process. The boxplots represent the distribution of distances from each candidate's home to the city of Santiago. There is one boxplot for each Chilean region. The horizontal axis plots the Chilean regions in Roman numerals. XV represents Arica and Parinacota, I Tarapacá, II Antofagasta, III Atacama, IV Coquimbo, V Valparaíso, RM Metropolitan Region of Santiago, VI O'Higgins, VII Maule, VIII Bío-Bío, IX Araucanía, XIV Los Ríos, X Los Lagos, XI Aysén, and XII Magallanes. The figure shows only a few outliers in the V Region (Valparaíso). These outliers are entirely accurate, as they correspond to some of the Pacific islands that fall under the Valparaíso Region's administration.

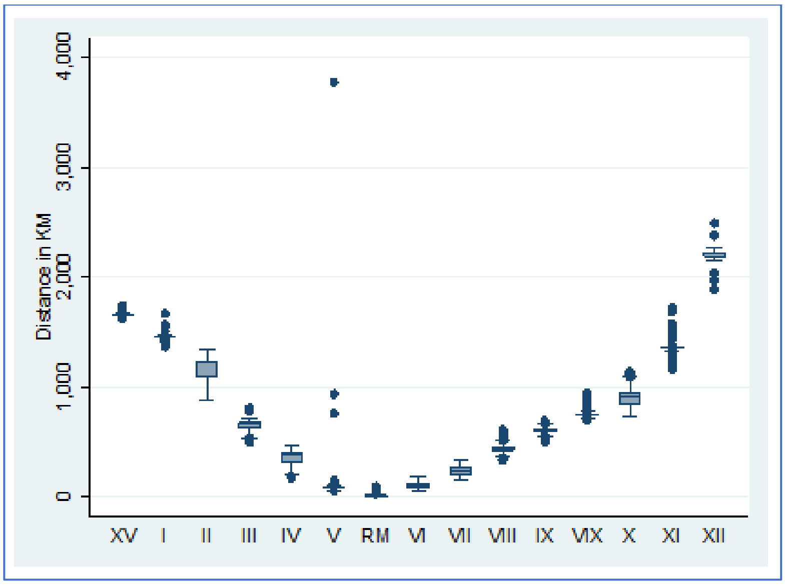


Figure 2. Distance to Santiago distributions after the georeferencing process.

4. Estimation strategy

We are interested in the effects that individual student characteristics have on the probability that recent high-school graduates—at least those who participated in the selection test—will access university education in Chile. This goal gives rise to two separate (sequential) procedures, visualized in the flow chart in **Figure 3**.

4.1. Heckman Probit models

We estimate two Heckman probit (Heckit) models. Both baseline models use the same sample population of (260,755) students potentially eligible for higher education in Chile in 2016.

4.1.1. Baseline Model 1

The modeling strategy assumption in Model 1 is that high school graduates' primary decision is whether or not to take the (mandatory) country-wide pre-selection test. Only in the second stage, conditional upon successfully passing this pre-selection test, must the student decide whether to apply to the university. In the absence of longitudinal data, we use the natural estimation strategy, a Heckman correction procedure.

More specifically, we use the Heckit model ([14]; see also [15]), which allows for the estimation of two probit models with controls for self-selection bias, which may arise due to exclusion of students who exit the application process for higher education (i.e., insufficient score on the PSU or no application to university). We find two examples of Heckit estimations close to those in this paper in [16-17].

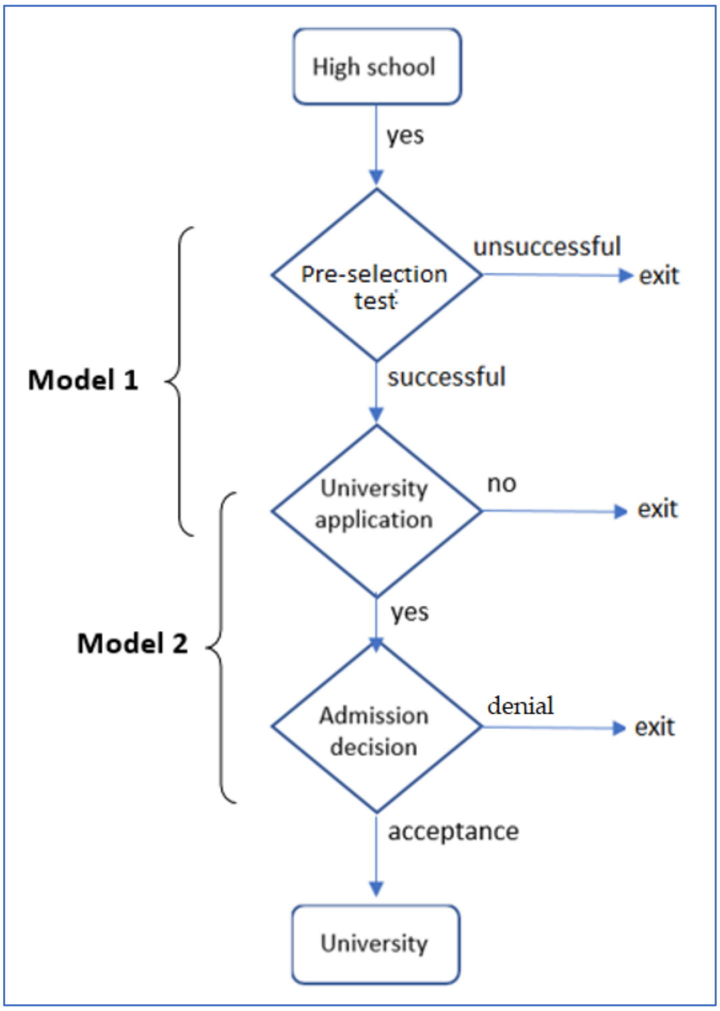


Figure 3. Flow chart of trajectory to higher education.

