UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

IMPLEMENTATION OF THE MATCHING MECHANISM FOR THE NEW SCHOOL ADMISSIONS SYSTEM AND MODELING OF THE SCHOOL CHOICE FOR CHILEAN FAMILIES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

NICOLÁS ANDRÉS ARAMAYO BENVENUTTO

PROFESOR GUÍA:
MARCEL GOIC FIGUEROA

MIEMBROS DE LA COMISIÓN:
JOSÉ CORREA HAEUSSLER
RICARDO MONTOYA MOREIRA

SANTIAGO DE CHILE
2018

ABSTRACT
BY: NICOLÁS ANDRÉS ARAMAYO BENVENUTTO
DATE: NOVEMBER 2018

IMPLEMENTATION OF THE MATCHING MECHANISM FOR THE NEW SCHOOL ADMISSIONS SYSTEM AND MODELING OF THE SCHOOL CHOICE FOR CHILEAN FAMILIES

Matching mechanisms for school assignment have been adapted on a global scale since the popular application in New York City in 2004, which used the Deferred Acceptance algorithm to assign thousands of students. Implementation of these systems is not a trivial task because of the large scale of the problem and educational regulations particular to each country. One of the goals of this thesis was to develop the implementation of the matching algorithm for the Chilean case and to show the nuances in the design of this mechanism. Then, taking advantage of the strategy-proofness of this system and the centralized data it produces, this work develops structural models for discrete choice to study families' preferences for schools, specifically, a Bayesian multivariate ordered probit with a hierarchical Bayesian structure to model heterogeneity in the school choice. Also, a methodology was developed to obtain choice sets for families using unsupervised learning techniques in the Coquimbo region where these structural models could be applied. This way, meta-analysis was conducted to evaluate what characteristics in school choice are consistently significant, where results indicate that, for example, while schools with poor academic performance are not preferred on average, schools with superior results on standardized test are only preferred by students with good academic records. The estimation of these preferences allows for a series of counterfactual analyses that can aid in the design and implementation of public policies and support the decision making of schools. A no-pricing policy was simulated for the Coquimbo region—which is actually in the process of being adopted by subsidized private schools in Chile—, where it was found that it would improve the social welfare of the assignment of the region, specially for families with children with disabilities, but would impact negatively students that are not in the lowest socioeconomic family groups.

ABSTRACT
BY: NICOLÁS ANDRÉS ARAMAYO BENVENUTTO
DATE: NOVEMBER 2018

IMPLEMENTATION OF THE MATCHING MECHANISM FOR THE NEW SCHOOL ADMISSIONS SYSTEM AND MODELING OF THE SCHOOL CHOICE FOR CHILEAN FAMILIES

Mecanismos de matching para la asignación escolar han sido adaptados a escala global desde la popular aplicación en la ciudad de Nueva York el 2004, la cual usó el algoritmo de Aceptación Diferida para asignar a miles de estudiantes. La implementación de estos sistemas no es una tarea trivial debido a la gran escala del problema y las regulaciones educacionales de cada país. Uno de los objetivos de esta tesis fue desarrollar la implementación del algoritmo de matching para el caso Chileno y mostrar los matices en el diseño del mecanismo. Luego, usando como ventaja que el sistema es strategy-proof y produce datos centralizados, este trabajo desarrolla modelos estructurales de elección discreta para estudiar las preferencias de las familias por colegios, específicamente, un probit multivariado ordenado Bayesiano con una estructura jerárquica Bayesiana para modelar la heterogeneidad en la elección de colegios. También, se desarrolla una metodología para obtener conjuntos de elección mediante técnicas de aprendizaje no supervisado en la región de Coquimbo en donde estos modelos estructurales pudiesen ser aplicados. De esta forma, se condujo un meta análisis para evaluar qué características son consistentemente significativas, en donde los resultados indican que, por ejemplo, mientras que los colegios con un bajo desempeño académico no son preferidos en promedio, colegios con resultados más altos en pruebas estandarizadas son preferidos sólo por estudiantes con buen historial académico. La estimación de estas preferencias permite realizar una serie de análisis contrafactuales que pueden ayudar en el diseño e implementación de políticas públicas y apoyar la toma de decisiones de los colegios. Una política de precio cero fue simulada para la región de Coquimbo—que de hecho está en el proceso de ser adoptada por los colegios particulares subvencionados en Chile—, en donde se obtuvo que produciría una mejora en el bienestar social de la asignación de la región, en particular para las familias con hijos con discapacidades, pero impactaría negativamente a estudiantes que no están en los estratos socioeconómicos familiares más bajos.

To a couple of faculty professors who told me I would never use what I learned at the University in a real world problem

Acknowledgements

I would like to thank my thesis professor, Dr. Marcel Goic, for his guidance and counsel, which allowed me to learn more than what I could've pictured from this experience, and for helping me develop something that I feel proud of, not to mention that he always showed the greatest willingness to answer all my—so many—questions every week for more than a year, even as they often ranged out of the scope of this work.

Thanks also to my friends and colleagues, Mario and Luis, for always providing help and support when it was needed.

Finally, I would like to thank my family. My brothers Hugo and Matías, my sister Francisca, and especially to both my parents, Jacqueline and Hugo. Your endless support and caring is always noted and deeply appreciated.

Table of Contents

List of Tables

List of Figures

Introduction

Matching mechanisms for school assignment consist in the use of a matching algorithm to produce an assignment between a number of school and students who wish to be assigned in some of them. This problem is not trivial, since students may have a strict order of preferences for the schools, and the schools may prefer some students over others, and the matching must satisfy some criterion to produce an outcome that aims to be optimal in this respect. And the choice between algorithms that have been popular for solving this matching problem also carries some trade-offs. The Deferred Acceptance algorithm produces an optimal stable matching, while also being a strategy-proof system (i.e. no agent can benefit from misrepresenting their preferences), meanwhile the Top Trading Cycles algorithm is Pareto efficient and strategy-proof, therefore it must be decided between which properties are more desirable for a particular educational system.

These centralized systems also produce benefits beyond the properties of the matching algorithms. In particular for the case of Chile, the new school admissions system ends large queues that formed every year outside of public schools all across the country by parents that wished to ensure a seat for their children. Also, the strategy-proofness of the system implies that the families obtain the greater benefit from the system if they submit their true preferences for schools in the system. This would produce a large and presumably clean source of preference data that was never before available in Chile, because the system was not centralized. This allows for researchers to make use of this data to understand what families—and what types of families—value and do not like about schools in different regions in the country. This also does not come without its difficulties. The particular structure of the data, censored strict ordered rankings, do not allow for state-of-the-art structural models to model the data generating process. In this work, models are developed to accommodate for these constraints while also comparing them with previous structural models often used in discrete choice decision problems, while taking advantage of a fully Bayesian specification to model the uncertainty about the decision making via probability distributions.

Modeling the choices is interesting by itself because it allows for understanding the underlying processes that make families decide between schools, specially accounting for heterogeneity between families and explicitly modeling school's attributes. But this also allows for counterfactual analyses that can aid in improving the system, the design and improvement of public policies, and support the efficient use of some resources of the schools. Any improvement on the aforementioned policies can impact positively from hundreds to millions of lives for the coming years.

Chapter 1

School Assignment by the Deferred Acceptance Algorithm

This first chapter describes the context under which the educational reforms were driven in Chile with respect to school assignment for establishments that receive support from the State, and also showing how similar reforms to the school admission mechanisms have been driven around the world. Later, the Deferred Acceptance algorithm is introduced along with an explanation of why it is appropriate for solving the school admission problem, showing how it fulfills the reform's goals, in addition to having desirable properties for the assignment it produces. Afterwards, the particularities and details of the algorithm's implementation for the 2017 admission process are shown, and how it was used for 5 of the country's regions satisfying multiple criteria of quotas and preferences that were designated in the educational reform.

1.1 Historic Context

Mechanisms for school assignment have been being implemented for years on a global scale before Chile adopted it by the means of the new educational reform. The pioneer city in which the system was implemented—and Chile attempted to replicate—was New York in the year 2004 for 8th graders. Since then, this system has been acclaimed by both the public and the scientific community, even contributing to the Memorial Nobel Prize in Economic Sciences won by Alvin Roth (one of the professors invited to solve the assignment problem in New York) in 2012. The cities that followed implementing the system were Camden, Denver, New Orleans, Newark, Washington DC, Boston, and Amsterdam.

The new Chilean educational reform, which has been being deployed since 2015 with a series of legislative projects, has as one of its main pillars ensuring the end of arbitrary selection by educational establishments in preschool, primary and high schools that receive support from the State, as well as ending profiting in these establishments and ensuring a free education for everyone. With this spirit in mind, the Deferred Acceptance algorithm (Gale and Shapley, 1962) has been being progressively implemented for the whole country, ending a mechanism that so far had been uncoordinated (given that families applied separately to every schools and the applications were not informed between schools), making way for a coordinated system (given that it outputs a unique assignment to the students taking into account the available seats of the schools and the applications altogether) in which families apply through a centralized online system so that students are assigned to a school through the formerly mentioned algorithm.

In 2016, the first pilot was put to test in the Magallanes region, assigning over 4,000 students of prekinder, kinder, first, seventh and ninth grade, where more than 58% of them were assigned to their first preference. In 2017, the system was deployed for the same levels in the Tarapacá, Coquimbo, O'Higgins and Los Lagos regions, and in every level for the region of Magallanes, assigning almost 77,000 students in over 2,100 schools, where more than 56% of students were assigned to their first preference. By the year 2019, it is expected that the system will work in all of the country's regions, and it is estimated that around 500,000 students will participate through this admission system.

1.2 Properties of the Assignment Algorithm

The algorithm that was used originally to produce the assignments in the city New York was the Deferred Acceptance algorithm which, when it was published in 1962 by David Gale and Lloyd Shapley, offered a solution to the problem of the stable matchings formally defined next. This algorithm works as follows: each student produces a list with his or her preferences for some schools, ranked strictly. Each school will have a list of priorities that is also strict for all the applicants wishing for a seat at that school, and each school will also have a fixed capacity—a fixed number of available seats. At the first iteration of the algorithm each student proposes his or her first school of preference, and each school will accept tentatively each proposal as long as it does not exceeds the school's capacity. If a school has more demand than available seats, it will accept all the applicants it can according to its capacity and its list of priorities for the students, assigning from highest priority to lowest priority. The rest of the students are rejected. In the second round, all students who were not accepted in the first round propose to their second highest preference, if they have any. Then, each school considers the new applications and the students tentatively assigned on the first round, and accepts tentatively again according to its priority list and capacity, even if that means rejecting a previously assigned student. This continues until none of the students' proposals are rejected. This can occur either due to the fact that all of the students are assigned to a school on their list of preferences, or that there are students who were not assigned (because of capacity constraints) but exhausted their applications to schools.

The mechanism implemented for school assignment is in fact the Deferred Acceptance algorithm, however, it incorporates significant modifications in order to satisfy the norms imposed by the educational reform. Before discussing these modifications and particularities that change the original version of the algorithm, its main properties will be presented, some of which will have a transcendent relevance in justifying the modeling of the decision making for the families on chapter 2.

The original assignment problem arises from a set of schools S and a set of students I. Each school $s \in S$ has the capacity to admit q_s students, and each student $i \in I$ has a strictly ordered list of preferences for a certain subset of schools. To avoid arbitrary discrimination by the schools, they are assigned a strict list of priorities on each one of the applicants which is randomly generated. It is worth noting that in the current system the list is not generated completely at random, but with priority criteria predefined by the State (which will be detailed further on). An assignment refers to pairs of students and schools (i, s), where the student i is and can only be paired with one school, in this case s, and each school can contemplate up to q_s pairings with students. This result will be referred to as a matching or assignment. Formally, a **matching** is a function $\mu : I \longrightarrow S \ \cup \ \{\emptyset\}$ such that $\forall s \in S, |\mu^{-1}(s)| \leq q_s$. Each student has a strict preference relation, $\succ_i$, between schools and not being assigned. If $s \succ_i s'$, then student i strictly prefers school s over school s', which his or her choices on the preference list. Also, if $s \succ_i \{\emptyset\}$, then student i strictly prefers school s over not being assigned. A similar notation will be used to represent the strict preference that schools have for students, $i \succ_s i'$, meaning that school s strictly prefers student i over student i'.

The list of priorities that define the ordered preferences of every school for the applicants is designed by what is known as **tiebreak rules**. A tiebreaker is a bijection $r : I \rightarrow \mathbb{N}$ that breaks ties for schools:

$$i \succ_s i' \iff r(i) < r(i') \tag{1.1}$$

The first important property to be studied is the notion of a stable matching that was mentioned before. We define the worst student of a school, $w(s) \in I, s \in S$, such that $\mu(w(s)) = s$ and $i' \succ_s w(s) \ \forall i', \mu(i') = s$, i.e. the lowest preferred matched student to a school. We also define a **blocking pair** as a tuple $i \in I$, $s \in S$ such that:

$$s \succ_i \mu(i) \tag{1.2}$$

$$i \succ_s w(s) \tag{1.3}$$

And we define that a matching has **individual rationality** if $\forall i \in I, \ \mu(i) \succ_i \{\emptyset\}$. Then, a matching is said to be **stable** if it is individually rational and there are no blocking pairs.

A stable matching is a desirable property for a problem like the one undertaken in school assignment since, without stability, in theory, the establishments could arrange with some families to switch the seat of a student who was assigned and who is less desirable for the school for another student who was not assigned a seat but is more desirable to that school and at the same time prefers to be assigned to that school instead of the one assigned to him by the system, therefore endangering the whole purpose of the assignment.

Following this thought, an algorithm can produce a matching which is stable, while another algorithm could generate another matching which is also stable but its not the same one, making it necessary to define what is a stable and optimal matching. Thus, a stable matching will be called **optimal** if each applicant is equally or better off in this assignment than in any other stable matching. A student is better off in his assigned match than to other school s if $\mu(i) \succ_i s$.

The Deferred Acceptance algorithm produces a stable and optimal matching (Gale and Shapley, 1962). This is one of the main advantages that makes it a favorable algorithm for solving the school assignment problem. Now, it is presented formally in algorithm 1.1.

Algorithm 1.1: Deferred Acceptance

```
Start with $\mu(i) = \{\emptyset\}$ $\forall i \in I$ and $|\mu^{-1}(s)| = 0$ $\forall s \in S$
While any $i$ such that $\mu(i) = \{\emptyset\}$ and $i$ still has proposals left:
  $s^*$ = first school on $i$'s list whom $i$ has not yet proposed
  if $|\mu^{-1}(s^*)| < q_{s^*}$:
      $\mu(i) = s^*$
  else:
      if $r(i) < r(w(s^*))$:
          $\mu(w(s^*)) = \{\emptyset\}$
          $\mu(i) = s^*$
      else:
          continue
```

Another property that is desirable is a matching algorithm that is strategy-proof. A matching mechanism is **strategy-proof** if none of the agents can benefit from misrepresenting their true preferences in their preference lists in an unilateral way. This means that the best strategy for the students (or the families) when they participate in the system is to reveal their true preferences for schools to the system. This is of utmost importance from the system's point of view, since any agent wanting to offer the service of helping to generate a strategy to help the applicants get into the best possible school for them would be fraudulent, given that the only useful strategy is revealing the true preferences. On the other hand, it justifies that the modeling of the decision making by the families or any other type of study that makes use of the data of the applications, because is using information that, in theory, contains the true preferences of the families. An example of a mechanism that is not strategy-proof is the one that is currently used in Santiago of Chile. A student that wants to get into the Instituto Nacional and the José Victorino Lastarria school, two of the capital's emblematic establishments, is under the obligation to strategize given that after rendering the admission tests for each school, the student must decide weather to enroll in the latter school before knowing the first one's answer. The Deferred Acceptance algorithm is strategy-proof (L. Dubins y D. Freeman, 1981; Roth, 1982).

The trade-off that results between using this matching algorithm and a different one, such as the Top Trading Cycles algorithm (Shapley and Scarf, 1974), is with respect to the efficiency of the assignment. A matching is said to be **Pareto efficient** if there is no other matching in which each student is in a worse or equal match, and at least one student in a strictly better match. The Top Trading Cycles algorithm is Pareto efficient and strategy-proof. Therefore, the choice of mechanism resides in deciding if stability is preferred over efficiency, or vice-versa.

1.3 Implementation of the Matching Mechanism

In June of 2015, the new School Inclusion Law was published, with its main objective being eliminating arbitrary selection from the schools that receive support from the State towards the students who wish to apply to them throughout the country. Therefore, the schools must accept all applicants as long as they have the seats to accommodate them, and the priority criteria for selection are determined by State-imposed rules, avoiding, for example, selection by a family socioeconomic status.

The applications must be submitted through this system for all those who wish to be admitted for the first time to a municipal educational establishment or one that receives State subsidy, students who wish to switch establishments (to one with the formerly mentioned characteristics), or those who wish to re-enter the school system. Through the website `www.sistemadeadmisionescolar.cl`, the Ministry of Education provides the families with access to apply to all the schools they wish, also containing information on the educational programs of all the establishments that participate in the system. It also provides recommendations on how to apply adequately, for example, they recommend applying to the greatest number of schools that the families like, or that the application must reflect the true order of preferences. On the other hand, they give the recommendation that if applying to schools with high demands, they should also apply to schools with lower demands (an image of this recommendation is shown in the annex 1). This could be misleading if misunderstood, because it could be interpreted as an indication to alter true preferences, and will be referred again in the the analysis of preference estimation in subsequent sections.

The alterations to the Deferred Acceptance algorithm to produce the match between applicants and schools, which also makes the case of the Chilean problem unique with respect to the systems in the rest of the world, consist in defining priorities criteria with which the establishments will admit the applicants. The first one is related to the tiebreaking rules. The ones to be considered for the assignment problem are whether to use single or multiple tiebreaking. A **single tiebreak** rule implies using the same tiebreaker for all the schools, while a **multiple tiebreak** uses different tiebreakers for each school. In practice, these strict tiebreak lists are random numbers assigned to each student indicating his or her priority in a certain school. For the present case, multiple tiebreak rules were implemented in order to avoid having a student with a bad priority in all establishments because of the use of a single tiebreak rule, although single tiebreaking rules have shown to have superior results in simulations that measured welfare properties (Abdulkadiroğlu et al., 2009). It is worth mentioning that the strategy-proof property is maintained using any set of tiebreak rules (Abdulkadiroğlu et al., 2009), and it has been shown (Kesten ,2004) that when the preference lists are strict—as is in the present case—there are no mechanisms that can be strategy-proof, Pareto efficient and Pareto dominate the Deferred Acceptance algorithm. So, this way the Deferred Acceptance algorithm keeps demonstrating to be the better choice for this particular context.

The other modifications applied to the algorithm consist in the implementation of priority quotas. In order to detail them, first it is necessary to define the kinds of applicants that the system can identify and treat differently:

- **Assured enrollment**: An applicant belongs to this class when they are applying to the school they are currently enrolled in
- **Special Admission Programs (PIE)**: Students that present special educational needs
- **Priority student**: Students who are within the most vulnerable third according to the Social Registry of Homes, or other requisites specified by the Ministry of Education to validate their economical situation
- **Academic excellence (AE)**: Students who have had historically high performance in their schools
- **Siblings**: Students who have an (older) sibling in the school to which he or she is applying
- **Staff member's children**: The child of a staff member of the establishment
- **Regular**: Students in this class do not possess any of the characteristics formerly described, and their priority is defined by the tiebreaker only

There are also special openings that a school can allocate to select students in accordance with the academic performance of a student. These schools are referred as **high demand schools**, and must have at least 30 years of operation, have twice as many applicants as seats in the admission process, and they must be in the top 20% of the results in the national measurements. These schools are classified into two categories:

- **High demand in regime (HDR)**: They can assign 30% to students belonging to the top 20% of academic performance in their schools
- **Transitory high demand (HDT)**: They can make a priority list of students to fill up to 85% of the vacancies (e.g standardized test) in accordance with what it is established by the school admission process

Other seats that schools can declare are for PIE and regular students. For the PIE seats, first there is an order for PIE students, then siblings, children of staff members, former students, and finally the rest of the applicants. For high demand schools, the high demand assignment proceeds first, and then the same order before mentioned for the rest of the students. The school must also have a quota for priority students, which is calculated from the regular seats and considers the same order as before, but correcting by the priority students who should have a higher preference. The regular seats also follow the former order.

It is important to mention that the applicants that have identical characteristic (as far as the system is allowed to observe) are differentiated in their priority for a school by the tiebreaker. The order in which the different types of students will be filling up the seats is:

- A student that belongs to the classes assured enrollment, PIE and priority tries to fill a seat in the following order:
  PIE, priority, regular, HDT and HDR
- Assured enrollment, PIE and priority:
  PIE, regular, priority, HDT and HDR
- Assured enrollment, non PIE and priority:
  Priority, regular, HDT, HDR and PIE

- Assured enrollment, non PIE and non priority:
  Regular, priority, HDT, HDR and PIE
- No assured enrollment, sibling and non priority:
  PIE, HDT, HDR, PIE, regular and priority
- No assured enrollment, regular:
  PIE, HDT, HDR, PIE, priority and regular

The Deferred Acceptance algorithm with specific-type quotas is also strategy-proof (Abdulkadiroğlu et al., 2003).

Following the rules afore mentioned, seats are tentatively assigned through a **choice function** (the function that generates the mapping from students and schools to matchings) until either all the applicants are assigned to a school, or no applicants have any proposals left (they were rejected in all of the establishments to which they applied), thus framing this choice function (which defines the priority criteria) inside the classic Deferred Acceptance algorithm. In a following round, the students can once again participate in the system, and in this last instance the system assures that all the students are assigned by proposing an assignment based on the distance to the nearest school (that satisfies minimum quality standards declared by the Ministry of Education) as a last resource, in order to guarantee that there is a school for every student in the country.

1.4 Results from the 2017 Process

In 2017, 76,821 students submitted applications through the system. In this instance there was an offer of 6,635 courses in 2,100 schools for the 5 regions before mentioned. The period of applications began on September 25th and closed on October 13th. The results of the assignment are shown on table1.1.

Assingment	Result	%
1° preference	43,179	56.2%
2° preference	11,146	14.5%
3° preference	5,311	6.9%
4° preference	2,318	3.0%
5° preference	1,014	1.3%
6° preference	457	0.6%
7° preference	172	0.2%
8° preference	105	0.1%
9° preference	44	0.1%
$\geq$ 10° preference	34	0.0%
Assured enrollment	6,365	8.3%
Distance	0	0.0%
Unassigned	6,676	8.7%

Table 1.1: General results for the school admissions in 2017

The results show that a low percentage of students was not assigned beyond their third preference, and a high percentage of students was not assigned. This could be explained by the low amount of declared preferences, being 4.05 the average of preferences, considering a universe of 50,905 students as a result of excluding rural zones (there is more common to have a lower number of applications and is actually allowed to only submit one preference, where the rest is obligated to submit at least two). This is remarked as low because it would be reasonable to assume that, in general, families value positively more than 4 schools, and as a result they should apply to all positively valued schools. Table 1.2 shows the percentage of students assigned as a result from different number of declared schools in their lists of preferences.

Number of declared preferences	% assigned
1	72.0%
2	76.4%
3	80.9%
4	82.5%
5	85.7%
6	86.3%
7	88.1%
8	90.2%
9	91.6%
≥ 10	95.1%

Table 1.2: Percentage of students assigned in accordance with the number of schools declared in their preference lists

As would be expected, by declaring more schools there is a higher chance of being assigned to one of them. However, these results allow for the empirical quantification of, for example, the difference in probability of being assigned by declaring 2 and 6 schools, which was nearly 10% higher.

Chapter 2

Models for Estimating Preferences in the School Admissions Process

This chapter describes the methodology and models to be used for modeling the decision making of Chilean families when applying to schools through the admission system. The following sections will show in detail the discrete choice models that were implemented in order to understand what is important (and what is not) when choosing schools—and by how much—, using the data from the school's admission process and data of the establishments that will be presented in chapter 3. This, in order to quantify not only the variables that affect the applications, modeling also the uncertainty through probabilities in a completely Bayesian treatment, but to also being able to study counterfactuals that can help support public policies, allowing to study the effects of policy changes in the school's admission system.

2.1 Why Use a Bayesian Approach?

The study and application of statistical inference has historically been dominated by the frequentist approach. As an example of this, the usual formula for estimating a model of latent utility—such as the logit or probit presented next— is carried out through maximum likelihood estimation. However, conducting inference through a Bayesian approach has become much more relevant in recent years due to numerous successful applications of these methods, advances in computing power (even though, as will be shown later, carrying out Markov Chain Monte Carlo simulations is still a very demanding computational exercise), and also some of the main criticisms against the Bayesian approach have lost strength, for example, the importance of choosing the priors, which carry subjectivity in the judgment of the researcher about the problem in question, as this is balanced out by extensive simulation and large availability of data (Gelman et. al., 2009). For the present work, the main advantages of using a Bayesian approach are the treatment of uncertainty through probability distributions that will be directly simulated via Gibbs sampling, and not being necessary to explicitly write the likelihood function, which for the present case could be difficult to specify.

The general problem addressed in the following sections consists in modeling the latent utility of a student i for a school s, w_{is}, through a set of covariates $\boldsymbol{X}_{is}$ that includes attributes of the students, the schools, and variables that depend on both agents (e.g. the distance between an applicant and the school to which he or she is applying). The observable variable that is aimed to be represented by this formulation is the ranking, y_{is}, an integer that indicates the position on which the student i values the school s relative to the other schools on his or her list of preferences. Even though it is allowed to rank as many schools as the student desires, it will always occur that there are schools that the student does not add to his list of preference—as he or she will naturally not rank all the schools in the country. Therefore, it will be assumed that the schools are not ranked by a student are strictly less preferred that the ones that were ranked, supporting this hypothesis also by the strategy-proofness of the matching system. The posterior distribution aimed to be recovered will therefore be that of the conditional latent utility $\boldsymbol{w}$ with respect to the parameters of the models and the observable variables.

2.2 Multivariate Probit

A **multivariate probit** model aims to represent a choice in which an agent i (in this case the students) can choose between p alternatives, where each alternative represents a different school. The agents chooses k alternatives, with $k \leq p$, and these choices are represented by binary variables y_{is}, which take the value of 1 if the agent ranks school s in his or her list of preferences, and 0 if not. It is assumed that the agents have a non observable appreciation or valuation for each one of the alternatives, w_{is}, which is specified as:

$$w_{is} = X_{is}\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \tag{2.1}$$

$$y_{is} = \begin{cases} 1, & \text{if } w_{is} \geq 0 \\ 0, & \text{if } w_{is} < 0 \end{cases} \tag{2.2}$$

The use of a Gaussian error term with 0 mean and a variance-covariance matrix $\boldsymbol{\Sigma}$ indicates that this is a probit model, and it is said to be multivariate because it allows to model a problem in which the agent can choose multiple options from the total of alternatives, as is the case for the present problem. This specification leaves out the rankings that the students grant the schools, but could be a good enough approximation if, either the rankings are uninformative or too noisy, or the information contained in the indication of whether they apply or not to the schools only allows for good estimations of the preference lists.

The benefit from using a probit model instead of a logit, which assumes the error term distributes extreme value type I, is that it does not allow for the estimation of systematic random variations, which are captured in the $\boldsymbol{\Sigma}$ matrix when using a probit model. The trade-off between both approaches is produced by the fact that the logit model is much simpler to estimate due to its closed form, and to estimate the probit simulations must be conducted. Nonetheless, only probit models will be considered.

In this model, $\boldsymbol{X}$ is a matrix of covariates of dimension $(n \cdot p, (n_d + 1) \cdot (p - 1) + n_a)$, in which n is the number of agents, p is the number of alternatives, n_d is the amount of covariates that are independent from the alternatives, and n_a is the number of specific covariates of each alternative (such as the school gender, for example). The term $1 \cdot (p-1)$ indicates the inclusion of fixed effects for each alternative with the exception of one, to avoid multicollinearity. The parameter vector β has dimensions $((n_d+1) \cdot (p-1) + n_a, 1)$, and the error term has dimension $(n \cdot p, 1)$.

In order to implement a Gibbs sampler that conducts Markov Monte Carlo Chain simulations, the priors for the parameters of the model, β and $\mathbf{\Sigma}$, must be specified. This will be done through the use of non-informative priors, reflecting the lack of knowledge on the uncertainty and around which values the coefficients of the covariates lie, and hoping this way the actual data guides the values that the distribution parameters should take. The specification of these parameters is the following:

$$\beta \sim \mathcal{N}(0, \boldsymbol{A}^{-1}) \tag{2.3}$$

$$\boldsymbol{A} = 0.01 \cdot \boldsymbol{I} \tag{2.4}$$

$$\mathbf{\Sigma} \sim \mathrm{IW}(\nu, V_0) \tag{2.5}$$

$$\nu = p + 3 \tag{2.6}$$

$$V_0 = \nu \cdot \boldsymbol{I} \tag{2.7}$$

Matrix $\boldsymbol{A}$ shows that the correlations between the parameters contained in β will not be made explicit beforehand by using the identity matrix $\boldsymbol{I}$ and, as was mentioned, non-informative normal distributions are specified for each parameter by setting a variance of 100. The variance-covariance matrix is specified as the inverse-Wishart distribution. This distribution is the prior conjugate of a multivariate normal distribution, which justifies its choosing. Its parameters represent the degrees of freedom, ν, and a finite density if $\nu \geq k+1$, with k the dimension of $\mathbf{\Sigma}$. The distribution is non informative as $\nu \longrightarrow 0$, therefore the value of ν is set to give a finite density and also to make it non-informative. V_0 is a positive definite matrix of scale.

With this specification, the Gibbs sampler for the multivariate probit model results from iterating as shown in algorithm 2.1.

Algorithm 2.1: Gibbs Sampler for the Multivariate Probit

```
Start with initial values for $w_0$, $\beta_0$ and $\mathbf{\Sigma}_0$
Sample $w_1|\beta_0, \mathbf{\Sigma}_0, y$:
  for $i$ in $\{1, ..., n\}$:
    for $s$ in $\{1, ..., p\}$:
      if $y_{is} = 1$:
        Sample $w_{is}|w_{i,-s}, \beta, \mathbf{\Sigma}$ from a positive truncated normal distribution
      if $y_{is} = 0$:
        Sample $w_{is}|w_{i,-s}, \beta, \mathbf{\Sigma}$ from a negative truncated normal distribution
Sample $\beta_1|w_1, \mathbf{\Sigma}_0$
Sample $\mathbf{\Sigma}_1|w_1, \beta_1$
Repeat until convergence is reached for several iterations
```

The implemented version for sampling from the conditional distributions of w, β y $\mathbf{\Sigma}$ follows the one in Edwards and Allenby (2003). The initial values, w_0, β_0 and $\mathbf{\Sigma}_0$, are vectors of 0s and the identity matrix is set for the variance-covariance matrix. The sampler is clearly a Markov Chain since the state of the random variables is updated only with information from the current state of the rest of variables, complying with being a stochastic process that follows the Markovian property. One of the important parameters that must be tuned is the number of iterations for the sampler. There must be enough iterations for the Markov Chain to reach the stationary distribution and sample from there (when neither the first nor second moments of the model's simulated parameters change as the iterations continue) in order to infer from the obtained distributions of the parameters.

The interpretation of the β coefficients that result from estimating a multivariate probit is that if the mean of a coefficient is positive and statistically significant, this reflects a positive effect on the underlying utility shown on the equation 2.1—more precisely, one can refer to the probability of it having a positive effect—, which translates into a bigger chance for the alternatives that contemplate this positive effect are actually chosen by the students. In the opposite scenario, if there is high certainty that a coefficient has a negative effect, this would indicate a tendency for not choosing the alternatives that consider this variable. The interpretation of the magnitude of the effect is direct: a bigger absolute effect indicates a higher probability for the school to being included in the preference lists, thus reflecting the importance of making the decision concerning which schools to apply to.

The main disadvantage of using the multivariate probit to model the data produced in the school admission problem is that it ignores the ranking in which certain schools are submitted by the students in their applications, and only taking into account if they include them or not it in their preference lists. This ignores information which would be reflected in the underlying utility, w_{is}, which should grant a higher value to a certain school s if it is positioned higher than other schools in the preference lists of the students, translating into information that the parameters β of the model could capture. It is for this reason that in the following section a version capable of modeling the true data generating distribution is presented, and the results of the Gibbs sampler shown in this section will be available to contrast differences in the fits of the models.

2.3 Multivariate Ordered Probit

The second model that was implemented is a **multivariate ordered probit**. This one incorporates the specification of a multivariate probit in which an agent chooses k among p alternatives, but adds that these k chosen alternatives have a strict order of preference. This models allows to represent how the preference data was generated for the school admission system, given that for each student there are p alternatives of schools to which he or she can apply, from which a decision is made to apply to a k number of them, and they are ranked in a way in which the first ranked school is the one that the student desires to enroll the most, the following is the second most desired, and so on. This produces a latent valuation with a strict order for the chosen schools. For the schools that are not chosen, the specification of a negative and orderless underlying valuation is maintained.

The specification of the latent utility is equal to the one shown in 2.1, but the vector of responses is represented in a different way:

$$y_{is} = \begin{cases} 1, & \text{if } w_{is} > w_{i,-s} \text{ and } w_{is} > 0 \\ 2, & \text{if } \exists\ s' \text{ such that } w_{is'} > w_{is} \text{ and } w_{is} > w_{i,-\{s,s'\}} \text{ and } w_{is} > 0 \\ ... & \\ 0, & \text{if } w_{is} < 0 \end{cases} \tag{2.8}$$

The Gibbs sampler for the multivariate ordered probit model results from iterating as shown in algorithm 2.2.

Algorithm 2.2: Gibbs Sampler for the Multivariate Ordered Probit

```
Start with initial values for $w_0$, $\beta_0$ y $\mathbf{\Sigma}_0$
Sample $w_1|\beta_0, \mathbf{\Sigma}_0, y$:
  for $i$ in $\{1, ..., n\}$:
    for $s$ in $\{1, ..., p\}$:
      if $y_{is} > 0$:
        Sample $w_{is}|w_{i,-s}, \beta, \mathbf{\Sigma}$ from an upper and lower truncated normal distribution
      if $y_{is} = 0$:
        Sample $w_{is}|w_{i,-s}, \beta, \mathbf{\Sigma}$ from a negative truncated normal distribution
Sample $\beta_1|w_1, \mathbf{\Sigma}_0$
Sample $\mathbf{\Sigma}_1|w_1, \beta_1$
Repeat until convergence is reached for several iterations
```

It should be noted that this model in a generalization of the multivariate probit, given that apart from considering if a student applies or not, it considers the additional information indicated by the ranking of preferences. The algorithm to conduct Markov Monte Carlo Chain simulations is similar to the multivariate probit model, with its main difference in how the conditional underlying utilities are sampled.

This process is still a Markov Chain since the way each utility is sampled is always conditional to the current state of the stochastic process (see algorithm 2.2), and is therefore also a Gibbs sampler.

The implementation of the sampler in order to obtain the conditional underlying utilities for the schools for students in an orderly fashion is shown in annex 2. The condmom function of this implementation obtains the first and second conditional moments of $w_{is}|w_{i,-s}, \mu_i, \mathbf{\Sigma}$, which is then used to sample from truncated normals, with trunNorm and rtrunSc obtaining the necessary samples from a truncated normal distribution on one and two sides, respectively. The vector of means, μ, is obtained as $\boldsymbol{X}\beta$, and is sliced so that μ_i corresponds to the means of the applicant in question. The implementation of the function that obtains the complete utility vector for each student is shown in annex 3.

The parameters for the priors are the ones previously used for the multivariate probit (see equations 2.3 - 2.7), reflecting again little knowledge about the distributions that generated the applications data for the admissions problem. A special consideration that must be had is with respect to the initial value of w_0, since initializing it in 0 for all alternatives—as was done for the multivariate probit—would generate an error in the first iteration by not being able to sample from an upper and lower truncated normal distribution with respect to the values of w_0 (i.e. to sample between 0 and 0). Therefore, the values are initialized equidistantly between 0 and 10 for the ranked options, and sampled in a uniform distribution between -10 and 0 for the non ranked options.

Implementations in C++ of the samplers were necessary because of the cumbersome computation needed for Markov Chain Monte Carlo simulations (see section 4.4 for the solving times) and, therefore, fast computation by the means of this low level programming language needed to be achieved. Also, there were not libraries available for the Bayesian multivariate ordered probit, and little literature was available for this model since often the multivariate probit and ordered probit are studied separately, so it is hoped that this would be a contribution to the field (the development and implementation of the Bayesian sampler) and to the literature with respect to an empirical use of the multivariate ordered probit.

Chapter 3

Data and Preference Model

The data presented in this chapter comes from the school admission process of 2017, the SIMCE databases (which contains the historical results of the schools in standardized tests) and the Personal Development Indicators databases[1].

For the work presented here and in the following chapters, only the biggest region (in terms of amount of applications and population) is considered. A map of the Coquimbo region and its administrative divisions is shown in figure 3.1. It has 3 provinces: Elqui (north), Limarí (center) and Choapa (south). Each province is divided by communes, with a total of 15 communes. It has a population of over 742,000 in accordance with the 2017 census.

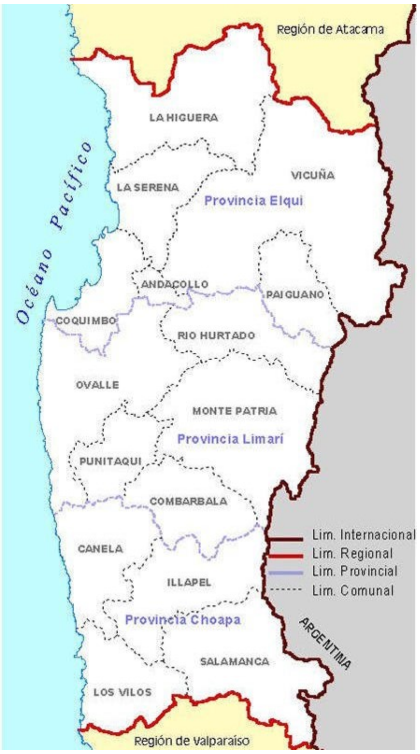


Figure 3.1: Geographic division of the Coquimbo region

The decision to model preferences only for this region comes from considering several factors: first, it is assumed that no families living hundreds or thousands of kilometers away from a particular school, like in the Magallanes region, are even considering in their choice sets schools in the Coquimbo region, and to some extent, not any other region because of the large distances between them. This could be validated from the actual data of the student's homes georeference but, as will be discussed in subsequent sections, that data is not clean. The Coquimbo region alone has an area of 40,580 km^2. In contrast, the city of New York, where a preference model was also estimated (Abdulkadiroğlu et.