In Model 1, if the students who must decide whether or not to apply to the university differ systematically from high-school graduates who did not pass the pre-selection test, the estimated coefficients of the determinants of the application decision are likely to be biased. To address this potential selection bias in the probit estimation, we estimate two probit models, where each model consists of a (probit) selection equation and a (probit) outcome equation (see also [18]):

$$y_{1i}^{apply} = (x_{1i}\beta_1 + u_{1i} > 0)\,, \tag{1}$$

$$y_{2i}^{pre-select} = (x_{2i}\beta_2 + u_{2i} > 0)\,, \tag{2}$$

where equation (1) is the main equation and equation (2) the selection equation. The binary outcome in equation (1), which is related to the student's decision whether to apply to university, is of course only observed if the student passes the PSU. The dependent variable in equation (2) is also a binary variable and takes a value of one for students who passed the PSU and zero otherwise.

It is further assumed that the error terms, representing idiosyncratic unobservable variables, are bivariate normal and independent of the explanatory variables (exogeneity) in both equations; that is:

$$u_{1i} \sim N(0,1)\,, \tag{3}$$

$$u_{2i} \sim N(0,1)\,, \tag{4}$$

$$corr(u_1, u_2) = \rho \,, \tag{5}$$

where the models are estimated using Maximum Likelihood (ML). The log likelihood is computed as follows:

$$\begin{aligned}\ln L = &\sum_{\substack{i\in S\\ y_i\neq 0}} w_i \ln\left\{\Phi_2\left(x_{1i}\beta_1 + \text{offset}_i^{\beta_1}, x_{2i}\beta_2 + \text{offset}_i^{\beta_2}, \rho\right)\right\}\\ &+ \sum_{\substack{i\in S\\ y_i=0}} w_i \ln\left\{\Phi_2\left(-x_{1i}\beta_1 + \text{offset}_i^{\beta_1}, x_{2i}\beta_2 + \text{offset}_i^{\beta_2}, -\rho\right)\right\}\\ &+ \sum_{i\notin S} w_i \ln\left\{1 - \Phi\left(x_{2i}\beta_2 + \text{offset}_i^{\beta_2}\right)\right\},\end{aligned} \tag{6}$$

where S is the set of observations for which y_i is observed, $\Phi_2(\cdot)$ is the cumulative bivariate normal distribution function (with mean [0 0]'), $\Phi(\cdot)$ is the standard cumulative normal, and w_i is an optional weight for observation i.

The selection-bias problem in equation (1) occurs when the error terms in the two equations are correlated ($\rho \neq 0$). The Heckit approach should correct for such selection biases by also estimating equation (2), thus providing consistent and asymptotically efficient estimates for the unknown parameters in the model.

4.1.2. Baseline Model 2

We follow the same approach as in Model 1. Along similar lines, the model encompasses the binary outcome in equation (7). This outcome is related to the DEMRE's final decision to admit or deny students access to higher education, where the binary outcome is only observed if the student in fact applies to university. The dependent variable in equation (8) is also a binary variable that takes a value of one for students who applied to university, and zero otherwise. If applicants admitted to university education differ systematically from high-school graduates who did not apply to university, the estimated coefficients of the determinants of the admission decision are likely to be biased.

To address this potential selection bias, we again estimate two probit models, each model consisting of a (probit) selection equation and an (probit) outcome equation:

$$y_{1i}^{admit} = (x_{1i}\beta_1 + u_{1i} > 0) \,, \tag{7}$$

$$y_{2i}^{apply} = (x_{2i}\beta_2 + u_{2i} > 0) \,, \tag{8}$$

under similar assumptions to those in Model 1.

4.2. Heckman Probit models with spatial effects

In this section, we augment the previous Heckit models to include spatially lagged explanatory variables to account for the student's neighborhood and to address the endogeneity problem caused by spatial autocorrelation. We use GeoDa software ([19]) to estimate the spatial lag variables and Stata's 'heckprob' command to estimate the Heckman probit models with sample selection.

More specifically, we consider two ways of including spatial effects in the Heckit model. First, we show that Moran's *I* test is calculated on the residuals of the Heckit model. Second, we examine the role of localized social interactions between "neighbors" (nearby in a spatial sense) that occur in the context of information exchange and social context related to participation in higher education.

4.2.1. Endogeneity issues and spatial autocorrelation test of the residuals

An endogeneity problem may arise because individual students who live in the same socio-spatial setting (social space) may act in a similar way because they share common unobservable factors or institutional environments ([20]). This phenomenon creates spatial dependence that reflects a situation in which a given student's values may be contingent on the values of students living nearby ([21]).