[1] This research used as a source of information the databases from the Agency of Education Quality. The author thanks the Agency of Education Quality for the access to the information. All the results from the study are responsibility of the author and in no way compromises the Institution

al., 2017), has only 789 km^2 of area. The smallest province, Choapa, has 12.8 times more area than New York. The biggest province, Elqui, has 21.4 times more area than the American city. The area would be comparable to something in the scale of communes (e.g. districts), between the sizes of Andacollo (310 km^2) and Paiguano (1,495 km^2). See figure 3.1 to get a sense of the scale of the problem. This prompts an additional difficulty in estimating the preference models, since choice sets will have to be designed specifically to account for this problem. Second, taking into account a large number of schools for the choice sets of the families makes the problem of estimating preferences intractable (this will also be discussed later in more depth in the results presented in chapter 4). Therefore, the data, model, and results presented concern only the Coquimbo region, being the region with more data available. Nonetheless, the methodology shown in this and the following chapters can be replicated for the other regions that participated in the system, and for future iterations of the admission process.

3.1 Schools

436 schools in the Coquimbo region participated in the 2017 school admission process. These schools sum up to 1,368 courses where students could be assigned. Table 3.1 shows the distribution of schools by province and commune. It can be seen that the most important communes in terms of school offer are La Serena, Coquimbo and Ovalle. In terms of the 3 provinces, the biggest in terms of school offers is Elqui, with 46.6% of the total of schools, followed by Limarí with 35.3%, and last Choapa with 18.1%.

Province	# schools	Commune	# schools
Choapa	79	Canela	18
		Illapel	24
		Los Vilos	15
		Salamanca	22
Elqui	203	Andacollo	7
		Coquimbo	76
		La Higuera	7
		La Serena	85
		Paiguano	8
		Vicuña	20
Limarí	154	Combarbalá	19
		Monte Patria	37
		Ovalle	66
		Punitaqui	15
		Río Hurtado	17

Table 3.1: Number of schools distribution by commune and province for the Coquimbo region

Figure 3.2 shows the schools as dots in a map of the region, distinguishing by different colors the province where a school belongs to: Elqui (red), Limarí (green) and Choapa (blue).

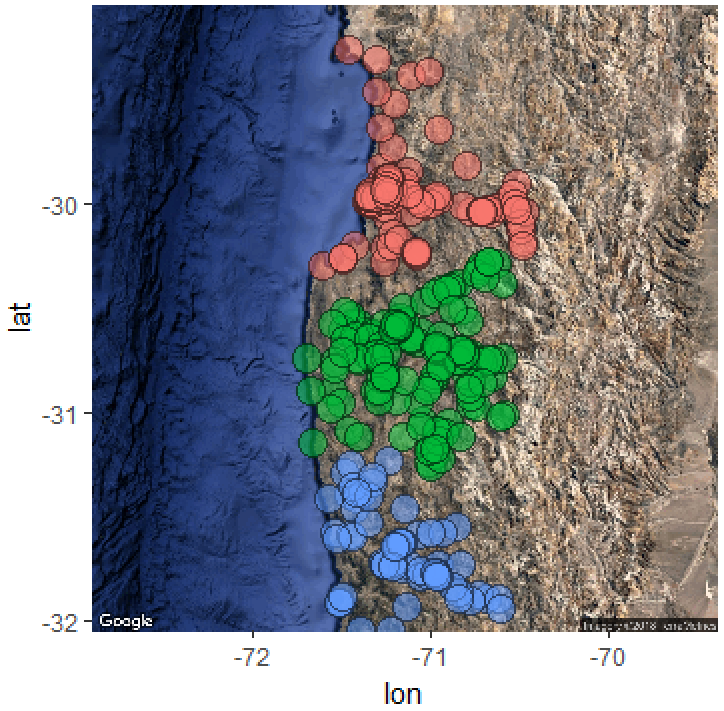


Figure 3.2: Map of schools in the Coquimbo region

It is important to remark that, as this was the first time that the region used the matching system to assign students, the law required that only some are allowed to participate, and then the following year all levels are considered. These levels corresponded to the ages 4 (pre-kinder), 5 (kinder), 6 (1st year of primary school), 12 (7th year of primary school), and 14 (1st year of secondary school). Figure 3.3 shows the distribution of courses with declared seats by level. Ages 4, 5 and 6 accumulate a 68.1% of the school supply, whereas ages 12 and 14 have the remaining 31.1%.

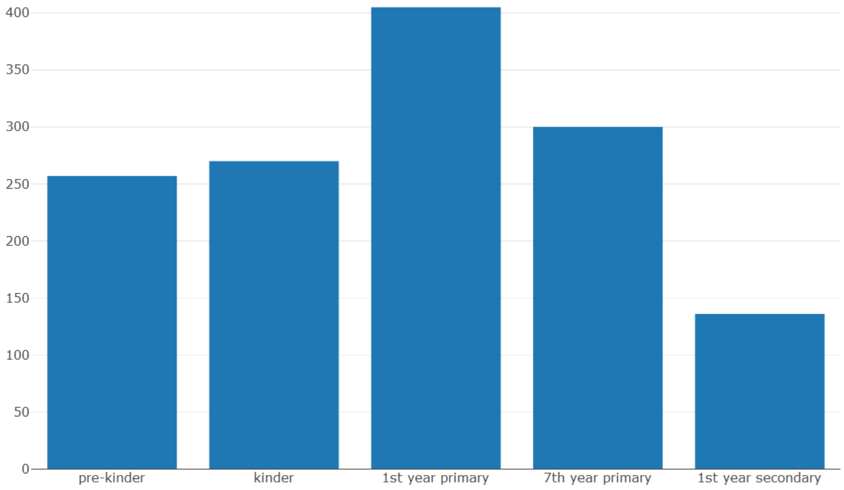


Figure 3.3: Histogram of courses with declared seats by level in the Coquimbo region

Before the matching can be computed, the schools have to declare the number of available seats that they are going to offer. These seats come as regular, PIE, priority, transitory high demand and regime high demand openings. For the 2017 process, no high demand seats were declared for this region. Figure 3.4 shows for each school (represented by dots) how many total seats it declared available and how many applicants did that school had. There is a positive correlation between the two variables (the Pearson correlation between them is .48), but one can see that there are many cases in which the demand vastly exceeds the supply. For example, the school with the highest demand has over 19 times the number of applicants over number of available seats. In 40.4% of the cases, the demand for seats exceeds the supply. In the plot the schools with over demand are shown in yellow dots.

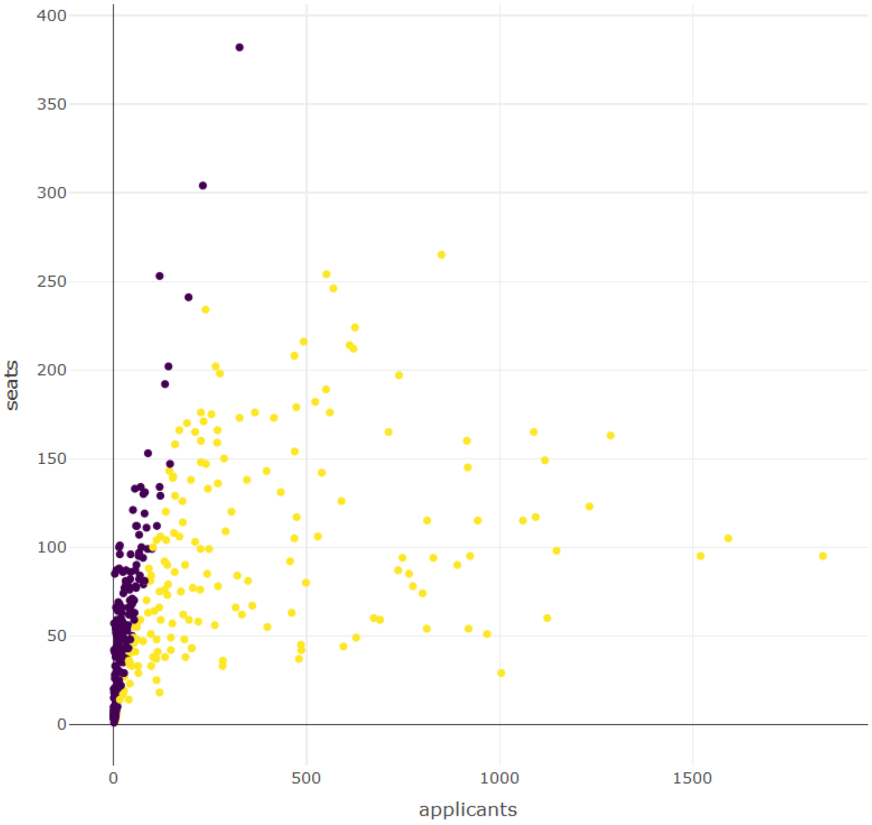


Figure 3.4: Applicants and declared seats by school

Schools can be classified by their socioeconomic status given by the income of the families of students that belong to each school. The groups that are used by the Ministry of Education are: low, medium low, medium, medium high and high. Figure 3.5 shows in a pie chart the distribution by these socioeconomic groups. It can be noticed that there is only an handful schools with a classification of medium high or high income (actually, only 1 classified as high). This is because in the system only public schools and schools that receive support from the State participate—not private schools. The lowest scored schools in terms of these categories, low and medium low, accumulated up to almost 70% of the school offer.

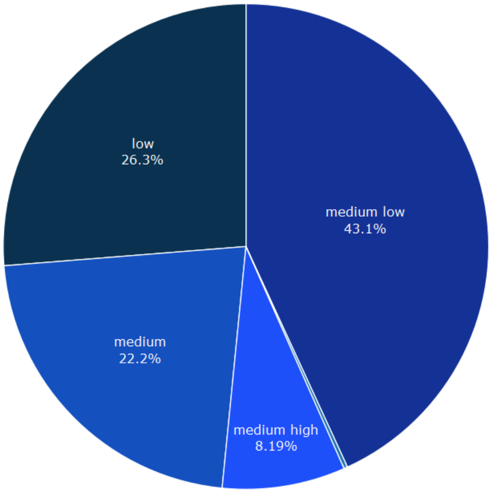


Figure 3.5: Distribution by socioeconomic group of schools

Schools can also be classified in urban or rural schools. Urban schools amount to 54.8% of the total schools. Figure 3.6 shows how each of this groups is composed by the socioeconomic classification presented before. It can be seen that rural schools contain 2.2 times as much low income schools, and urban schools contain 5.3 times more medium and medium high schools.

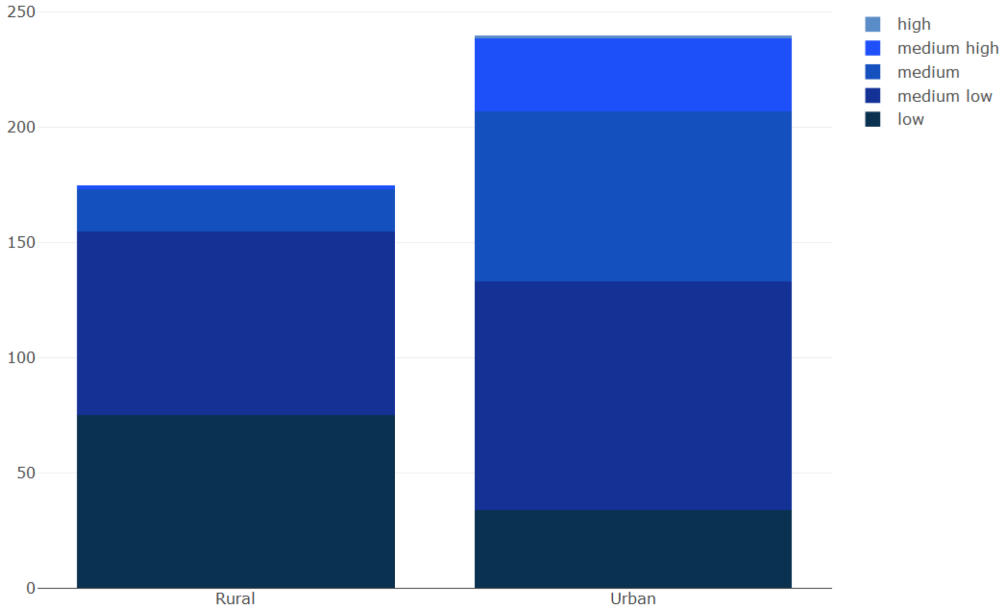


Figure 3.6: Rural and urban schools with their SEG distribution

Schools that receive support from the State and charge monthly payments to the families are called subsidized private schools. They amount to 15.3% of the total schools. A histogram of the distribution of the price variable is shown in figure 3.7, which ranges from \$0 (all the public schools) to \$99,946 CLP (about \$150 USD). This price is charged for each children that a family has in the school, so it can be a significant amount by noting that the minimum wage in Chile is $276,000 (about $420 USD). The mean price for the subsidized private schools is $50,497.

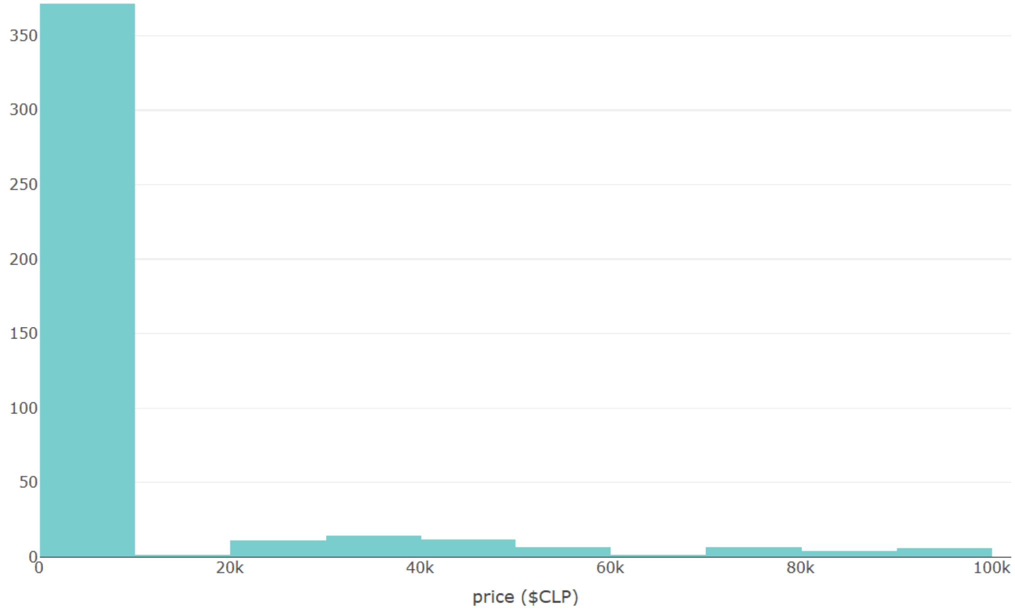


Figure 3.7: Histogram of prices by school

To provide a proxy for the academic quality of the schools, the SIMCE's results for the math and language tests for the years 2014 to 2016 for the levels 4th year of primary, 6th year of primary and 2nd year of secondary are used for each school. This tests are standardized and measured every year since 1968 in Chile and are used to improve the assignment of resources for the schools and to evaluate the achievement with respect to the active curriculum. To obtain an aggregate measure for each school in the Coquimbo region, the average over the 3 years are taken for all the available registered observations, since, for example, not all schools have primary and secondary levels, and there are also missing observations for some schools. The distribution of scores, which ranges from a minimum of 178 to a maximum of 334, is shown in figure 3.8. These distributions seem fairly similar. In fact, the mean for the language scores is 248.1 with a standard deviation of 19.5, and for the math scores is 240.1 and 23.7 respectively, reflecting that the math scores tend to be lower in average, but the distribution contains a heavier tail, as can be seen in the figure. Also, the skewness for each distribution is .17 and .59, reflecting again (although both have heavier tails to the right side of the distribution) that there are more schools that can get far from the mean scores in a positive way—specially in math.

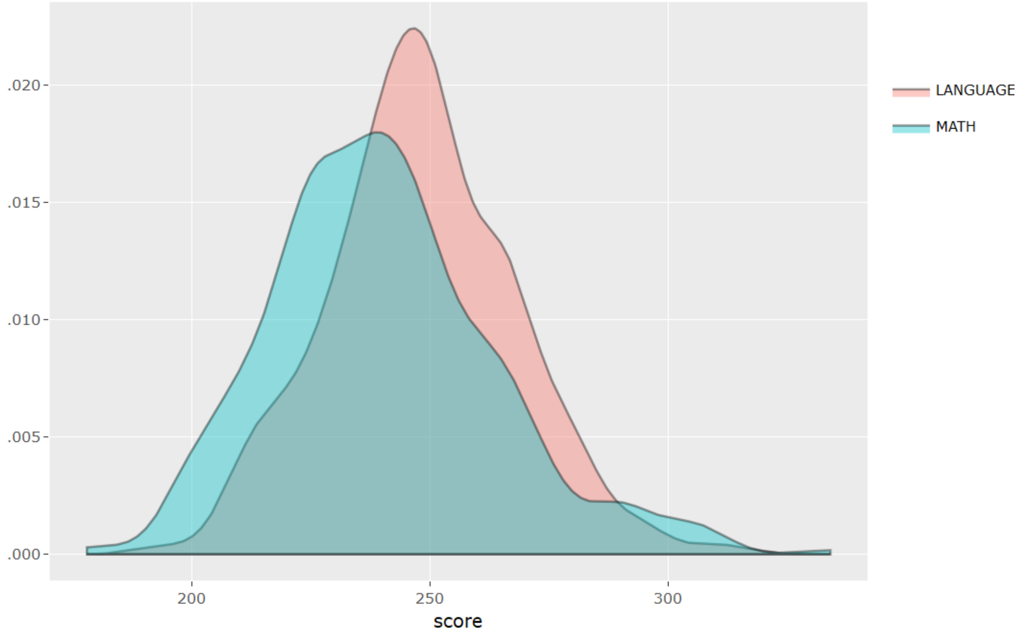


Figure 3.8: Density plot for the average math and language SIMCE results for schools

The correlation between the two variables is very high (.76 Pearson correlation) meaning that schools tend to have either good or bad results at both tests. The academic quality variable to be specified in section 3.4 will use the mean between language and math for each school.

There are also yearly measurements by the Agency of Education Quality about more qualitative attributes of the schools. These are called the Personal Development Indicators (PDIS) and record four variables assigning a score from 0 to 100 for each school. These variables are:

- Academic self-esteem and motivation of the school (**SELF-ESTEEM**)
- Climate of coexistence of the school (**CLIMATE**)
- Participation and citizenship formation of the school (**PARTICIPATION**)
- Healthy life habits of the school (**HABITS**)

As with the SIMCE's variables, means are computed for across all the available observations on levels 4th year of primary, 6th year of primary and 2nd year of secondary for the years 2014 to 2016. The variables are all positively and heavily correlated (see figure 3.9), with the healthy habits and climate of coexistence of a school being the highest correlated variables.

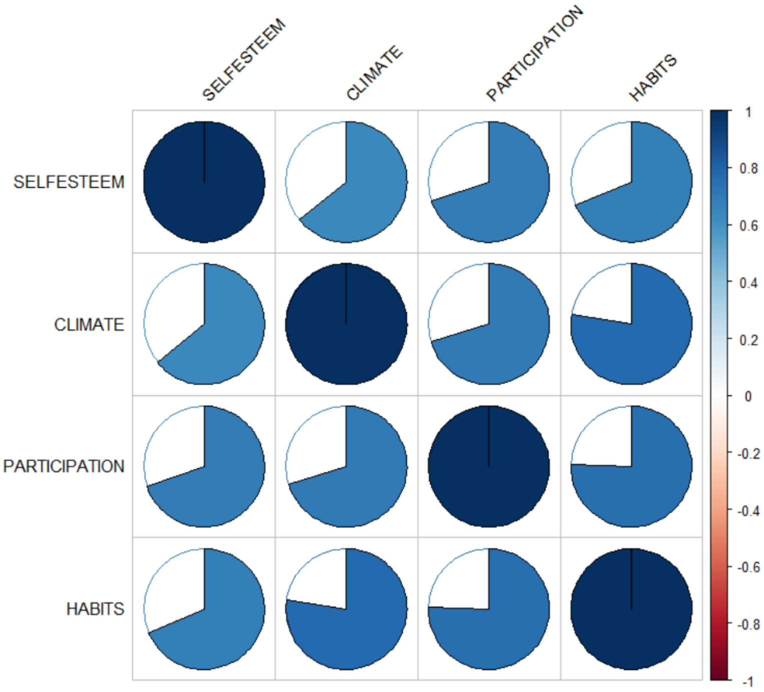


Figure 3.9: Correlation matrix of the PDIS for the schools in the Coquimbo region

The distribution of the PDIS scores is shown in figure 3.10. All the PDIS, with the exception of the participation score, are skewed positively, meaning that the distribution of schools is heavier towards positive scores. The thin left tail is a result of a school with participation score more than 6 standard deviations away from the mean, and that does not occur for the rest of the indicators.

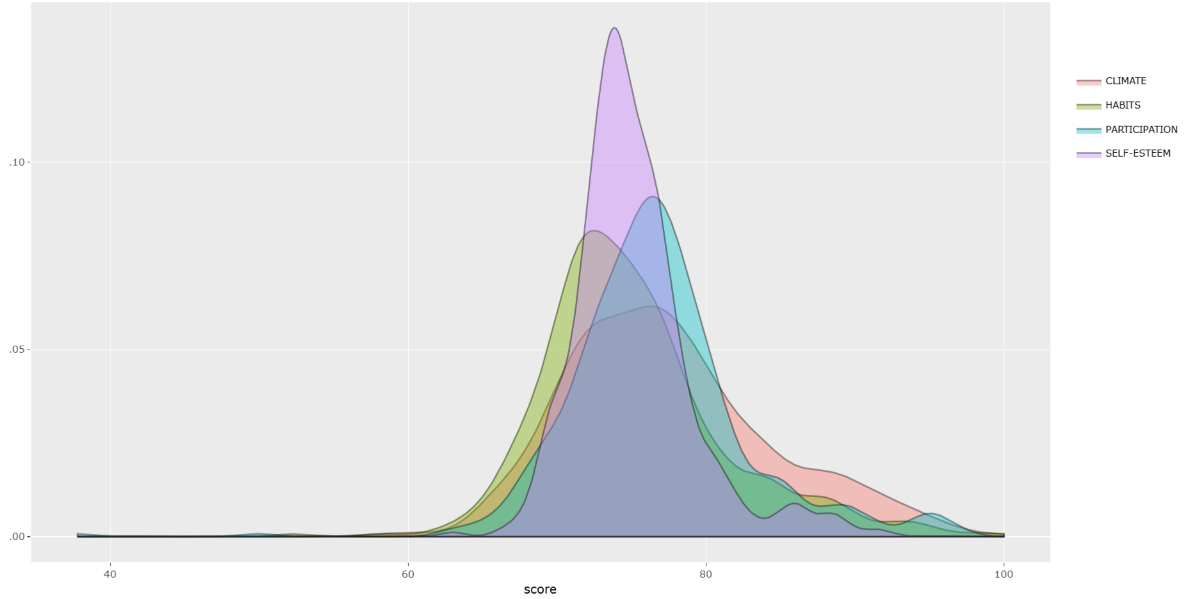


Figure 3.10: Density plot of the PDIS for the schools in the Coquimbo region

The inclusion of these variables in the later models comes from wanting to know if the families take into account, for example, the climate of coexistence within the school. It would be expected that, if they did, the schools with low scores would be ranked lower, because they would not want to send their children to a poor environment. This also prompts the question if these variables actually can measure what they intend—given that they are qualitative attributes—that will be addressed in the results of the preferences estimation.

3.2 Students

In the 2017 admissions process, 20,115 students sent applications for schools in the Coquimbo region, accumulating up to 26.2% of the total applicants that participated in the system in the 5 regions. As was described in chapter 1, these students are classified with respect to their families' socioeconomic situation, special educational needs and grades in order to compute their priority in the school's lists. Figure 3.11 shows how many students belong to each set, noting that 49.3% of the total of applicants have a priority condition (PRI), 2.6% have special educational needs (PIE) and 7.3% are of high excellence in terms of academic performance (AE). The biggest interaction comes from students that are both PRI and AE, those being 3.9% of the total applicants of the region.

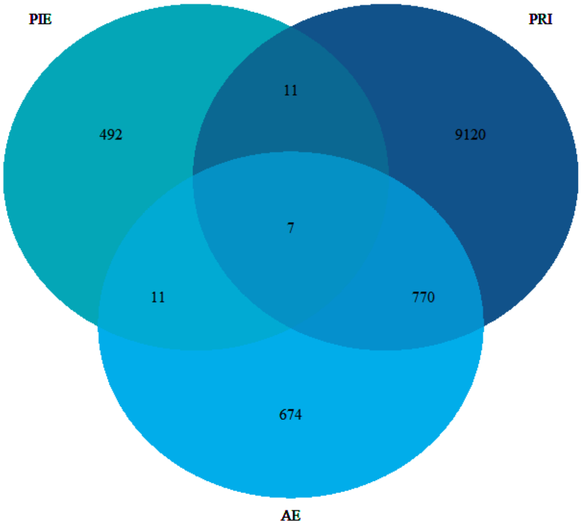


Figure 3.11: Venn diagram of the sets of PIE, PRI, and AE applicants

The number of applicants per level has the following distribution: 6,573 for pre-kinder, 3,038 for kinder, 3,829 for 1st year of primary, 1,848 for 7th year of primary and 4,827 for 1st year of secondary. This distribution is shown in figure 3.12 by also classifying into groups of "kids" (ages 4 to 6) and "teenagers" (ages 12 and 14). It can be noted that, by dividing by these ages, the kids group sum up to 66.8% of the total of applicants. This division is deemed to be important since one would want to verify if families value attributes of a school differently is their children are kids or teenagers. For example, one could think that teenagers can get farther away from home to go to school, and thus valuing less a rural school and more a school closer to urban zones of the city.

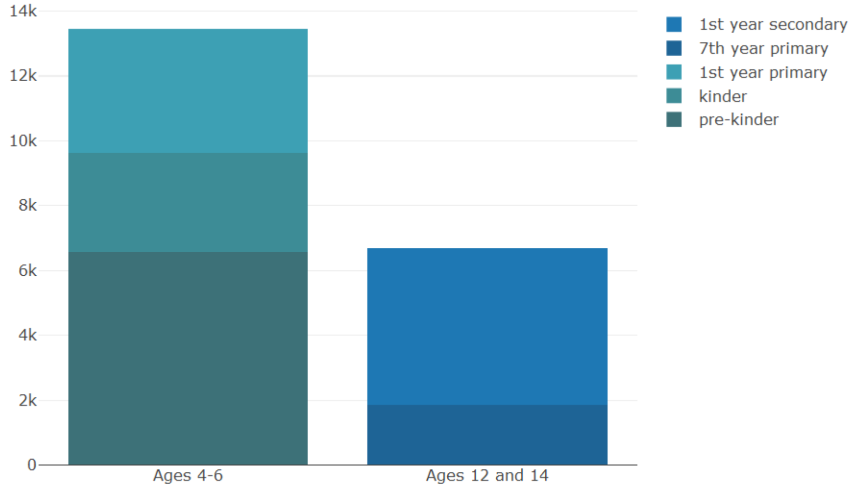


Figure 3.12: Distribution of applicants by the level to which they are applying, and also by the age that corresponds to those levels

It is only natural to think that the farther away a school is from an student's home, the least likely it will be that the student will end up submitting an application for that school. Figure 3.13 shows the distribution of distances according to the applications sent for schools in the Coquimbo region, and it can be seen that there is a considerable number of applications where the distance between the student's home and the school seems inconceivable. In fact, 6.2% of the applications have a distance of over 100 kilometers, and 4.6% of over 1,000 kilometers. This could have several explanations: some applications are sent from other regions in the country and, when accepted, the student will start living the region of the school that accepted him, the applications could have been sent by a relative that lives far away, or the student's address was recorded wrongfully. The latter is actually a certainty in some cases, as there are applications with a distance of over 13,000 kilometers, and the length of the country is 4,270 kilometers. As these cases are indistinguishable, it becomes a necessity to give the variable a treatment. It will be assumed that applications with a distance strictly over 150 kilometers were a result of incorrect recording, which accounts to 6.2% of the total student-school distances. These data points will be replaced by the mean of the choice set's average distance between all the remaining applications.

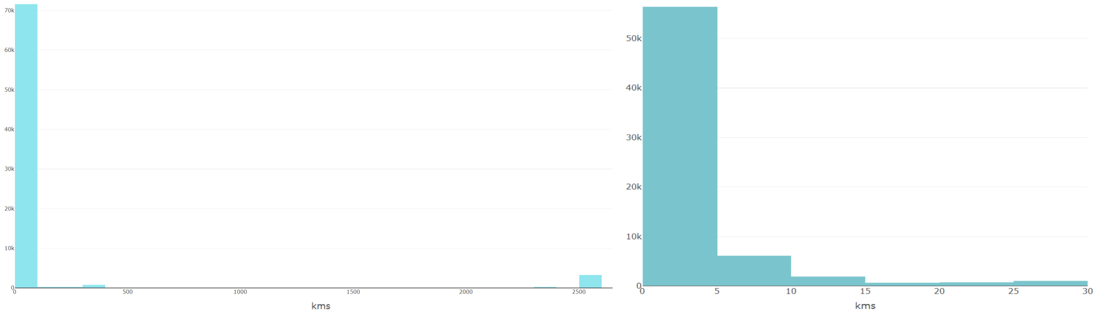


Figure 3.13: Distribution of distances for the applications in the Coquimbo region (*Left*) In bins of 100 kms, up to 3,000 kms of distance (*Right*) In bins of 5 kms, up to 30 kms of distance

3.3 Preferences

There was a total of 73,603 applications for schools in the Coquimbo region, with a mean of 3.66 applications per student, which is a bit higher than the 3.59 mean coming from the whole sample. Student must declare at least 2 preferences with the exception for rural zones, where they can declare only 1. The distribution of ranked schools is shown in figure 3.14, and ranges from 1 to 27 declared preferences. 73.0% of the students apply to 2 or more schools, 47.0% apply to more than 3, and 28.3% rank more than 4. This high rate of decrease shows that only a handful of families rank a “high” number of schools, which accounts to the result of the matching in terms of students that were not assigned, as was briefly discussed in chapter 1, and will also affect the preference model, which benefits from having more ranked schools (when revealing the true preferences), specially for the cases where only a few schools were ranked.

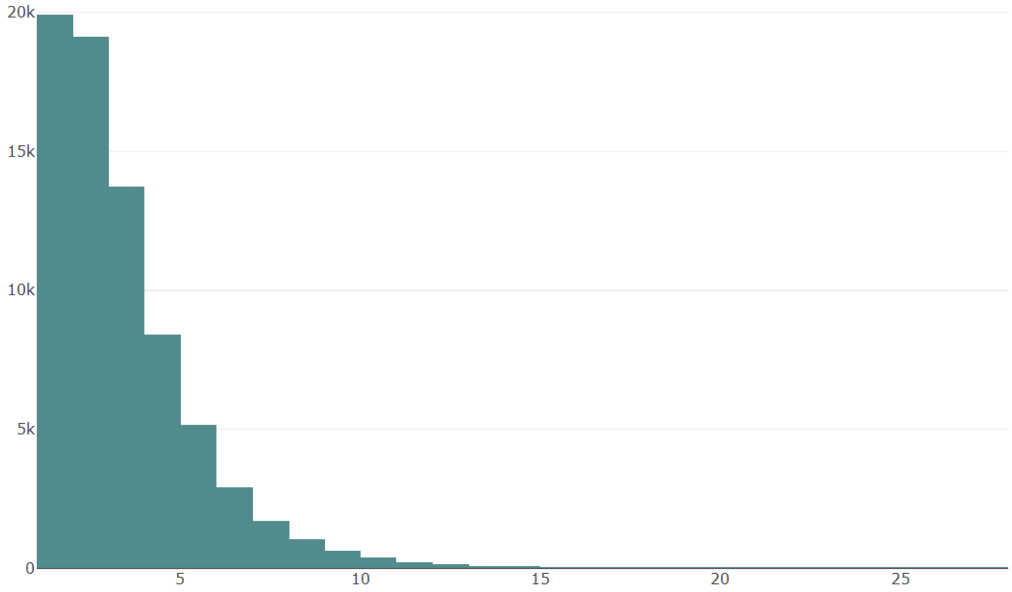


Figure 3.14: Histogram of ranked schools in the Coquimbo region

3.4 Specification of the Model

In this section the specification of the model that will be used in all the following chapters is presented, and was designed in accordance with (1) the data and patterns already described in the previous sections, (2) interest in following some questions or hypothesis about the families behaviours, and (3) studying variables of interest that could help in the design of public policies, or at the least, have a study that quantifies the effect that some covariates have in school choice so they can be taken into account when designing these policies.

The model follows the notation from chapter 2 where the latent utility models where presented:

$$w_{is} = \alpha_s + \sum_{k=1}^{m} \beta_k X_s + \sum_{k=m+1}^{m+m'} \beta_k \boldsymbol{X}_{is} + \varepsilon_i$$

w_{is} is the latent utility of student i for school s, α_s is a fixed effect for each school to account for the variables that are not present in the model and how all the students on average value that, X_s are school covariates, and $\boldsymbol{X}_{is}$ are covariates that depend on the school and the particular student, such as interactions or the distance between school s and student i. The $\boldsymbol{\beta}$ vector is the one that it is aimed to estimate. This model is actually divided into 2 separate models: one with fixed effects per school and the $\boldsymbol{X}_{is}$ matrix of covariates, and one with an intercept, α, the vector of school covariates X_s and also $\boldsymbol{X}_{is}$. This is because it is suspected from preliminary simulations that the non-observable attributes captured by the fixed effects could be collinear with the observable attributes specified for the schools. This measure is a conservative one, and it will be addressed again in section 4.5.2.

The variables represented in X_s and $\boldsymbol{X}_{is}$ are detailed next in table 3.2.

		Variable
X_s	(1)	number of applicants over number of declared seats for school s
	(2)	number of declared seats by school s
	(3)	if school s allows both genders, 1, or only one gender, 0
	(4)	if school s is of medium low SEG
	(5)	if school s is of medium SEG
	(6)	if school s is of medium high SEG
	(7)	if school s is of high SEG
	(8)	if school s is rural, 1, if urban, 0
	(9)	price of school s
	(10)	if school s is in the higher 25% for the SIMCE tests
	(11)	if school s is in the lower 25% for the SIMCE tests
	(12)	PDIS score for academic self-esteem for school s
	(13)	PDIS score for climate of coexistence for school s
	(14)	PDIS score for participation and citizen formation for school s
	(15)	PDIS score for healthy habits for school s
$\boldsymbol{X}_{is}$	(16)	priority dummy for student i & dummy of medium low SEG for school s
	(17)	priority dummy for student i & dummy of medium SEG for school s
	(18)	priority dummy for student i & dummy of medium high SEG for school s
	(19)	priority dummy for student i & dummy of high SEG for school s
	(20)	level dummy for student i & rural dummy for school s
	(21)	AE dummy for student i & dummy of 25% higher SIMCE tests for school s
	(22)	AE dummy for student i & dummy of 25% lower SIMCE tests for school s
	(23)	priority dummy for student i & price of school s
	(24)	PIE dummy for student i & price of school s
	(25)	level dummy for student i & price of school s
	(26)	distance between student i & school s

Table 3.2: Covariates for the preference model

Some variables that merit an explanation for including them are:

- **(1)** to measure the "popularity" of a school and to see if it has an effect on the resulting ranking. Also, it is represented this way instead of using the number of applicants alone to avoid an endogenous effect with the rankings
- **(2)** to check if families consider the capacity of the schools when applying. Since the assignment is strategy-proof, they should not, as it would not make any difference. The effect of this variable is also made cleaner by including the "popularity" variable
- **(20)** to try to account for the effect that could come from rural schools being valued differently by families for "kids" and for "teenagers"
- **(21)-(22)** to capture if this particular segment, the students with high grades, value differently than the rest the academic quality of schools
- **(23)-(25)** to capture if these particular segments value differently a school that can discriminate by price, and also to make cleaner the effect of price alone when it is included

Now, with the variables and specification of the model all set, the latent utility models proposed in chapter 2 will be estimated and the results analyzed in the following chapters.

Chapter 4

Clustering of Choice Sets and Estimating Preferences

In this chapter we study the definition of the choice sets. Then, we estimate student preferences for the Limarí province. The reason for working only with this province will be detailed in section 4.1, but the main idea is to use this as a baseline for choosing the methodology that will be applied. Evaluating all the decisions for how the preference models will be estimated for the whole region is too computationally expensive, and the second biggest province in the Coquimbo region should provide a good enough approximation to which methodology works best.

4.1 Clustering of the Choice Sets

The first question to tackle is if the preference model should and could be jointly estimated for the whole region. As was previously discussed, it does not seem logical to estimate with the available data how would a family that lives in the Choapa province value a school in the Elqui province because the distance could be more than 380 kilometers[1]. The admissions process data backs this intuition, as only 0.78% of the applicants ranked a school in 2 different provinces, and 0.005% applicants ranked schools in the 3 provinces. A simultaneous evaluation of all schools not only seems innecessary, but also increases dramatically the computational complexity of the estimation routines. As matter of fact, the problem becomes intractable as more schools are considered in the choice sets on the families. As will be shown in section 4.4, which contains the results of estimating the models for the Limarí province, considering choice sets of more than 30 schools takes days for the Gibbs sampler to finish iterating—and that is only one choice set.