We thus calculate the Moran's I test statistic to assess the presence of spatial error autocorrelation in the Heckit model. The general form of Moran's I is given by:

$$I = \frac{Q^*}{\tilde{\sigma}_{Q^*}} , \tag{9}$$

where:

$$Q^* = \hat{u}_n' \mathbf{W}_n \hat{u}_n , \tag{10}$$

in which $\hat{u}_n$ is the n × 1 vector of the generalized residual of the Heckit model; **W** is the familiar n × n spatial weights matrix, which reflects the vicinity relations among the n spatial observations, where the main diagonal is equal to zero by convention; and $\tilde{\sigma}_{Q_n^*}$ is a normalizing factor ([22]). The generalized residual values of the Heckit model are calculated as follows:

$$\hat{u}_i = y_{1i} - x_{1i}'\hat{\beta}_1 + \frac{\hat{\sigma}_{12}}{\hat{\sigma}_2^2}\left\{\frac{\phi(x_{2i}'\hat{\beta}_2)}{\Phi(x_{2i}'\hat{\beta}_2)}\right\} \text{ for } i = 1,2,\dots,n , \tag{11}$$

where $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\sigma}_{12}$ are the maximum likelihood estimates of the variable parameters and inter-equation residual covariance, respectively; and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution functions of the standard normal distribution. The term in curly brackets, known as the inverse Mills ratio, coincides with the generalized residual of the probit model ([23]).

4.2.2. A spatial Heckit model

The spatial version of Heckit Model 1 (university application) takes the form of a cross-sectional spatially lagged SLX model ([24]; see also [25,26]). This model incorporates an augmented outcome equation to account for the latent spatial structure of the decision-making process, reflected by y_{1i}, given by:

$$y_{1i}^{apply} = (x_{1i}\beta_1 + (\mathbf{W}x_{1i})\gamma_1 + u_{1i} > 0) , \tag{12}$$

where x_{1i} contains the usual explanatory variables (as in the standard Heckit model), **W** is the spatial weights matrix indicating "nearest neighbors", $\mathbf{W}x_{1i}$ are the spatial lagged variables representing local "spillovers", and γ_1 is an additional vector of unknown parameters to capture interaction spatial effects ([21]).

Local spatial spillovers are appropriate when the proper spatial range of the explanatory variables is the location and its immediate neighbors (but not beyond); that is, the range of neighbors considered in the reference space—for example, only direct neighbors, not neighbors' neighbors ([27]). This concept is in line with [12]'s claims, for whom the nearby environment of the students' address (family, school, neighborhood) constitutes its main spatial contextual reference.

Similarly, we extend the selection equation in Model 1 by including an SLX term, given by:

$$y_{2i}^{pre-select} = (x_{2i}\beta_2 + (\mathbf{W}x_{2i})\gamma_2 + u_{2i} > 0), \tag{13}$$

For Heckit Model 2, university application becomes the selection criterion. We thus extend the selection equation to include only SLX terms, not the outcome equation, because admission depends on the decision of the DEMRE (not of the student):

$$y_{1i}^{admit} = (x_{1i}\beta_1 + u_{1i} > 0) , \tag{14}$$

$$y_{2i}^{apply} = (x_2\beta_2 + (\mathbf{W}x_{2i})\gamma_2 + u_{2i} > 0) , \tag{15}$$

5. Estimation results

5.1. Baseline models

This section presents the results for two baseline models.

- Baseline Model 1

1. Main equation:

$$\begin{aligned}\Pr(APPLICATION = 1|PRE\text{-}SELECTION = 1) =\\ &= \beta_0 + \beta_1 FEMALE + \beta_2 Log(LIT\_SCORE)\\ &+ \beta_3 Log(MATH\_SCORE) + \beta_4 Log(GRADE\_PTS)\\ &+ \beta_5 Log(DISTANCE) + \beta_6 [Log(DISTANCE)]^2\\ &+ \beta_7 WORKING + u_1 ,\end{aligned} \tag{16}$$

2. Selection equation:

$$\begin{aligned}\Pr(PRE\text{-}SELECTION = 1)\\ &= \gamma_0 + \gamma_1 FEMALE + \gamma_2 Log(GRADE\_PTS)\\ &+ \gamma_3 RURAL + \gamma_4 Log(DISTANCE) + \gamma_5 WORKING\\ &+ \gamma_6 SIBL_{UNIV} + \gamma_7 SOCIAL\_CAP + u_2 ,\end{aligned} \tag{17}$$

This first baseline model uses the variables "RURAL", “SIBL_UNIV”, and "SOCIAL_CAP" as exclusion criteria (instruments) that correlate with selection (“PRE-SELECTION”) but not with the binary outcome in the main equation (“APPLICATION”).

- Baseline Model 2

1. Main equation:

$$\begin{aligned}\Pr(ADMISSION = 1|APPLICATION = 1) =\\ &= \beta_0 + \beta_1 FEMALE + \beta_2 Log(LIT\_SCORE)\\ &+ \beta_3 Log(MATH\_SCORE) + \beta_4 Log(GRADE\_PTS) + u_1 ,\end{aligned} \tag{18}$$

2. Selection equation:

$$\begin{aligned}\Pr(APPLICATION = 1)\\ &= \gamma_0 + \gamma_1 FEMALE + \gamma_2 Log(LIT\_SCORE)\\ &+ \gamma_3 Log(MATH\_SCORE) + \gamma_4 Log(GRADE\_PTS) + \gamma_5 RURAL\\ &+ \gamma_6 Log(DISTANCE) + \gamma_7 [Log(DISTANCE)]^2\\ &+ \gamma_8 WORKING + \gamma_9 SOCIAL\_CAP + u_2,\end{aligned} \tag{19}$$

This model uses the variables "RURAL", "DISTANCE", "WORKING", and "SOCIAL_CAP" as exclusion criteria (instruments) that correlate with selection (“APPLICATION”) but not with the binary outcome in the main equation (“ADMISSION”).