Revisiting these two arguments results in that covering a whole province in one swipe still is not good enough. The smallest province, Choapa, has more than 10,000 km^2 and 79 schools, and Limarí and Elqui have about 30% and 60% more area and almost 2 and 2.6

[1] Distance obtained via Google Maps

times as many schools, respectively. Therefore, clustering approaches are deemed necessary to obtain choice sets for the families that: (1) make sense in terms of distance between students and schools and (2) are feasible to estimate.

The first and most natural approach would be to use the districts or communes, which are shown in table 3.1. The benefit of this approach is that the clusters are fixed and need not be estimated, but the clusters are blind to how the applications are allocated.

The second proposed clustering is obtained via unsupervised learning over a representation of the applications data, i.e., given the data $\boldsymbol{X}$, a representation $\hat{\boldsymbol{X}}$ is computed to then optimize some criterion that will produce a mapping from the schools, S to a fixed number of clusters, C: $f(\hat{\boldsymbol{X}}) : S \to C$. The representation consists in using an occurrence matrix to leverage where the actual applications are allocated. This matrix contains all the intersections of applicants whom apply to two particular schools. Let $\boldsymbol{A}$ be the occurrence matrix of dimensions $(|S|, |I|)$, where each row has a value of 1 for every student that submitted an application for the school, and 0 otherwise. A matrix that contains the intersection of applications for every school can be obtained from $\boldsymbol{A}$: the co-occurrence matrix. Let $\boldsymbol{V}$ be the co-occurrence matrix of dimensions $(|S|, |S|)$, computed as

$$\boldsymbol{V} = \boldsymbol{A} \cdot \boldsymbol{A}^T \tag{4.1}$$

The representation used to perform the actual clustering uses the co-occurrence matrix $\boldsymbol{V}$ and also incorporates a spatial dimension. For each school, a geographic location (i.e. a latitude and a longitude) is computed as shown in equations 4.2 and 4.3, where $\boldsymbol{lat}$ is a vector containing the latitudes for all schools, and $lat^*{}_s$ is the entry corresponding to school s, and equivalently $lon^*{}_s$ is computed for the longitude terms of each school. This produces a position for each school that is weighted by the position of other schools with common applicants.

$$lat^*{}_s = \frac{\boldsymbol{lat} \cdot V_{(s,:)}}{||V_{(s,:)}||_1} \tag{4.2}$$

$$lon^*{}_s = \frac{\boldsymbol{lon} \cdot V_{(s,:)}}{||V_{(s,:)}||_1} \tag{4.3}$$

The resulting clusters are shown in section 4.3, but the justification for using these 2 schemes is the following: clustering by communes uses only geographic information that is already well established, and clustering by the co-occurrence matrix attempts to form choice sets for similar families using how they allocate their preferences, and at the same time mixing it with the geographical locations.

4.2 Evaluation Metrics

Before going any further, a series of evaluation metrics for both the clustering methods and for the latent utility models is proposed. These metrics seek to give a sense of how well formed could the choice sets be, and how precisely are the preference models fitting the data for each model specification.

The first metric proposed next evaluates the clusters by measuring how much information is contained inside the clusters and is not lost with the other clusters.

- **Out-of-cluster applications**: measures how many applications leave each cluster and therefore cannot be used when estimating a particular choice set. An OCA equal to 0 means that the cluster is completely self-contained and no students are applying to schools outside that cluster.

$$OCA = \frac{\sum_{c' \neq c} \sum_{i,s} \sum_{s' \neq s} \mathbb{1}\{\text{if } i \text{ ranks } s \text{ in cluster } c \text{ and } s' \text{ in cluster } c'\}}{\sum_{i,s} \mathbb{1}\{\text{if } i \text{ ranks } s \text{ in cluster } c\}} \tag{4.4}$$

This metric has the obvious fault that is perfect when all the schools are considered at once, so it should be measured against the number of schools per cluster, aiming to more reduced choice sets that preserve as much information as possible.

Next, a set of metrics is proposed that is more standard in terms of trying to ascertain the fit of the model and can give a direct comparison between models. It is important to remark that the OCA metric is independent of the model for estimating preferences to be used, while the next metrics evaluate the models.

- **Accuracy**: gives a general measure of the number of hits in terms of predicting if a student is going to rank a school or not.

$$Accuracy = \frac{\sum_{i,s} \mathbb{1}\{\text{if } i \text{ ranks } s \text{ and } w_{i,s} > 0\} + \mathbb{1}\{\text{if } i \text{ does not rank } s \text{ and } w_{i,s} < 0\}}{|I| \cdot |S|} \tag{4.5}$$

- **Recall**: measures if the model is retrieving the actual ranked schools.

$$Recall = \frac{\sum_{i,s} \mathbb{1}\{\text{if } i \text{ ranks } s \text{ and } w_{i,s} > 0\}}{\sum_{i,s} \mathbb{1}\{\text{if } i \text{ ranks } s\}} \tag{4.6}$$

A separate metric is proposed that cannot be used to compare models, as it will measure the performance of the resulting ranking provided by the multivariate ordered probit. This metric is necessary because the metrics already presented do not measure rank, which is the way the data is originally presented and it is still important, even if not comparable, to get a sense of how well is the rank being represented.

- **Spearman correlation**: measures the relationship between two ordinal variables by how well described they can be using a monotonic function, meaning that it will measure the direction of increase or decrease of the submitted rank, y_i (that includes at the last position the schools that were not ranked), and the predicted rank, $\hat{y}_i$.

$$Spearman\ correlation = \frac{1}{|I|} \cdot \sum_i \frac{\text{cov}(y_i, \hat{y}_i)}{\sigma_{y_i} \cdot \sigma_{\hat{y}_i}} \tag{4.7}$$

4.3 Instances for the Limarí Province

Some preprocessing of the choice sets and preference lists is needed, since considering schools with only 1 application, for example, results in a heavier computational burden (for each iteration latent utilities must be sampled for all students for that particular school) and the extreme tail of the distribution could have a noisy behaviour that affects the overall results. Also, in the ideal case, all of the preferences would be contained in each cluster, meaning that no student is considering or applying to some school outside his or her choice set, but this is not possible when using a clustering approach.

To tackle the first problem, only schools with at least 3 applicants will be considered when the cluster has 20 schools or less, and schools with at least 15 applicants will only be considered for the rest of the clusters (**School Filter**). For the second problem, a heuristic is proposed that consist in trying to contain as much from each preference list while trying to cut as least students as possible. This is done by setting that no more than 25% of students applying to a cluster can be cut, while maximizing the minimum length of the preference list (**Student Filter**), meaning that if the filter is set to more than 50%, only students with more than half of their preference list of ranked schools is within the cluster will be considered. It is important to note that this does not necessarily means that a student that is cut will never be consider, only that the student may be used in another cluster where more complete information is had for his or her preference list.

4.3.1 Communes

The Limarí province has 5 communes: Combarbalá, Monte Patria, Ovalle, Punitaqui and Río Hurtado. A map of the schools divided into these districts is shown in figure 4.1.

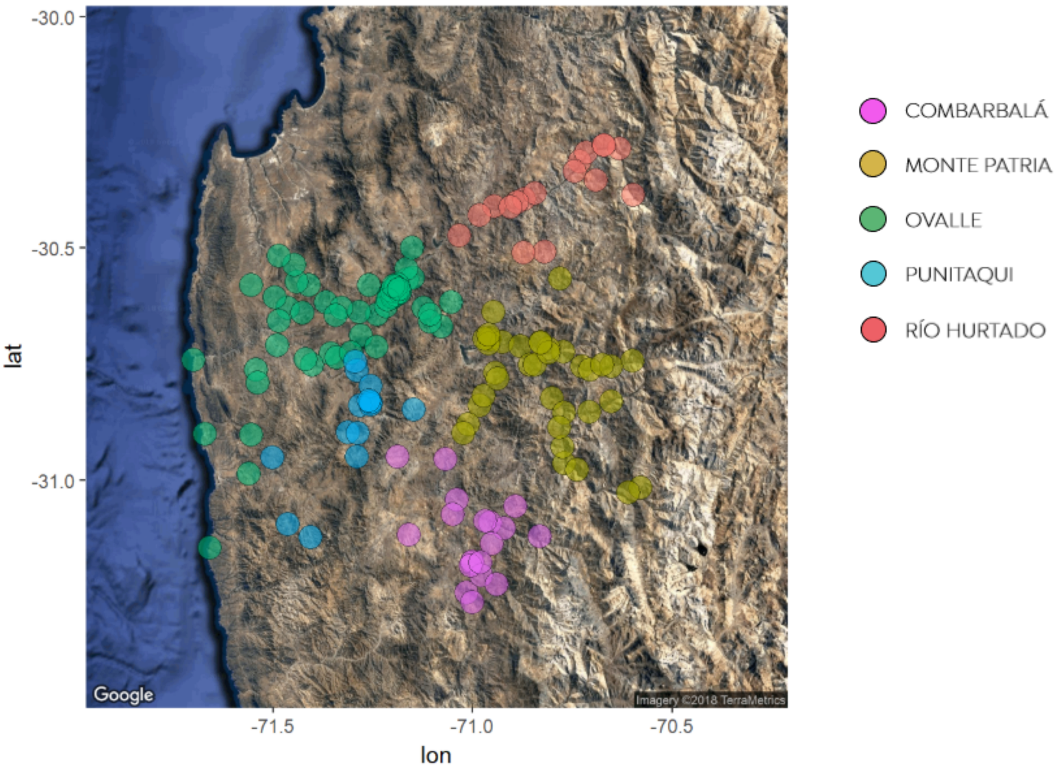


Figure 4.1: Schools by commune in the Limarí province

The metrics corresponding to these divisions and the filters applied in terms of minimum applicants per school and minimum length of preference list are presented in table 4.1. The column "School Filter" refers to the rule applied to consider schools that have a minimum amount of applicants, and the next column shows the resulting number of schools from this filter that will be used for the choice sets. The column "Student Filter" shows the minimum percentage of ranked schools in the cluster of the total amount of ranked schools that will be considered for each student, filtering preference lists that are too incomplete for a particular cluster. The next column shows the number of applications for schools in that cluster after applying these filters, and the final column shows the out-of-cluster applications metric presented in section 4.2.

Cluster	School Filter	# Schools	Student Filter	# Applications	OCA
Combarbalá	≥ 3	13	100%	468	.18
Monte Patria	≥ 15	16	$> 50\%$	1,496	.47
Ovalle	≥ 15	36	100%	13,964	.05
Punitaqui	≥ 3	8	$> 50\%$	471	.52
Río Hurtado	≥ 3	9	$> 65\%$	100	.41

Table 4.1: Clustering by commune summary

The first thing that stands out is that the Ovalle commune is significantly larger than the rest of the communes, as it has more than twice the number of schools and about 10 times the number of applications than the second largest commune. It can also be noted that, as a reflection of the OCA, the two communes with lowest OCA could filter only for students that had their whole preference list inside the commune, and the two communes with largest OCA had to allow for students with only half or more of their preference list inside the cluster.

4.3.2 Clustering by Applications Centroids

As discussed in section 4.1, a geographic location is computed for each school to be used as a representation of the school's location and applications allocation for a clustering algorithm. Figure 4.2 shows the location of schools in the Limarí province and also the locations of the same schools after computing the representations of the applications centroids, only for schools with at least 3 applicants. It can be seen that most of the schools tend to move towards centers in the middle of the province, and one in the south zone of the province. Also, some schools tend to remain in their original far-off location. It is worth mentioning that a couple of schools did not have shared applications with any other school, so their location was not moved at all.

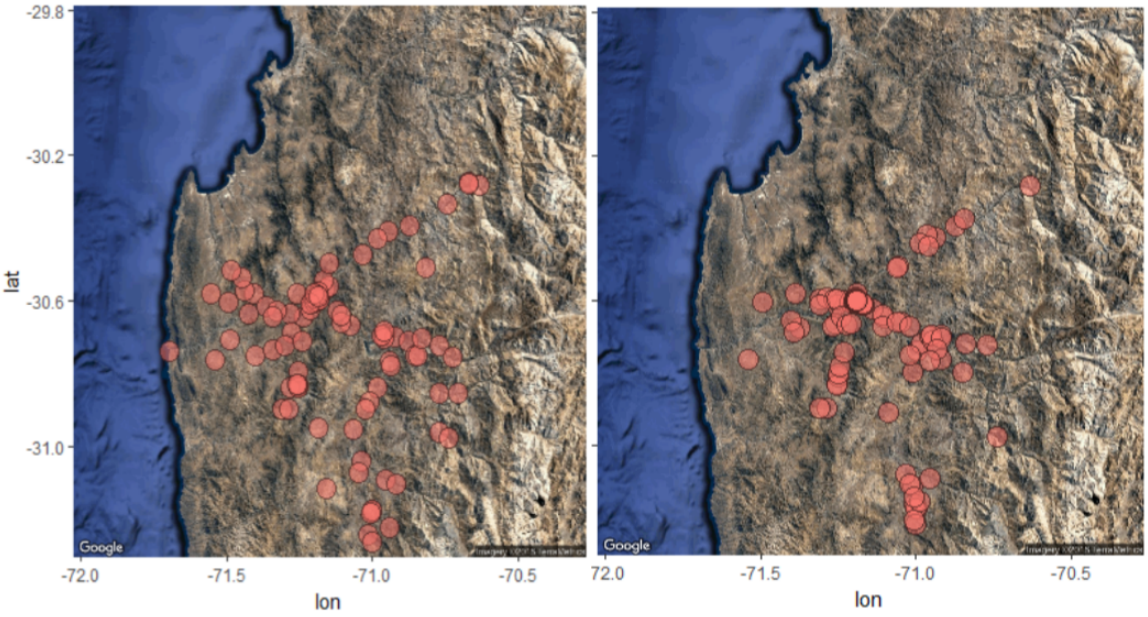


Figure 4.2: (*Left*) Schools in the Limarí province with at least 3 applicants (*Right*) Same schools after representing each school by its applications centroid

The clustering technique applied was hierarchical clustering with Ward's linkage, which is a form of agglomerative clustering where, at each iteration, the two clusters with minimum intra-cluster variance are merged into one cluster, beginning with all data points as an individual cluster. The measure of distance from which the variance is computed is the Euclidean distance between any two data points. The number of clusters is a hyperparameter to be tuned.

The result of applying this technique is shown in figure 4.3 for 5 clusters of schools, where first the computed clusters are shown in their original location, and next the points in the representation previously mentioned where one can see how the clustering algorithm performed the segmentation.

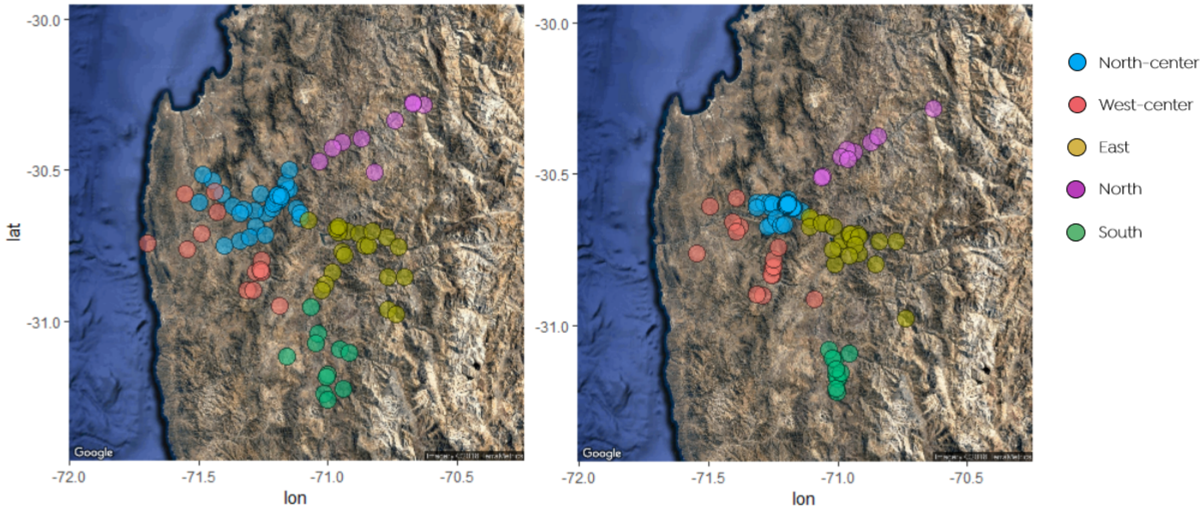


Figure 4.3: (*Left*) Schools in the Limarí province with at least 3 applicants clustered by the applications centroids (*Right*) Representation of the schools to which the clustering algorithm is applied

The details of these 5 clusters, such as the number of applicants, schools and the OCA metric, are shown in table 4.2. It can be seen that the OCA is similar to the one obtained for the communes for the resulting clusters, and more so, some clusters turned out to be communes or very close to some beign exactly one commune. This shows that at least the communes can be recovered (if the allocations of preferences allows so) and more flexibility can be obtained by clustering, because the number of clusters can be increased or decreased. The fact that again a really big cluster is obtained is suspected to be a result of a very difficult problem—the clustering of the co-occurrence matrix—in this particular setting, because the applications can be very intertwined between students and schools and it simply does not happen that there are easy clusters where all the applications are contained. A dimensionality reduction via multi-dimensional scaling was performed on the co-occurrence matrix (shown in annex 4), where it can be seen that, apart from a few outliers, the vectors of intersections of the applications for the schools are clustered tightly together, backing up the intuition that this clustering problem was difficult to solve in a way that is completely satisfactory.

Cluster	School Filter	# Schools	Student Filter	# Applications	OCA
North-center	≥ 15	36	100%	14,026	.05
West-center	≥ 3	15	$> 50\%$	576	.66
East	≥ 15	16	$> 50\%$	1,496	.48
North	≥ 3	9	$> 65\%$	100	.44
South	≥ 3	11	$> 50\%$	390	.26

Table 4.2: Clustering by applications centroid summary

4.4 Results for the Limarí Province

In this section, the results of estimating the preference models is presented for the two clustering schemes, two specifications of the models, and the two structural models with the objective of evaluating all of them before applying a particular methodology for the whole region. As was previously mentioned in section 3.4, there are two model specifications: considering a fixed effect for each school in the choice set and interactions between schools and student covariates (**FE**), and considering an intercept with all the school covariates and interactions (**SC**), which were presented in table 3.2. The results presented next show the Markov Chain Monte Carlo simulations, all of which have a burn-in period of 2/3 of the simulations, thus reporting only the remaining third of the 1 million iterations performed for each model, specification and cluster. Table 4.3 presents the results for the communes and table 4.4 reports the results for the clustering by applications centroid.