Table 2 shows the results of the baseline models. For both models, we found a statistically significant selection of unobserved factors (non-negative correlation between the errors of the main and selection equations). That is, (i) applicants to university are systematically different from the students who did not pass the pre-selection test (Model 1), and (ii) applicants to university who are granted admission to higher education are likely to be systematically different from students who did not apply to university.

In the application model, as expected, the variables with considerable influence on the probability of getting into university are the test results, and the mathematics test has the most significant effect on the probability that a student will apply to university. Grades are another variable with considerable influence on the probability of a student applying to university.

Students’ characteristic variables are also important to explaining the probability of both being preselected and applying to university. Women are less likely to pass the PSU, but once they have passed this exam, they are more likely than men to apply to university. After the application process, however, the probability of being accepted into higher education is significantly lower for women.

Table 2. Estimated coefficients from Heckman probit – Baseline models.

	Baseline Model 1	Baseline Model 2

A: Main equations	Pr(APPLICATION=1 | PRE-SELECTION=1)		Pr(ADMISSION=1 | APPLICATION=1)	
FEMALE	0.176***	(22.1)	-0.248***	(-35.6)
Log(LIT_SCORE)	2.054***	(60.6)	0.886***	(34.9)
Log(MATH_SCORE)	2.488***	(67.8)	1.257***	(50.1)
Log(GRADE_PTS)	0.625***	(16.0)	0.585***	(25.3)
WORKING	-0.155***	(-12.1)	–	–
DISTANCE	0.056***	(27.5)	–	–
DISTANCE squared	-0.002***	(-15.3)	–	–
Constant	-31.924***	(-82.6)	-16.118***	(-82.2)
B: Selection equations	Pr(PRE-SELECTION=1)		Pr(APPLICATION=1)	
FEMALE	-0.256***	(-44.8)	0.144***	(24.4)
Log(LIT_SCORE)	–	–	2.700***	(139.9)
Log(MATH_SCORE)	–	–	2.069***	(111.5)
Log(GRADE_PTS)	3.470***	(196.9)	1.210***	(62.2)
DISTANCE	-0.019***	(-28.9)	0.041***	(28.4)
DISTANCE squared	–	–	-0.001***	(-12.5)
RURAL	-0.573***	(-29.2)	-0.170***	(-9.4)
WORKING	-0.056***	(-6.2)	-0.078***	(-9.4)
SIBL_UNIV	0.170***	(26.7)	–	–
SOCIAL_CAP	0.196***	(138.7)	0.028***	(23.0)
Constant	-21.252***	(-193.9)	-37.220***	(-259.9)
Athrho	-0.267***	(-13.8)	-2.261***	(-61.8)
Rho	-0.261***		-0.979***	
No. of observations	260,775		260,775	
No. of censored observations	103,398		120,256	
No. of uncensored observations	157,377		140,519	
Likelihood Ratio test	206.4***		7,603.0***	

Note: *t* statistics in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001.

Additionally, working students are less likely to pass the PSU and apply to university. This phenomenon could reflect lower priority and/or interest in educating these candidates or higher opportunity cost of accessing the university.

As to familial variables, students with siblings at university are more likely to pass the PSU, as are those with higher levels of social capital (a composite variable of parental education and family income). Candidates living in rural areas are less likely to pass the PSU.

Finally, distance to the Santiago plays a different role in the pre-selection and application equations. On the one hand, the probability of passing the PSU declines linearly with distance to Santiago, but once students pass this exam, there is a non-linear positive relation between distance to Santiago and probability of applying to university (**Figure 4**).

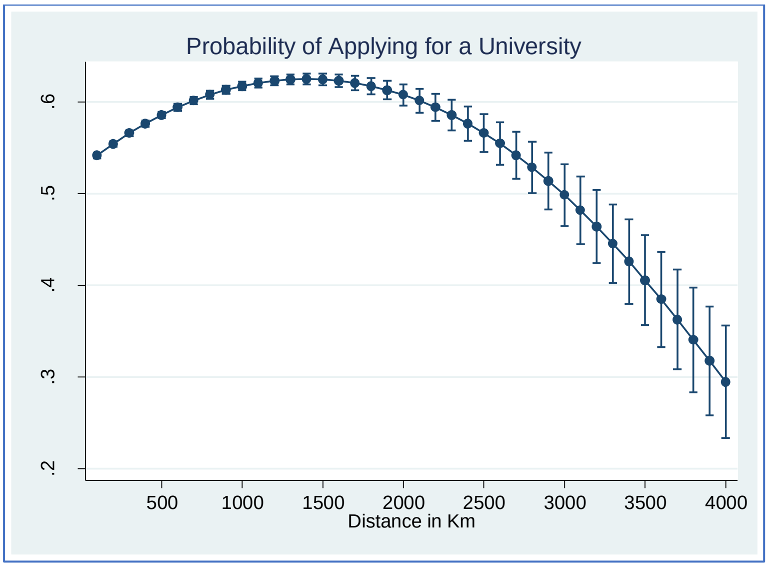


Figure 4. Effect of distance from students' home to Santiago city on the probability of applying to university.

Students living in peripheral areas are viewed as more likely to apply to university. The distance variable shows an inverted U effect on applying to the university. That is, when distance to Santiago increases, the probability of applying to university increases to a threshold of 1,400 kilometers (Antofagasta region, in the North and Aysén region, in the South), at which point it begins to decrease. Additionally, the probability of applying is increasingly variable, from the distance threshold to the country limits. This variability can be explained by the many private universities concentrated in the Metropolitan Region. The private universities offer their residents more opportunities to access higher education once they pass the PSU than are available to those living farther from Santiago, who opt to apply to the traditional higher education system.

5.2. Spatial models

5.2.1. Specification of the spatial weights matrix

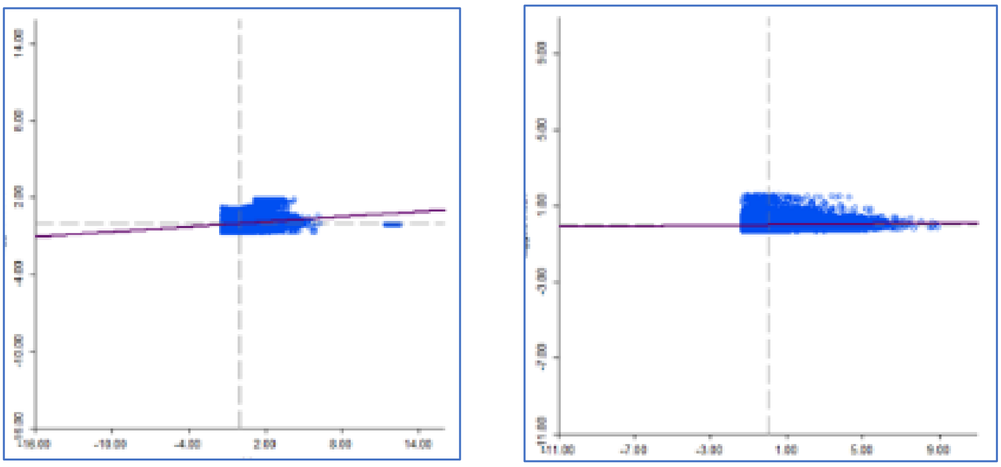
Based on the spatial distribution of the individual students in Chile, we identify spatial neighborhoods to be used in the construction of the spatial weights matrix. We use an exploratory approach to characterize the structure of the spatial weights matrix **W**. More specifically, based on the addresses of all the students, we performed Thiessen polygonization to define the spatial contiguity within the neighborhood. The most frequent number of neighbors was three, covering around 41% of the total number of students. Taking this into consideration, we specified three different **W** matrices: (i) dispersed matrix (few neighbors): 3 neighbors; (ii) dense matrix (more neighbors): 100 neighbors; (iii) very dense matrix: 300 neighbors.

Table 3. Results of Moran's I test on the baseline model residuals.

Baseline model	Nearest neighbor #	Moran's *I*	z-Value	Pseudo p-value
Model 1	3	0.082	61.0	0.001
	100	0.074	276.8	0.001
	300	0.066	453.9	0.001
Model 2	3	0.014	10.2	0.001
	100	0.015	57.0	0.001
	300	0.011	73.1	0.001

Table 3 shows the results from the Moran's *I* test of the residuals of the baseline models—equations (9) and (10)—using an inferential process based on the permutation approach (9,999 permutations). Moran's I is statistically significant for the three types of **W** matrices, and the dense matrix shows the most significant *z*-value.

Despite the high significance of the tests, with *pseudo* p-values at 0.001, the slope of the regression line in the Moran scatterplot exhibits a weak and uniform pattern of spatial association. The Moran scatter plot was first outlined in [28]. It consists of a plot with the spatially lagged variable on the *y*-axis and the original variable on the *x*-axis. The slope of the linear fit to the scatter plot equals Moran's *I*. **Figure 5** represent these plots for the dense and most significant spatial weights matrix, W_{300}.



(a) (b)

Figure 5. Moran's I test on the model residuals for W_{300}.

Spatial autocorrelation analysis of the baseline model residuals thus demonstrates the existence of statistically significant spatial neighborhood effects influencing the university access process in Chile, especially in the vicinity of 300 neighbors. This spatial effect is not very strong for the global spatial structure, however, suggesting the existence of local spatial autocorrelation—spillovers—in this phenomenon. As stated above, the SLX model estimates spatial local spillovers, which are suitable for our purpose here. Additionally, [29] pointed out that when spatial dependence is weak, the best fitting specification might be the SLX model.

5.2.2. SLX Heckit model results

Next, we present the specification of the SLX Heckit models for a spatial weights matrix of 300 neighbors (W_{300}).