Model	Specification	Cluster	Time [hrs]	Accuracy	Recall	Spearman
MVOP	FE	Combarbalá	.8	88.3%	34.6%	.44
MVP	FE	Combarbalá	2.5	88.3%	34.8%	-
MVOP	SC	Combarbalá	.8	88.5%	44.9%	.47
MVP	SC	Combarbalá	2.1	88.3%	62.0%	-
MVOP	FE	Monte Patria	3.6	87.0%	6.2%	.35
MVP	FE	Monte Patria	8.7	87.0%	6.6%	-
MVOP	SC	Monte Patria	3.2	87.0%	6.5%	.35
MVP	SC	Monte Patria	7.5	86.8%	6.8%	-
MVOP	FE	Ovalle	30.3	89.0%	1.1%	.16
MVP	FE	Ovalle	30.9	89.1%	1.3%	-
MVOP	SC	Ovalle	20.9	89.4%	5.9%	.28
MVP	SC	Ovalle	22.9	89.4%	6.2%	-
MVOP	FE	Punitaqui	.2	82.6%	56.9%	.60
MVP	FE	Punitaqui	.9	82.6%	57.7%	-
MVOP	SC	Punitaqui	.3	83.5%	64.8%	.62
MVP	SC	Punitaqui	1.0	83.5%	64.8%	-
MVOP	FE	Río Hurtado	.1	87.0%	39.0%	.52
MVP	FE	Río Hurtado	.5	87.2%	43.0%	-
MVOP	SC	Río Hurtado	.1	86.8%	39.0%	.51
MVP	SC	Río Hurtado	.5	86.7%	43.0%	-

Table 4.3: Preference estimation results for the Limarí province clustering by communes

Model	Specification	Cluster	Time [hrs]	Accuracy	Recall	Spearman
MVOP	FE	North-center	32.7	89.2%	0.9%	.15
MVP	FE	North-center	32.4	89.3%	1.5%	-
MVOP	SC	North-center	20.1	89.1%	6.2%	.29
MVP	SC	North-center	23.8	89.3%	6.1%	-
MVOP	FE	West-center	3.1	91.0%	28.3%	.77
MVP	FE	West-center	3.1	92.1%	41.3%	-
MVOP	SC	West-center	2.8	92.6%	55.9%	.72
MVP	SC	West-center	3.4	93.1%	55.9%	-
MVOP	FE	East	7.9	86.7%	1.9%	.30
MVP	FE	East	7.2	87.0%	6.7%	-
MVOP	SC	East	7.6	87.0%	6.5%	.35
MVP	SC	East	6.3	86.9%	7.1%	-
MVOP	FE	North	.3	87.0%	39.0%	.52
MVP	FE	North	.3	87.2%	43.0%	-
MVOP	SC	North	.3	86.8%	39.0%	.51
MVP	SC	North	.3	86.7%	43.0%	-
MVOP	FE	South	1.6	89.7%	48.2%	.51
MVP	FE	South	2.0	89.6%	48.2%	-
MVOP	SC	South	1.7	89.6%	57.7%	.54
MVP	SC	South	2.1	89.6%	48.5%	-

Table 4.4: Preference estimation results for the Limarí province clustering by applications centroids

The results seem competitive in the sense that it is not immediately clear if the MVOP provides significant improvements in accuracy and recall overall, and also the communes and clusters are fairly similar, which would evidently produce similar results. Although it should be noted that the Fixed Effects specification struggles in some clusters and is unable to outperform the School Covariates specification in most cases. This is addressed in more detail in section 4.5.2. The high Spearman correlation in some choice sets indicate that the relative order of rankings is being recovered with the revealed preferences, most notably in the west-center cluster.

The high accuracy and low recall obtained for the largest clusters suggest that there is not enough information at the individual level (i.e. interactions) to correctly learn all the selected preferences by each student, which is also an effect of having classes that are heavily unbalanced (a lot more schools are not selected in comparison to the ones the students rank for each choice set) and results in high accuracy but low recall. However, the school's attributes are still retrieved by the samplers (see annex 5 to review the convergence of some covariates). It is important to remark that the problem that the MVOP tries to solve is much harder, since it has to correctly order the estimated preference lists while also predicting which schools are going to be in each preference list. Second, the data is noisy as a result of dividing the choice sets and thus, some preference lists are subsetted, only remaining the relative ordering of preferences but not the *true* ordering of preferences. This affects the MVOP underlying assumptions but not the MVP's, because the latter only has to figure out if a school is going to be in a preference list or not, regardless of its true order.

By now several times it was mentioned that the problem of estimating preferences considering too big choice sets was intractable, and from the results one can see how the problem scales as more schools are considered in the choice sets, and the execution time does not scale linearly. This comes from whenever a school is added to a choice set, for each applicant a latent utility must be estimated in every iteration. The Ovalle commune and the north-center cluster would take about 4 days for each of the instances to be completed, so, to address this issue, 20% of the applicants for each of these instances were selected randomly to be able to obtain the results in a reasonable amount of time—although it can still take more than an entire day. Considering that this instance has only 36 schools and the Limarí province has 154, not to mention that the whole region has 436 schools, the problem is computationally infeasible for bigger choice sets.

Before concluding which methodology is preferred, the results of an instance will be detailed. In table 4.5, the estimated coefficients with their respective uncertainty estimates and significance levels are reported for the west-center cluster for the MVP and MVOP for the SC specification. They are also compared with a multinomial probit that uses only the first ranking (see section 4.5.1 for more details).

	MNP			MVP			MVOP		
Variable	Mean		Std	Mean		Std	Mean		Std
(0) Intercept	5.40		9.48	5.67		9.09	9.06		8.64
(1) Popularity	1.23	·	0.64	1.09	·	0.59	1.10	*	0.52
(2) Seats	0.01		0.03	0.02		0.03	0.01		0.03
(3) School gender	-		-	-		-	-		-
(4) Medium-low SEG	-6.45	***	1.75	-3.54	*	1.77	-2.25		1.42
(5) Medium SEG	-		-	-		-	-		-
(6) Medium-high SEG	-		-	-		-	-		-
(7) High SEG	-		-	-		-	-		-
(8) Rural	-2.26		3.59	-6.27		3.82	-4.55	·	2.77
(9) Price	-		-	-		-	-		-
(10) Top SIMCE	1.95		1.85	0.94		2.04	0.56		1.89
(11) Bottom SIMCE	-4.03	***	1.85	-3.33	*	1.37	-2.46	**	0.84
(12) PDI self-esteem	0.42	*	0.19	0.26		0.22	0.14		0.16
(13) PDI climate	0.19		0.15	0.38	*	0.19	0.17		0.16
(14) PDI participation	-0.59	*	0.26	-0.61	*	0.29	-0.37		0.25
(15) PDI habits	-0.06		0.33	-0.09		0.30	-0.05		0.23
(16) PRI - medium-low SEG	-0.09		0.37	0.04		0.41	-0.02		0.36
(17) PRI - medium SEG	-		-	-		-	-		-
(18) PRI - medium-high SEG	-		-	-		-	-		-
(19) PRI - high SEG	-		-	-		-	-		-
(20) Level - rural	1.04	·	0.59	2.89	*	1.13	2.33	**	0.77
(21) AE - top SIMCE	-4.37	·	2.43	-1.02		0.95	-1.22		0.88
(22) AE - bottom SIMCE	-0.54		1.07	0.67		1.33	0.53		0.91
(23) PRI - price	-		-	-		-	-		-
(24) PIE - price	-		-	-		-	-		-
(25) Level - price	-		-	-		-	-		-
(26) Distance	-0.10	***	0.01	-0.10	***	0.01	-0.08	***	0.01
Significance level: (·) 90%, (*) 95%, (**) 99%, (***) 99.9%									

Table 4.5: Results for the west-center zone cluster, and school covariates specification for the MNP, MVP and MVOP models

It is important to notice that the results from the models should not be analyzed under the same light. Although they are similar in that they model latent preferences of students for schools in their choice sets, the fact that the MVOP considers the ranking produces a different interpretation for the estimated coefficients (see section 2.3). Nonetheless, the 3 models produce similar parameter estimates with similar levels of uncertainty around their means.

First, the MNP produces some differences with both the MVP and MVOP that would be a result of not considering the rest of the rankings. Most notably, the popularity, rural and level-rural interaction covariates reduce their significance levels, meaning that their effect on including schools with these characteristics in the preference lists when considering only first rankings is less conclusive. Being a low performing SIMCE school becomes a more relevant characteristic in the first choice, as well as being a medium-low school in socioeconomic terms. The interaction between top SIMCE schools and academic excellence students (AE) becomes more significant as well with a negative effect—which would be counterintuitive—, and the academic self-esteem of the school recovered had a positive effect.

Leaving the MNP behind, it was obtained that both the MVP and MVOP models produce very similar statistical fits, and they agree in the means, standard deviations and statistical significance for the most part. One difference, for example, is that the MVP shows that schools of medium-low socioeconomic group are valued negatively with respect to schools of low socioeconomic group, while the MVOP shows that there is no concluding effect. They both agree that popularity and rural schools for younger students have a positive impact on the preference lists, while being a lowest performing school on the SIMCE test will have a negative impact. They produce very similar results regarding the distance variable, which has a negative and highly significant effect—as would be expected. The MVOP shows some confidence regarding the estimated preferences for rural schools, saying those schools will in general be ranked lower than urban schools, and shows also higher significance regarding that popular schools will be ranked higher than what the MVP accounts for. The MVP shows a statistically significant effect for two of the personal development indicators, saying that schools with better climate of coexistence are preferred, while schools where the parents have high citizen participation are not. The MVOP shows there is no concluding effect regarding these two variables.

In this specification, the intercept represents the valuation of including a fictional school that has all covariates set to 0—although it is not possible for the seats variable to be 0—, and for all models the mean valuation is not significantly different from 0.

In figure 4.4 convergence plots are shown for some of the covariates in the MVOP model presented before. This serves to illustrate that the Gibbs sampler converges to the stationary distribution—in these cases very quickly—and to show graphically the uncertainty about the parameter estimates presented in table 4.5.

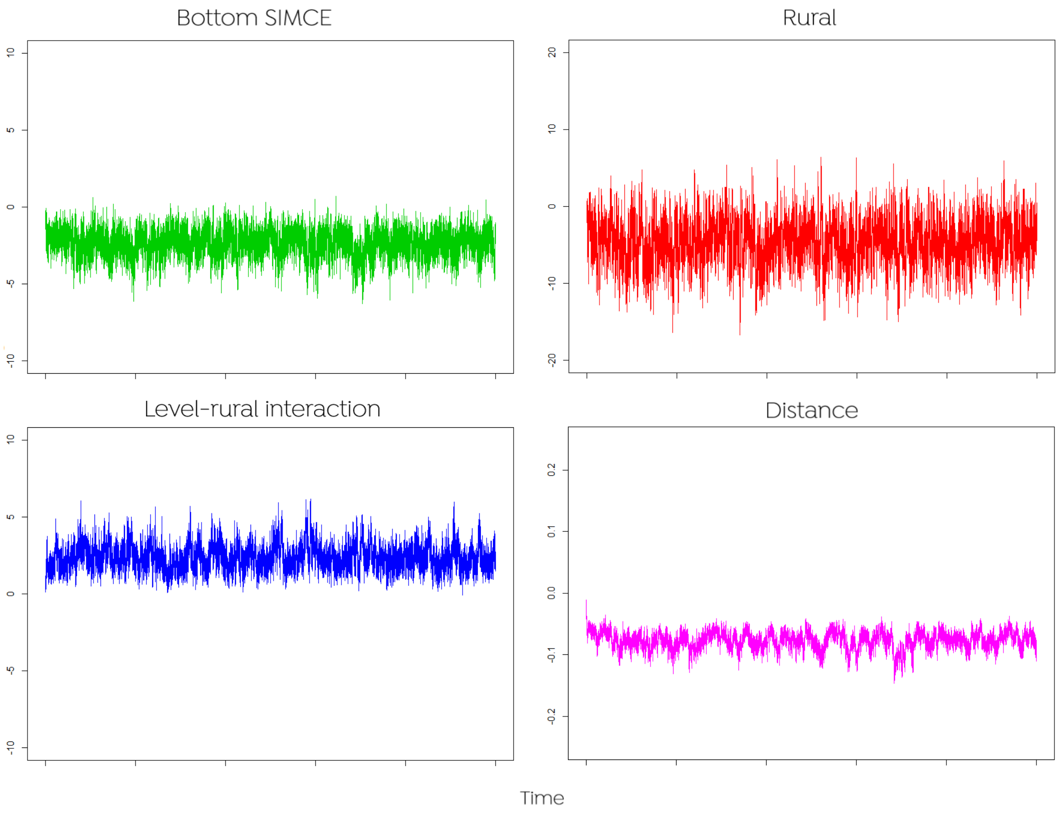


Figure 4.4: Convergence plots of the Gibbs sampler for the MVOP model with the west-center zone instance, for the bottom SIMCE, rural, level-rural interaction and distance covariates

To summarize, both the MVP and MVOP models produce very similar statistical fits and very similar conclusions regarding the effects of the different specified covariates. While there are some differences, they are to be expected since the two models are not modeling the exact same underlying processes by which the preference lists are created.

Now, to choose which specification and clustering scheme is adequate moving forward with the whole region of Coquimbo, the evaluation metrics are obtained as a weighted average of the amount of applications per instance. Table 4.6 reports these summary statistics.

Clustering	Model	Specification	Accuracy		Recall		Spearman	
Communes	MVP	Fixed Effects	88.7%	8°	4.6%	6°	-	
		School Covariates	89.0%	6°	9.7%	**1°**	-	
	MVOP	Fixed Effects	88.6%	8°	4.3%	7°	0.20	3°
		School Covariates	89.0%	5°	8.9%	4°	0.30	2°
Applications Centroids	MVP	Fixed Effects	89.2%	4°	4.7%	5°	-	
		School Covariates	89.2%	2°	9.1%	3°	-	
	MVOP	Fixed Effects	89.2%	3°	3.3%	8°	0.20	4°
		School Covariates	89.3%	**1°**	9.6%	2°	0.32	**1°**

Table 4.6: Summary results for the Limarí province

Although there is no absolute winner, the MVOP with School Covariates specification on the applications centroid clustering achieves the best results in accuracy and Spearman correlation, with a close second place in recall. But this is not the only reason to choose this setting for the rest of the region. The MVOP is a generalization of the MVP, so it is expected to attain at least some improvement in statistical fit. Also, using a clustering algorithm instead of the communes provides flexibility to desing choice sets, and in this particular case it showed that it can also recover similar clusters to the communes, so, again, it should lead to improvements when finding adequate choice sets.

Consequently, from here on the MVOP will be used for the rest of the region, with an School Covariates specification, and the choice sets will be determined via an agglomerative clustering algorithm of the applications centroids.

4.5 A Note on the Specification

In this section, a couple of points are reviewed in order to provide a more satisfactory justification for some choices made during the modeling stages. First, the probit model that uses only the first ranked school will be further reviewed and used to check the effect that the rest of the rankings have on statistical fit and parameter estimates. This comes from concerns over the possibility of too noisy rankings that do not reflect the actual complete preferences of the families. Next, the same instance is estimated using jointly a fixed effect for each school and the school covariates to review the impact of the conservatory measure of separating the specifications.

4.5.1 Multinomial Probit

Fitting the MVOP only with the first ranked schools is equivalent to fitting a MVP (this was also checked in this instance for consistency). This model is also known as a multinomial probit, where out of n discrete alternatives only 1 is chosen by each decision maker. This should be an easier problem in terms of retrieving these preferences since only the most important—presumably—information regarding the rankings is provided. The accuracy and recall achieved were 95.9% and 53.0%, respectively. The recall in predicting if an applicant is going to submit their first choice was of 68.5% by both the MVP and MVOP, which is significantly higher than the recall obtained by the MNP. This suggests that the information the rest of the rankings provide is important to predict the first choice, and serves to compare more fairly the fit of the 3 models.

Revising in more detail the rankings for this cluster, it can be seen from figure 4.5 that there is substantial information being left out, as there is also noisy information that would enter the models when considering all ranked schools, most notably, a student submitted at least 11 preferences, but preferences 7 to 10 did not even end up in this cluster. It should be noted that this is a result of using any clustering scheme, as some preference lists will always be partitioned.

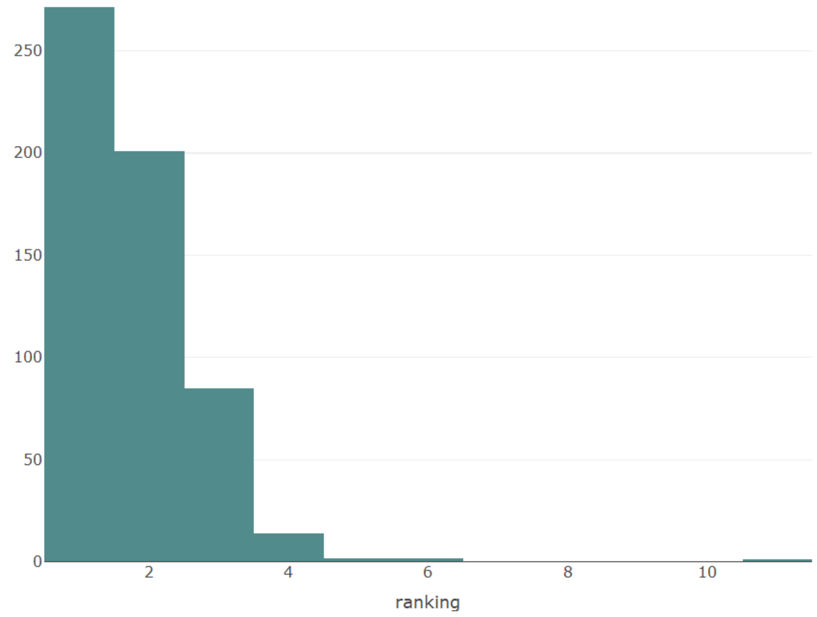


Figure 4.5: Histogram of schools by submitted ranking for the Limarí west-center cluster

In summary, it would seem that the rankings provide a great deal of information, as it would even change the meaning for some of the obtained estimates if not considered, but there is also perturbed information that results from clustering choice sets. This sheds some light in why the MVOP underperforms the MVP in some instances, but it is still not a concluding answer to this concern.

4.5.2 Fixed Effects and School Covariates Specification

The preference models were estimated for the west-center Limarí cluster using fixed effects and school covariates. The resulting accuracy and recall did not provide an improvement over considering the specifications separately , and it is a first indication that there are no gains in statistical fit from combining the different specifications.

Out of the 14 fixed effects—there are 15 schools in this cluster— none was 95% statistically significant different from 0. This would indicate that most of all characteristics of the schools that families give importance when choosing them seem to be captured by the specified covariates, meaning that mixing fixed effects will result in little information to be captured that is not already provided by the observable variables.

Since there are no obvious gains from using this mixed specification in this example, it would seem that using them separately is the right course of action for the rest of the region.

Chapter 5

Results for the Coquimbo Region

In this chapter, the estimation of school preferences is completed for the whole Coquimbo region, adding to the already estimated preferences in the Limarí province the provinces of Choapa and Elqui. By fullfilling this, a complete picture of how the families allocate their preferences for schools is obtained, accounting for factors attributable to both the schools and the families, while also accounting for other factors that could not be specified and/or are not observable that could have had an impact in the school choice decision.

Next are presented the results for each province according to the methodology proposed in chapter 4, followed by an analysis and discussion of these results.