- Spatial Model 1
  1. Main equation:

$$\begin{aligned} &\Pr(APPLICATION = 1|PRE\text{-}SELECTION = 1) = \\ &\quad = \beta_0 + \beta_1 FEMALE + \cdots + \beta_8 \mathbf{W}_{300} Log(GRADE\_PTS) \\ &\quad + \ \beta_{10} \mathbf{W}_{300} WORKING + u_1 \ , \end{aligned} \tag{20}$$

  2. Selection equation:

$$\begin{aligned} &\Pr(PRE\text{-}SELECTION = 1) \\ &\quad = \gamma_0 + \gamma_1 FEMALE + \cdots + + \gamma_8 \mathbf{W}_{300} SIBL_{UNIV} \\ &\quad + \gamma_9 \mathbf{W}_{300} SOCIAL\_CAP + u_2 \ , \end{aligned} \tag{21}$$

This model uses the statistically significant variables "SIBL_UNIV" and "SOCIAL_CAP" as exclusion criteria (instruments) that correlate with selection ("PRE-SELECTION") but not with the binary outcome in the main equation ("APPLICATION").

- Spatial Model 2
  1. Main equation:

$$\Pr(ADMISSION = 1|APPLICATION = 1) \\ = \text{No spatially lagged variables included}\ , \tag{22}$$

  2. Main equation:

$$\Pr(APPLICATION = 1) \\ = \gamma_0 + \gamma_1 FEMALE + \cdots + \gamma_{10}\mathbf{W}_{300}WORKING \\ + +\gamma_{11}\mathbf{W}_{300}SOCIAL\_CAP + u_2, \tag{23}$$

This model uses the statistically significant variables "WORKING" and "SOCIAL_CAP" as exclusion criteria (instruments) that correlate with selection ("APPLICATION") but not with the binary outcome in the main equation ("ADMISSION").

Table 4 shows the main outcomes of the SLX Heckit models. In spatial model 1, each candidate's social capital and siblings already enrolled in the university have a positive effect on the probability of successfully passing the PSU. Additionally, the spatial neighborhood of the 300 nearest candidates leverages the positive impact of these variables. That is, the existence of nearby applicants with high levels of social capital and siblings at the university influences a candidate positively to pass the PSU. Hence, a "good" social environment matters for success on the PSU.

Having good high-school grades and not having a job are significant in explaining candidates' probability of both passing the PSU and applying to university. In this case, however, students are also affected by their closest neighbors' performance in high school and professional situation. The spatial effect of the variable "WORKING" in particular is significantly higher, indicating that the presence in a student's neighborhood of many high-school peer graduates already working will discourage him/her to apply to university after passing the PSU.

In spatial model 2, social capital also has a significant positive effect on a student's probability to apply to university, and having a job has a significant negative effect, as shown above. In this case, the candidates' spatial neighborhood will influence their decision to apply to university through social capital and professional situation. As to ultimate acceptance to a university and a career, spatial effects are not relevant, since this decision must be taken exclusively by the university.

A student's spatial vicinity is thus crucial to ensuring that a candidate both pass the PSU exam and apply to university. Specifically, four variables have local spillovers: social capital, professional situation, having siblings studying at a university, and having good marks in high school. The best environment for a student to succeed in accessing higher education in Chile includes peers who got good marks in high school and do not have a job, and high social capital and siblings already studying at a university. Conversely, the worst neighborhood for a higher education candidate includes peers who got bad grades in high school and have a job, and lack of good social capital status or of siblings at university. We would also add other environmental variables with a negative impact on higher education accessibility, such as living in rural areas and/or ultra-peripheral regions, over 1,400 km. far from Santiago. It is thus important for the state to foster policies that motivate students living in working-class neighborhoods to see the university as a valid option for their personal advancement. Students who live in environments with good secondary schools, where students earn good grades, are more likely to apply than others. This finding strengthens support for the above-mentioned need to foster public policies to improve the quality of secondary education in the most disadvantaged neighborhoods.

The results indicate a problem of gender inequality. Although women's likelihood of applying to university is higher than men's, their probability of ultimately being accepted is significantly lower. This is clearly a serious problem that merits further in-depth study to see whether bias exists in the model used or in the type of education women receive. The effect of the candidates' social capital, while statistically significant, does not seem as relevant as initially expected, perhaps due to its correlation with the effect of the variable of student grade-point average in secondary education and score on the PSU.

Table 4 Estimated coefficients of the spatial Heckman probit models.