5.1 Results for the Choapa Province

The Choapa province, being the smallest among the 3 provinces in the Coquimbo region, had 4 resulting clusters, which are displayed in figure 5.1. It can be seen from the images that some schools move far away from their original locations to a particular cluster as a result of considering its allocation of preferences by the students that are shared with other schools of the province. Most notably, schools from the north-coast cluster, which some are dozens of kilometers away, cluster tightly in the north part of Choapa. The resulting clusters are fairly homogeneous in the sense that there is no huge cluster as in the Limarí province—the center zone cluster contains 49% of the preferences, while in Limarí the north-center cluster contained 85% of the preferences—, and furthermore, the OCA metrics seem promising (see table 5.1); the weighted average of them by the number of applications is .06, while the weighted average OCA for the Limarí province was .12.

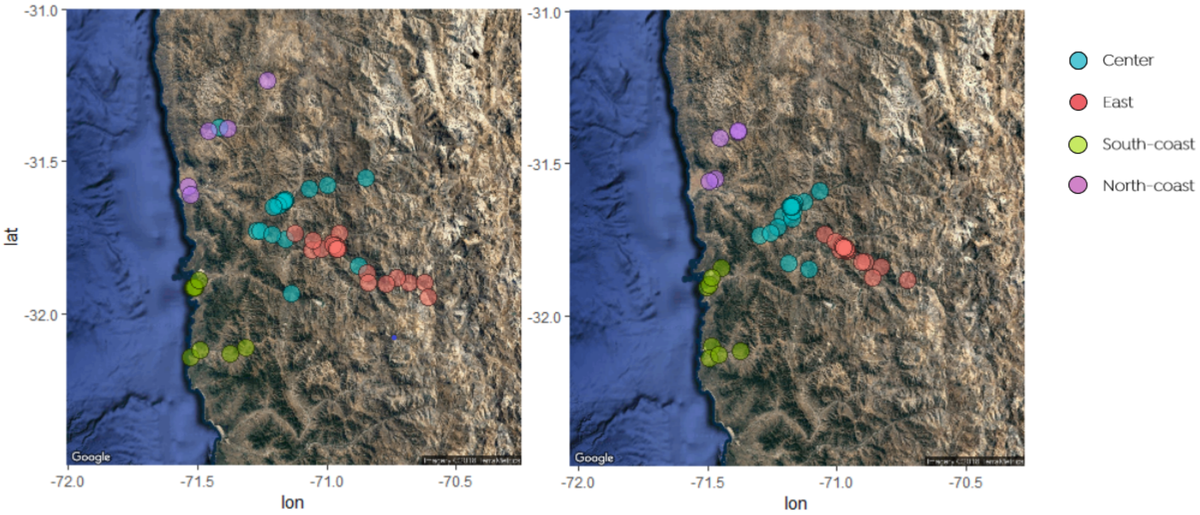


Figure 5.1: (*Left*) Schools in the Choapa province with at least 3 applicants clustered by the applications centroids (*Right*) Representation of the schools to which the clustering algorithm is applied

Cluster	School Filter	# Schools	Student Filter	# Applications	OCA
Center	≥ 15	15	100%	2,198	.07
East	≥ 15	12	100%	1,202	.04
South-coast	≥ 3	8	100%	913	.07
North-coast	≥ 3	5	100%	212	.01

Table 5.1: Clustering by applications centroid summary in Choapa

In table 5.2 the evaluation metrics as a result of using the MVOP with a school covariates specification are shown. The north-coast cluster shows exceptionally high goodness of fit in all 3 metrics, and the remaining clusters show similar results in these metrics. The east cluster showed a lower recall than the rest of the clusters that could not be diagnosed by the OCA, since it has a better OCA than the center and south-coast clusters.

Model	Cluster	Specification	Time [hrs]	Accuracy	Recall	Spearman
MVOP	Center	School Covariates	6.7	86.2%	33.6%	.44
	East		2.8	80.8%	22.7%	.31
	South-coast		2.1	80.3%	35.8%	.39
	North-coast		.2	87.9%	84.4%	.71

Table 5.2: Preference estimation results for the Choapa province clustering by applications centroids

By aggregating these metrics as a weighted average of the number of applications in each cluster, the results displayed in table 5.3 are obtained. It can be seen that a much higher average recall is achieved than in the Limarí province, which could be attributable to the lack of choice sets no greater than 15 schools, less than half of what is obtained for the biggest cluster latter province.

Model	Specification	Accuracy	Recall	Spearman
MVOP	School Covariates	83.7%	33.5%	.41

Table 5.3: Summary results for the Choapa province

5.2 Results for the Elqui Province

The Elqui province is the largest province of the Coquimbo region. It has 32% more schools participating in the admissions system than the Limarí province, and almost 2.7 times the amount of applications (ranked schools from the preference lists). This resulted in heavily unbalanced clusters, which are shown in table 5.4. The north-center and coast-center cluster sum up to almost 95% of the applications and could not be separated further by a clustering algorithm. Figure 5.2 shows that these clusters result in highly dense areas in terms of school offer, so the clustering algorithm struggled to provide more balanced clusters since the applications in these zones are not easily separable. It can be noted that the south-center and east clusters have a fairly high number of schools, but the low number of applications in comparison to the largest clusters seem to suggest these are less competitive areas.

The weighted OCA is .22, the highest of all 3 provinces. This results from the clustering algorithm not being able to cleanly separate the schools in choice sets in the center zone of the region.

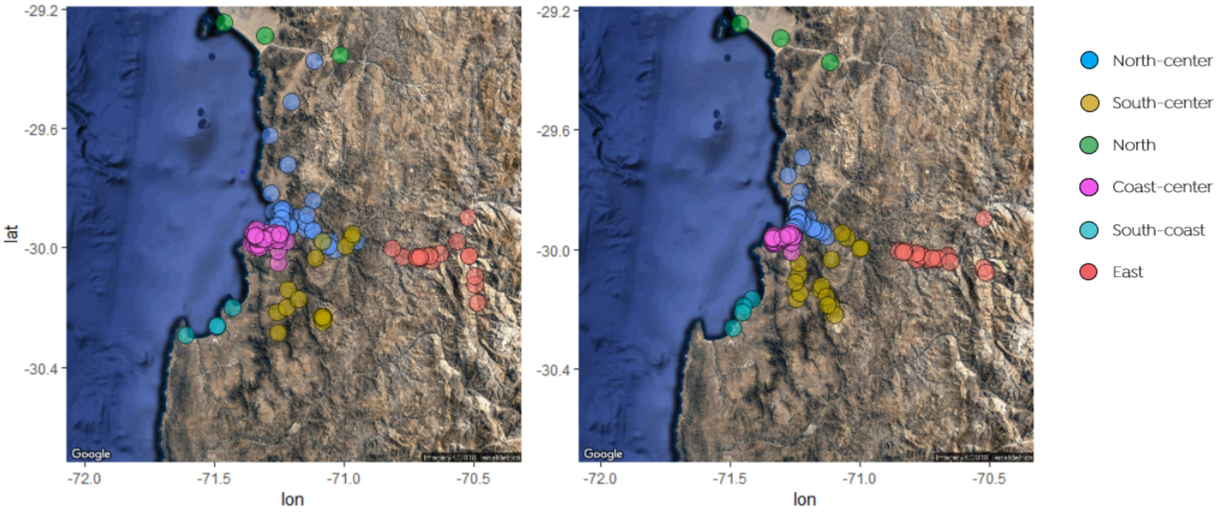


Figure 5.2: (*Left*) Schools in the Elqui province with at least 3 applicants clustered by the applications centroids (*Right*) Representation of the schools to which the clustering algorithm is applied

Cluster	School Filter	# Schools	Student Filter	# Applications	OCA
North-center	≥ 15	78	$> 85\%$	23,497	.18
South-center	≥ 3	16	$> 50\%$	624	.59
North	≥ 3	3	100%	15	.06
Coast-center	≥ 15	67	100%	18,371	.26
South-coast	≥ 3	4	100%	432	.20
East	≥ 3	19	100%	1,233	.19

Table 5.4: Clustering by applications centroid summary in Elqui

To fit the MVOP to the two large clusters of the province, 20% of the applicants were sampled randomly for each cluster—as in the Limarí province—, because otherwise the algorithm would have taken days to be completed. The results by cluster are shown in table 5.5. It can be seen that for the two largest clusters the samplers were unable to retrieve the positive valuations of the schools, thus high accuracy but recall equal to 0 was achieved. This was already diagnosed for the north-center cluster in the Limarí province, and the discussion provided in section 4.4 still applies to this setting, only now the problem was more severe. The explanation for this is fairly simple, and it results from considering 2 factors. First, a lot of information regarding the schools was provided, but scarce information about the students. As most schools are not ranked for each student, the mean valuation for any school is negative. If more detailed information regarding the students was available, a difference could have been made to study at the individual level what one student values positively and negatively, but only segments of students—or families—as a whole could be studied, with only the most specific information being the distance (which was also a corrupted variable). Second, in comparison to the school system in New York, it is mandatory to apply to at least 6 schools. This provides a lot more positive valuation than what is present in the Coquimbo assignment. In figure 3.14 it was shown that only 28.3% of the families rank 4 schools or more, thus far less information was provided than in the more mature system in the American city. Nonetheless, the sampler was again able to retrieve the desired covariates posterior distributions (see annex 6 and 7 to check the convergence of the samplers), which is what this study is mostly concerned about.

Model	Cluster	Specification	Time [hrs]	Accuracy	Recall	Spearman
MVOP	North-center	School Covariates	54.4	94.8%	0.0%	-
	South-center		3.2	90.6%	38.5%	.53
	North		.0	77.8%	66.7%	.50
	Coast-center		41.8	95.4%	0.0%	-
	South-coast		.1	90.5%	89.2%	.75
	East		6.4	89.3%	22.5%	.41

Table 5.5: Preference estimation results for the Elqui province clustering by applications centroids

The final results are displayed in table 5.6. The highest accuracy was achieved between the 3 provinces, but also the lowest recall and Spearman correlation as a result of underperforming in these metrics for the largest clusters.

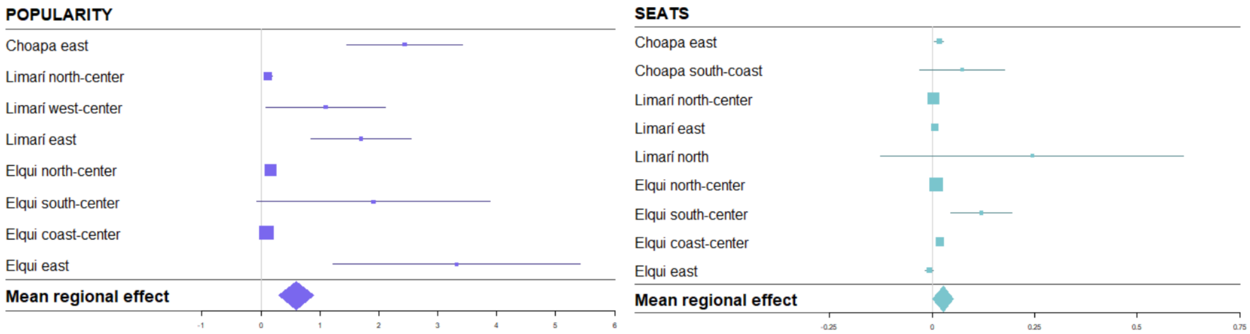
Model	Specification	Accuracy	Recall	Spearman
MVOP	School Covariates	94.8%	2.1%	.03

Table 5.6: Summary results for the Elqui province

5.3 How Do Families Value Schools?

The question of how families value schools, meaning how they appreciate different attributes of the schools and what makes them rank schools in a particular order, and how this generalizes to the entire region considering the differences that are produced by valuations in its different zones, can now be addressed. As a result of estimating the preferences for different choice sets, heterogeneity in the preferences for every attribute can be appreciated at the cluster level and at the regional level. These results are displayed via forest plots in figures 5.3 - 5.9. The results are reported only for the parameter estimates that have at least 75% of their posterior distribution greater or lower than 0. This is done to show only the parameter estimates for which the valuation has a high degree of certainty of it to be of a particular sign. However, the **mean regional effect** is also reported as the weighted average—by the number of schools in each choice set—of all the means of the posterior distribution for the parameter estimates.

Figure 5.3a displays the valuation in the different clusters for the popularity of the school, which was measured as number of applicants over declared seats (i.e. the overdemand or underdemand of the school), and figure 5.3b shows the valuation regarding the number of declared seats by a school. As it is to be expected, the popularity effect is shown to be positive, meaning that a more popular school is probably going to be ranked higher than a less popular school. Annex 1 reports that the Ministry of Education advised to include in the rankings less popular schools, and the recovered effects seem to suggest that this was not the case. Although this variable could be argued to be endogenous (since it is possibly correlated with unspecified variables that account for the over or underdemand), it serves as a control for the number of declared seats—the supply that the schools provide to system—since an explanatory analysis showed positive correlation between number of applicants and seats (see figure 3.4). The latter variable showed a positive and statistically significant effect for the majority of the clusters that can be appreciated across the entire region, providing a positive mean regional effect. This implies that a school that opens more seats is more likely to be included in the preference lists. This could be a signal of evidence for strategic decisions in the school choice, but, as was previously discussed in section 1.3, this should not matter when choosing schools because the system is strategy-proof. The outcomes of these two covariates can provide aid in the design of information campaigns for the families to adequately apply to schools when participating in the system.



(a) Popularity of schools

(b) Number of declared seats of schools

Figure 5.3: Valuation for popularity and seats in schools in the Coquimbo region

An important variable to be considered is the distance (Abdulkadiroğlu et al. 2017 even uses distance as a baseline for latent utility). This is because schools that are very far away from a student are probably not going to be ranked in that student's list. This negative valuation for the distance is recovered with high significance levels for almost all clusters, and is shown in figure 5.4. This attribute becomes more relevant for schools, as its program designs or any action taken to attract more students is limited by the distance to the students homes, and it is significantly different for each zone in the Coquimbo region. For example, students in the south-center zone fo Elqui are far less willing to travel than in the east zone of Limarí.

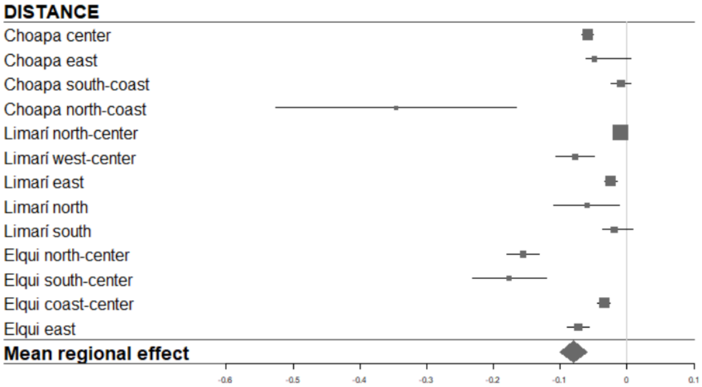
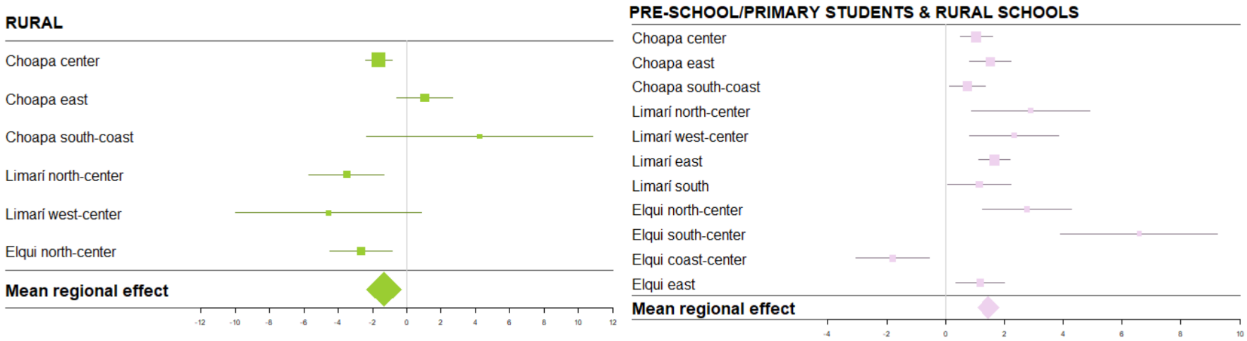


Figure 5.4: Distance between applicants and schools valuation in the Coquimbo region

Naturally, a difference was to be expected between urban and rural schools, considering also that latter type of school is prominent in the region (see figure 3.6). Figure 5.5a shows that the average valuation for these schools is negative in most clusters, and also when looking at the entire region. However, two clusters in Choapa report positive valuation for these schools over urban ones. Contrasting this to the fact that the density of the province is 9.0 [hab/km^2] and the Elqui province—a more urban zone—has a density of 29.4 [hab/km^2][1], it makes sense that in some sectors of this province people prefer rural schools. It should also be noted that statistically significant effects are mostly not present for the Elqui province.

[1] National Institute of Statistics, 2017

Although the mean effect of rural schools for the region is negative, this valuation changes drastically when considering it for pre-school and primary students only. Figure 5.5b shows that students of a younger age in almost all clusters have a positive and statistically significant valuation for rural schools. This again provides information for schools and public policy design to account for which families would prefer a particular type of school, accounting for the different levels of heterogeneity that were found in the populations across clusters. This effect could be interpreted under two lights: parents of younger students are less willing to let their kids travel to a urban zone (note that the distance effect is controlled), or that some rural schools could only be admitting pre-school and primary students.



(a) Rural schools

(b) Rural schools and pre-school and primary students

Figure 5.5: Valuation for rurality of schools in the Coquimbo region

Qualitative variables that could measure the environment of the schools were introduced by the means of the Personal Development Indicators (section 3.1). In figure 5.6 the results for the participation and citizenship formation of the families score for each school are displayed. The recovered effect shows that families often prefer schools that have higher score on this indicator, and the region overall also shows a positive valuation of this variable. The other 3 PDIS showed a less consistent effect. For example, consider the climate of co-existence score of the schools. A school with a better climate score should be preferred than a school with worse score, or at the very least, that no significant effect is to be found. It happens that statistically significant negative effects were found for several clusters for the other 3 PDIS, meaning that the valuation is the opposite of what one would expect. This would indicate that the scores are not capturing correctly what they are supposed to (with the exception of the participation score), so a more profound study of these results could assist on improving the PDIS that the Agency of Education Quality measures annually.