	Spatial Model 1		Spatial Model 2	
A: Main equations	Pr(APPLICATION=1 | PRE-SELECTION=1)		Pr(ADMISSION=1 | APPLICATION=1)	
FEMALE	0.177***	(22.3)	-0.248***	(-35.6)
Log(LIT_SCORE)	2.047***	(60.2)	0.886***	(34.9)
Log(MATH_SCORE)	2.455***	(66.6)	1.258***	(50.1)
Log(GRADE_PTS)	0.602***	(15.4)	0.587***	(25.3)
WORKING	-0.145***	(-11.3)	–	–
DISTANCE	0.052***	(24.1)	–	–
DISTANCE squared	-0.002***	(-13.1)	–	–
Spatial lag W_{300}WORKING	-0.799***	(-6.2)	–	–
Spatial lag W_{300}Log(GRADE_PTS)	0.367**	(3.2)	–	–
Constant	-33.747***	(-40.1)	-16.138***	(-82.3)
B: Selection equations	Pr(PRE-SELECTION=1)		Pr(APPLICATION=1)	
FEMALE	-0.256***	(-44.8)	0.144***	(24.4)
Log(LIT_SCORE)	–	–	2.698***	(139.6)
Log(MATH_SCORE)	–	–	2.061***	(110.8)
Log(GRADE_PTS)	3.491***	(197.1)	1.207***	(61.8)
DISTANCE	-0.020***	(-30.2)	0.040***	(26.1)
DISTANCE squared	–	–	-0.001***	(-11.6)
RURAL	-0.537***	(-27.4)	-0.168***	(-9.3)
WORKING	-0.065***	(-7.2)	-0.076***	(-9.1)
SIBL_UNIV	0.160***	(25.1)	–	–
SOCIAL_CAP	0.175***	(114.4)	0.0236***	(17.4)
Spatial lag W_{300}WORKING	–	–	-0.396***	(-5.1)
Spatial lag W_{300}SOCIAL_CAP	0.048***	(10.5)	0.020***	(7.3)
Spatial lag W_{300}SIBL_UNIV	1.256***	(18.2)	–	–
Constant	-21.746***	(-193.2)	-37.099***	(-257.7)
Athrho	-0.269***	(-13.5)	-2.252***	(-62.1)
Rho	-0.262***		-0.978***	
No. of observations	260,775		260,775	
No. of censored observations	103,398		120,256	
No. of uncensored observations	157,377		140,519	
Likelihood Ratio test	195.7***		7,600.6***	
Likelihood Ratio test – spatial [d.f. = 4]/ [d.f. = 2]	1,795.6***		97.0***	

Distance to Santiago seems to have an inverted U-shaped effect, such that the greater the distance between students and Santiago, the greater students' probability of applying to and being accepted at university. This trend continued up to a threshold distance of 1400 kilometers, or to the regions at the geographical extremes of Chile—region II in the north and region XI in the south. As stated above, the best university education system is clearly concentrated in the central region of Chile, as are many private universities that do not form part of the PSU admission system and to which students can apply if they are not admitted to the traditional universities. This U shape again demonstrates the importance of improving the quality of secondary school instruction in the ultra-peripheral Chilean regions, especially the rural ones, to increase the rates at which students in these regions enter the university.

6. Conclusions

This paper has two main objectives, implementation of a spatial Heckman probit model and validation of the importance of geography to education. This paper uses a discrete sample selection model augmented with local spatial autocorrelation effects to explore the role of spatial interaction role of Chile's educational economy.

The model was respecified as a spatial cross-regressive (SLX) model, which adds the spatially lagged explanatory variables to the standard Heckit model. This model can absorb and explain the effect of spatial dependence or proximity among the students on their probability of accessing higher education. This model can be estimated directly using the maximum likelihood method appropriate to the Heckit model, given the exogenous character of the lagged spatially explanatory variables and the model's suitability for explaining local spatial externalities even when they occur with weak intensity, as in this model.

The application process to enter the universities has been explained in two parts, or models: the first is based on students' decision to take the PSU test; this first step leads to the second, the DEMRE's decision on the student's application.

Our hope is that this paper will serve as a starting point for future research that estimates a SAR-Heckit model to quantify global spatial autocorrelation effects, their direct and indirect effects at national level, and new specifications of the spatial weights matrix based on social networks or socioeconomic factors.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, J.Q., L.P. C.C. and P.A.; methodology, J.Q., L.P., C.C. and P.A.; software, J.Q., L.P. and C.C.; validation, J.Q., L.P., C.C. and P.A.; formal analysis, J.Q., L.P. and C.C.; investigation, J.Q. and L.P.; resources, J.Q.; data curation, J.Q. and C.C.; writing—original draft preparation, J.Q. and L.P.; writing—review and editing, C.C.; project administration, J.Q.; funding acquisition, C.C. and P.A. All authors have read and agreed to the published version of the manuscript."

Funding: This research was funded by FONDECYT Project, grant number 1171230 and eS-MiData Project, grant number 447379001.

Data Availability Statement: Data and scripts are available in the Spatial and Regional Economics research group (ECONRES) of the Universidad Autónoma de Madrid at https://econresuam.wordpress.com/opendata.

Acknowledgments: We are grateful to participants at the 9th Seminar of Spatial Econometric Jean Paelinck | 6-7 October 2017, Cartagena, Spain and the 2021 RSAI World Conference | 25-28 May 2021, Virtual Conference, for their valuable comments on several versions of this paper. We also acknowledge Andrés Vallone's contribution with the postal address geocoding process.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Espinoza, O. Creating (in) equalities in access to higher education in the context of structural adjustment and post-adjustment policies: The case of Chile. *High Educ* **2008**, 55(3), 269-284. https://doi.org/10.1007/s10734-007-9054-8
2. Koljatic, M.; Silva, M.; Cofré, R. Achievement versus aptitude in college admissions: A cautionary note based on evidence from Chile. *Int J Educ Dev* **2013**, 33(1), 106-115. https://doi.org/10.1016/j.ijedudev.2012.03.001
3. OECD. Access and Equity. In *Reviews of National Policies for Education: Tertiary Education in Chile, 2009*; OECD Publishing: Paris, France, 2009; pp. 73-121. https://doi.org/10.1787/9789264051386-5-en
4. Pearson. *Final Report Evaluation of the Chile PSU*, 22nd of January 2013.
5. Castells, M. *La era de la información* (Vol. I). Alianza: Madrid, Spain, 2001.
6. Casanova Cruz, D. Equity of access to higher education: The "Class Rank" as a mechanism of inclusion in the Chilean admission system. *Education Policy Analysis Archives* **2015**, 23, p. 72. https://doi.org/10.14507/epaa.v23.1908

7. Brunner, J. SIES: Tres preguntas y la responsabilidad de las universidades. In *La Segunda*, 19th of June 2002; p. 14.
8. Orellana, M.; Moreno, K. *Inclusión a la universidad de estudiantes meritorios en situación de vulnerabilidad social*. UNESCO, 2015.
9. Looker, E.D.; Lowe, G.S. Post-secondary access and student financial aid in Canada. *Canadian Policy Research Networks Inc.* **2001**, 14, 1-12.
10. Finnie, R.; Wismer, A.; Mueller, R.E. Access and barriers to postsecondary education: Evidence from the youth in transition survey. *Canadian Journal of Higher Education* **2015**, 45(2), 229-262.
11. Coleman, J.S. Social capital in the creation of human capital. *Am J Sociol* **1988**, 94, S95-S120.
12. Furstenberg, F.F.; Hughes, M.E. Social capital and successful development among at-risk youth. *J Marriage Fam* **1995**, 57, 580-592. https://doi.org/10.2307/353914
13. Madrid, S. "Diversidad sin diversidad": Los colegios particulares pagados de élite y la formación de la clase dominante en una sociedad de mercado. In *Mercado escolar y oportunidad educacional. Libertad, diversidad y desigualdad*, Corvalán, J., Carrasco, A., García-Huidobro, J.E., Eds.; Colección Estudios en Educación, Ediciones UC, Santiago, Chile; 2016; Chapter 9; pp. 269-299.
14. van de Ven, W.P.M.M.; van Praag, B.M.S. The demand for deductibles in private health insurance: A probit model with sample selection. *J Econometrics* **1981**; 17, 229-252. https://doi.org/10.1016/0304-4076(81)90028-2
15. Wooldridge, J.M. *Econometric Analysis of Cross Section and Panel Data, Second Edition*. The MIT press, Cambridge, Massachusetts/London, England; 1981; pp. 813-814.
16. Pastore, F. To study or to work? Education and labor market participation of young people in Poland. *Eastern Eur Econ* **2012**; 50(3), 49-78. https://doi.org/10.2753/EEE0012-8775500303
17. Ahlin, L.; Andersson, M.; Thulin, P. Human capital sorting: The "when" and "who" of the sorting of educated workers to urban regions. *J Regional Sci* **2018**; 58(3), 581-610. https://doi.org/10.1111/jors.12366
18. Morrissey, K.; Kinderman, P.; Pontin, E.; Tai, S. Web based health surveys: Using a two-step Heckman model to examine their potential for population health analysis. *Soc Sci Med* **2016**; 163, 45-53. https://doi.org/10.1016/j.socscimed.2016.06.053
19. Anselin, L. *GeoDa. An Introduction to Spatial Data Science*. https://geodacenter.github.io (accessed on 24 / 11 / 2021).
20. Ioannides, Y.A.; Topa, G. Neighborhood effects: Accomplishments and looking beyond. *J Regional Sci* **2010**, 50(1), 343-362. https://doi.org/10.1111/j.1467-9787.2009.00638.x
21. LeSage, J.; Pace, R.K. *Introduction to spatial econometrics*. Chapman and Hall/CRC, Boca Raton, FL, U.S.; 2009.
22. Kelejian, H.H., Prucha, I.R. On the asymptotic distribution of the Moran *I* test statistic with applications. *J Econometrics* **2001**; 104(2), 219-257. https://doi.org/10.1016/S0304-4076(01)00064-1
23. Vella, F. Estimating models with sample selection bias: A survey. *J Hum Resour* **1998**, 33(1), 127-169. https://doi.org/10.2307/146317
24. Halleck Vega, S.M.; Elhorst, P. The SLX model. *J Regional Sci* **2015**, 55(3), 339-363. https://doi.org/10.1111/jors.12188
25. Fischer, M.M.; Scherngell, T.; Reismann, M. Knowledge spillovers and total factor productivity: Evidence using a spatial panel data model. *Geogr Anal* **2009**, 41(2), 204-220. https://doi.org/10.1111/j.1538-4632.2009.00752.x
26. Chasco, C.; Le Gallo, J. Hierarchy and spatial autocorrelation effects in hedonic models. *Economics Bulletin* **2012**, 32(2), 1474-1480.
27. Anselin, L. Spatial externalities, spatial multipliers, and spatial econometrics. *Int Regional Sci Rev* **2003**, 26(2), 153-166. https://doi.org/10.1177/0160017602250972
28. Anselin, L. The Moran Scatterplot as an ESDA Tool to assess local instability in spatial association. In *Spatial Analytical Perspectives on Gis in Environmental and Socio-Economic Sciences*; Fischer, M., Scholten, H., Unwin, D., Eds.; Taylor & Francis, London, U.K., 1996; pp. 111–25. https://doi.org/10.1201/9780203739051-8
29. LeSage, J. What regional scientists need to know about spatial econometrics. *Review of Regional Studies* **2014**, 44(1), 13-32. https://doi.org/10.52324/001c.8081