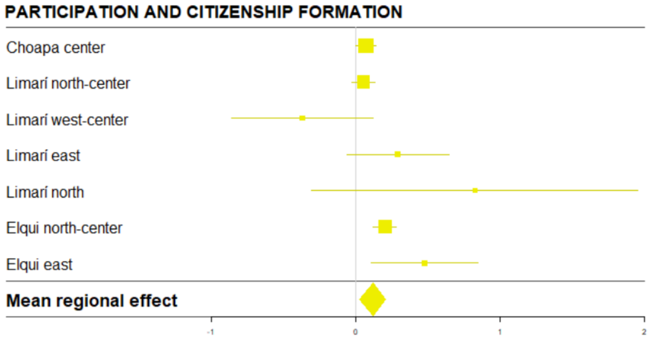
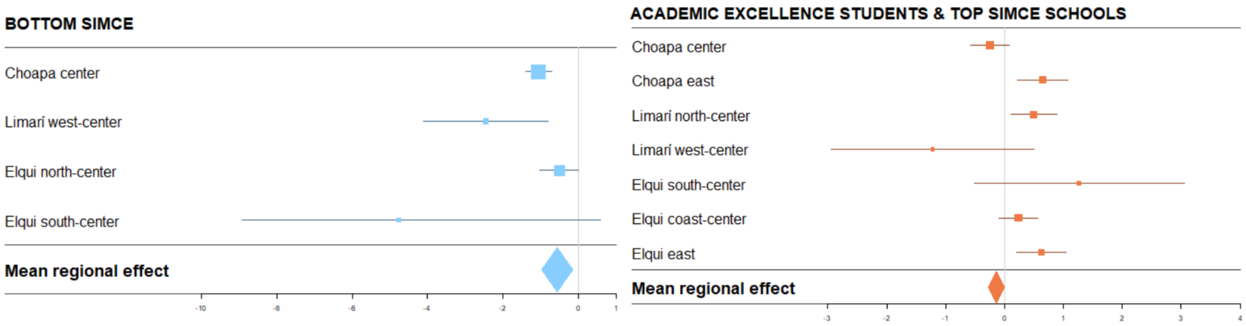


Figure 5.6: Participation and citizenship formation of the families in schools valuation in the Coquimbo region

As a measure of the academic quality of the schools, a non-linear version of the SIMCE test average scores were used, because using the scores directly did not have a consistent significant effect, but it did on the extremes of the distribution. Regarding this effect, it was found that some clusters showed significant negative valuation for the schools in the lowest 25% of the scores distribution, with drastically different levels of intra-cluster heterogeneity, which are shown in figure 5.7a. Also, the mean regional effect was negative. For the top performing schools, a more consistent effect was found for the students of academic excellence. Figure 5.7b shows that in several zones the top performing schools were preferred over the rest for these students. This would mean that in these zones, students of high grades look for schools with better SIMCE scores. Nonetheless, the mean regional effect was negative. Delving deeper into these results could lead to significant impact in the way efforts for improving test scores are planned for schools in different zones of the region. Schools could perform the exercise of estimating how much higher score they would need to increase the demand for seats in the school, or how much demand they could lose by performing worse in next year's SIMCE tests.



(a) Low performing SIMCE schools

(b) High performing SIMCE schools and academic excellence students

Figure 5.7: Valuation for academic quality of schools in the Coquimbo region

It was of interest to study if the socioeconomic group of the families that attended the schools could have an impact in the decision of applying or not to a school. This is because it is hoped that these variables could capture the economic environment of the families of students that assist to the schools, and to help control for the price variable. Specifically, it was found that prioritary students—the most vulnerable socioeconomic group—value negatively, and with high confidence, schools of medium-high socioeconomic level. This indicates that prioritary students induce in self-selection to schools of lower socioeconomic groups. The results are displayed in figure 5.8.

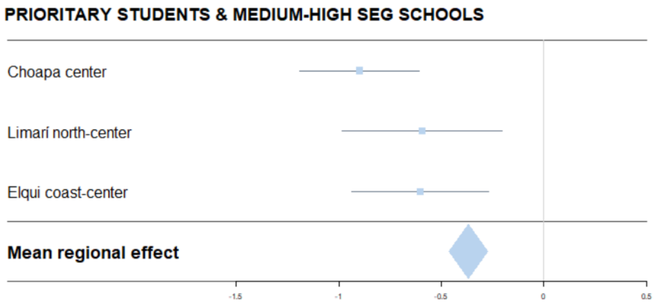
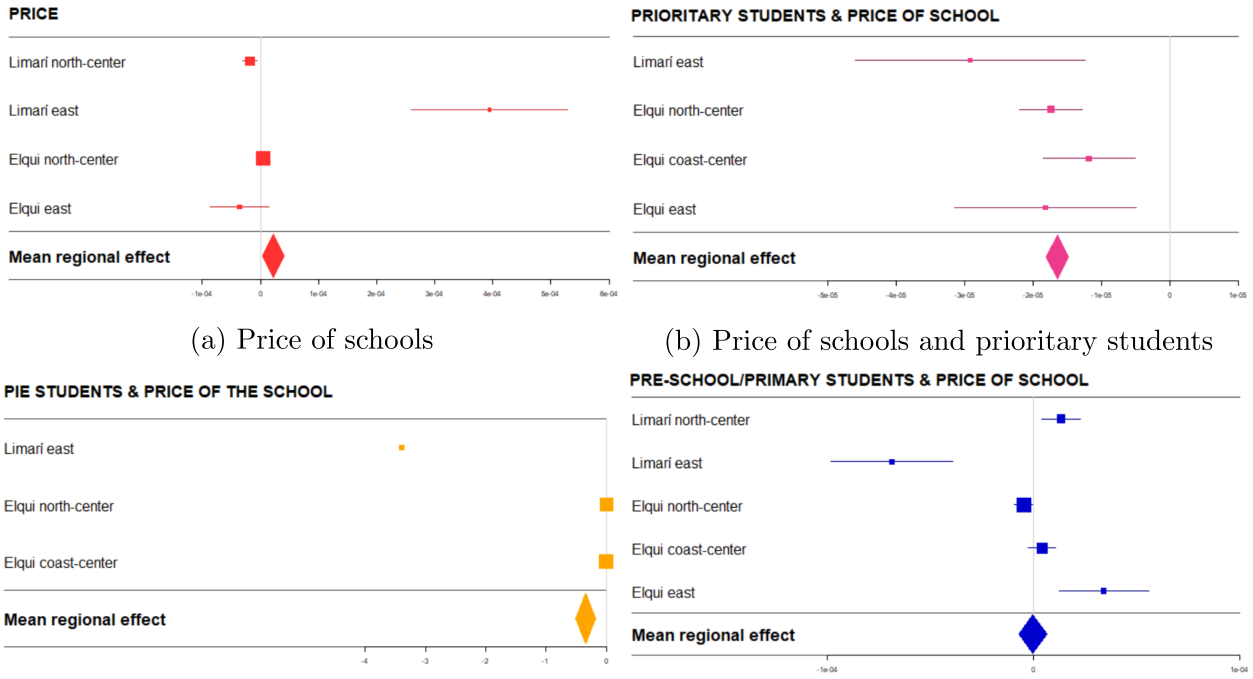


Figure 5.8: Medium-high socioeconomic group of the schools valuation by prioritary students

This leaves the price covariates. It has been studied that failing to account for the price endogeneity problem results in bias in the parameter estimates (Berry, Levinsohn, and Pakes, 1999), since attributes unspecified by the researcher are likely determining the price (i.e. captured in the error term). Thus, a positive valuation for the price, which if modeled incorrectly, would imply that people like to pay more and more, and the suppliers should charge higher prices. A complete picture of how different types of families value the price of schools in the Coquimbo region is displayed in figure 5.9.

Here, this argument is examined with care since it was actually found that the price variable has a mean regional positive effect (see figure 5.9a), although not for all zones. Even at the province level, it was found that in some zones of Elqui and Limarí (in the Choapa province no schools with price participated in the system) a highly confident positive valuation for the price is estimated, but in other zones of the same provinces the inverse effect was found. Moreover, the estimated models control for what it is logical to expect is being captured in the price regarding other school attributes: academic quality, economic environment, other qualitative indicators (PDIS), and distance. Not only that. Figure 5.9b shows highly significant negative effect for the price of the schools in all provinces when considering only prioritary students, and a much more strong mean valuation in the region than for the mean of the population. This indicates that families of low income do not value positively schools with price (with different means and levels of heterogeneity per zone), which can be interpreted as a result of having lower income than the rest. When considering only students with special educational needs (PIE), again only negative significant valuations were obtained—although in fewer zones—, meaning that families of students with disabilities are not keen on schools with prices (figure 5.9c). Also, when studying how the families with younger applicants (pre-school and primary) value the price, less consistent—but still statistically significant—effects are to be found. In some zones, families value positively the price of the school and in others negatively, varying even across provinces (figure 5.9d). These results show that some segments of families, depending on where they live, or their income, or if their young children are in need of a new school, or if their children suffer from some disabilities, can value positively or negatively the price of the school. Considering also all the mentioned variables that can help clean the endogenous price effect, there seems to be strong indication that the captured effect is actually retrieving a willingness to pay, or a statistically significant valuation for being in a segregated environment. Nonetheless, the endogeneity price problem is not rigorously modeled, thus leaving space for future research to provide an approach such as Bayesian BLP (see Yang, Shen, and Allenby, 2003).



(a) Price of schools

(b) Price of schools and prioritary students

(c) Price of schools and PIE students

(d) Price of schools and pre-school and primary students

Figure 5.9: Valuation for price of schools in the Coquimbo region

Some covariates that were left out of the analysis, such as the gender of the school or some levels of SEG of the schools, because of inconsistent or not statistical significant effects were found, although the Markov chains still arrive at the stationary distributions (see annex 8 for the convergence plot of the gender covariate across multiple clusters).

Several potential studies can be addressed from the presented results to help in the design or improvement of public policies, or to aid in the decision making that schools make on limited budgets given the understanding of how families apply to schools, and have been this far only partially studied. In this work, the impact of a policy regarding the price of the schools is simulated to measure the benefits or drawbacks on the social utility of the 2017 assignment in Coquimbo in chapter 6.

Chapter 6

How Does a Pricing Policy Affect Social Welfare?

One of the main promises of the new school admissions system was to end profiting in schools that receive support from the State and to end arbitrary discrimination in the selection processes. This included that these schools must cease to charge prices to families that enroll their children in these establishments. But, is this a *good* policy? It has already been shown that some families seem to prefer a school with price—but not as expensive as a private school. Therefore, does a no-price policy produces more benefits than drawbacks in the social welfare of the assignment? Who receives more benefit and who would end up being more harmed by this policy? These questions are aimed to be answered in this chapter for the 2017 assignment on the Coquimbo region.

6.1 Pricing Policy in the School System Reform

Law 20,845, which provides the regulations for school inclusion of the students by eliminating shared financing of the educational establishments—the subsidized private schools—and prohibits profiting in schools that receive support from the State, establishes that schools that subscribe to this regime must decrease their charges to the parents annually in parallel to receiving increasing subsidies from the State, until no charges are to be made to the parents. These prices are available in the admissions process, therefore the parents should be perfectly aware of how much the school will charge monthly for each children that is enrolled. The schools, aware of this change in the law, must decide either to submit to this regime, convert immediately to a free public school, or convert to a private school. This implies that the schools must evaluate which system is going to be of greater benefit to them, but this could be difficult to estimate if the preference of the families are not known. This is also a concern for the Ministry of Education, because one of its objectives should be to maximize the social welfare under the policies it promotes, and therefore the implementation and nuances of the current law must be dealt with a rigorous empirical analysis to make decisions that satisfy this goal.

In section 6.2, a methodology is developed for estimating the impact of this law if all schools were to stop charging prices under the 2017 Coquimbo assignment.

6.2 Measuring the Impact of a No-Price Policy in Public Schools

The impact of the no-price policy will be now evaluated for the 2017 assignment in the Coquimbo region. This way, the counterfactual simulation consists in evaluating the social welfare of the assignment that already occurred compared to the same assignment but now without the subsidized private schools charging prices, everything else being equal.

It first becomes necessary to define the welfare of the assignment in order to determine the impact a no-price policy would have on the families. The estimated preference models allow for this, since a valuation for the schools by the students was obtained for the entire region. Therefore, the mean individual welfare of a school assignment is defined as

$$W(\mu) = \frac{1}{|I|} \cdot \sum_{i \in I} w_{i,\mu(i)} \tag{6.1}$$

Recalling that $\mu(i)$ is the matching function evaluated for student i, and $w_{i,\mu(i)}$ is the latent utility of student i for school $\mu(i)$, equation 6.1 represents the mean welfare for an individual on I.

The individual utility is not observed, but an expectation for this value can be computed using the posterior means obtained for the covariates of the models. Thus, the expectation of 6.1 is

$$\hat{W}(\mu) = \frac{1}{|I|} \cdot \sum_{i \in I} \mathbb{E}[w_{i,\mu(i)} \mid \hat{\boldsymbol{\beta}}_c, X_{\mu(i)}, \boldsymbol{X}_{i,\mu(i)}] \tag{6.2}$$

Where $\hat{\boldsymbol{\beta}}_c$ is the vector of posterior means obtained for a school in choice set c (in this case, $\mu(i) \in c$), $X_{\mu(i)}$ is the vector of school covariates for school $\mu(i)$, and $\boldsymbol{X}_{i,\mu(i)}$ the vector of interactions between student i and the school, as specified in table 3.2. To measure welfare of different segments of families, for example that of a prioritary student, I is changed to I^{PRI}, which contains only prioritary students. Recall figure 5.9 for the price effect on different segments.

Finally, the expected impact of a no-price policy on individual welfare is

$$\hat{W}^{price}(\mu) - \hat{W}^{no-price}(\mu) = \frac{1}{|I|} \cdot \sum_{i \in I} \left(\mathbb{E}[w_{i,\mu(i)} \mid \hat{\boldsymbol{\beta}}_c, X_{\mu(i)}, \boldsymbol{X}_{i,\mu(i)}] - \mathbb{E}[w_{i,\mu(i)} \mid \hat{\boldsymbol{\beta}}_c, X^*_{\mu(i)}, \boldsymbol{X}^*_{i,\mu(i)}] \right) \tag{6.3}$$

Where $X^*_{\mu(i)}$ and $\boldsymbol{X}^*_{i,\mu(i)}$ have their price covariates set to 0. Equation 6.3 indicates, if greater than 0, that the mean welfare is better under the assignment with prices, and the contrary otherwise.

Using 6.3, the counterfactual simulation is performed for the entire region considering different segments of students, and the results are shown in table 6.1. The second column represents the % improvement on individual welfare for each segment of students, and the third column represents the number of students whom obtained a positive improvement over the number of students that perceived a change in their valuation of their assignment.

Segment	△ % welfare	△ % positive
Prioritary	+3.16%	97.9%
Not prioritary	−0.60%	48.0%
Pre-school and primary	+1.64%	66.5%
Pre-school and primary, not prioritary	−0.23%	51.7%
High-school	+0.66%	61.7%
High-school, not prioritary	−1.34%	35.4%
PIE	+6.83%	100.0%
Academic excellence	+1.40%	61.7%
Everyone	**+1.36%**	**65.3%**

Table 6.1: Counterfactual simulation results of the no-price policy in Coquimbo

The results show that almost every prioritary student would perceive an improvement over their current assignment, while the students that are not prioritary would perceive an overall slight negative effect, although only for half of them. This further consolidates the preliminary findings: prioritary students are hurt by the school's prices, while some students appreciate being in a segregated environment and thus their valuation for their assignment would be harmed if this would no longer be a discriminative—by price—environment. This effect holds for pre-school, primary and high-school, but it specially devalues the preference for subsidized private schools for high-school students that are not prioritary, resulting in a negative effect for almost 65% of them. On the other hand, PIE students see the greatest benefit from this policy, estimating a highly positive impact for everyone of them. It also results positive for academic excellence students, providing a positive benefit for more than 60% of them (although their interaction with price was not controlled). Overall, the regional effect of a no-price policy would be positive for the 2017 school assignment, benefiting more than 65% of the students of the region that participated in the system.

Further analysis of what these results imply for the school admissions system are discussed in Conclusions.

Conclusions

Matching mechanisms for school assignment can help improve the way families try to enroll their children into the best possible school they see fit. Abdulkadiroğlu et al. 2017 showed that in the New York assignment improvement in social welfare was obtained by changing from an uncoordinated assignment to a coordinated one. In the case of Chile, the mechanism used before was uncoordinated and also not centralized, producing long queues that could last for entire nights for parents to ensure a seat for their children, heavily inducing strategizing by the parents. In this work, it has been shown how to successfully implement and design the matching mechanism to not only satisfy the desirable properties of the Deferred Acceptance algorithm that has made it widely popular—like eliminating all incentives for strategic behaviour—, but also to account for all the regulations imposed by the educational reform, making the Chilean case unique among its pairs.

Further analysis into the data that was produced from the school assignment can help to obtain insight into how families respond to the system and to understand how they allocate their preferences according to observable characteristics from both the schools and the families. For example, in the New York system it is mandatory to submit at least 6 schools, while in Chile the minimum amount is 2, and 1 in rural zones. In the 2004 assignment in NYC almost 70,000 applicants participated, and in the 2018 assignment in Chile almost 77,000. Nonetheless, for the first one 542,666 applications were received, and only 276,112 in the latter. Is this evidence of poor understanding of the system in Chile? This certainly prejudices any models that try to infer families preferences as less information is provided to the system, while also probably leaving more students unassigned. But the systems are not directly comparable. NYC is a highly dense urban city, while in Chile several rural zones participated, and therefore is not clear how a new rule for the system would impact the welfare of the assignment. Should the minimum amount of applications be raised? Or only in urban zones? Or should it be raised differently in urban and rural zones? To address questions of this nature, models that can help understand and quantify the effects of any policy must be developed.

The preference data submitted to the system is more complex than what is usually found when modeling discrete choices. A multivariate ordered probit was developed to account for the fact that the preference data is expressed only for some schools by each family (the multivariate part) and it is a strict ranking (the ordered part). This was compared to the classic method for modeling discrete choice, the multivariate probit—both implemented in their Bayesian versions—, where it was found that some improvements can be achieved by explicitly modeling the rankings, and it is to be expected that greater benefits can be obtained by an ordered structural model if more rich rankings data is provided to the system. Moreover, the scale of the problem indulged in the need for clustering schools to find choice sets adequate for groups of families, since it is not logical to make families choose schools that are hundred or even thousands of kilometers away, adding an extra difficulty to the problem while also making it substantially different than the NYC problem. Replicating Abdulkadiroğlu et al. 2017 methodology applied for the system in NYC would make the Chilean version of the problem intractable.

These models allowed to find strong indication of strategizing by the families, which seemed to be influenced by the number of available seats of the schools—even though the system is strategy-proof—, that families value positively the participation and citizenship formation of the school and negatively schools in the bottom quartile of average SIMCE scores, but not such a strong indication for schools in the top quartile. There was also indication of self-selection according to the socioeconomic group of the families that attended the schools, as students of lower resources strongly disliked schools of the medium-high group, even after controlling by the price of the school. The price of the subsidized private schools also had an heterogenous effect depending not only on the family's economical situation, but also on their children's age, if their children had a disability, and also on where they live across the Coquimbo region. A positive mean effect was found for the price in the Coquimbo region, but negative for prioritary and PIE students. For families with children applying to pre-school or primary, positive and negative significant effects were found, implying that some families value positively to let their children be in a segregated environment, while hurting other families that are not keen on paying, probably because of their economic situation.

In finding all these primitives, several counterfactual analyses could be carried out to measure the impact of different policies on social welfare, and in this work a no-price policy for all schools in the Coquimbo was explored, as it is a concern for the coming years since this policy is part of the new educational reform to prevent segregation in schools that receive support from the State. It was found that this policy would provide an improvement of 1.36% in welfare in the current school assignment, benefiting more than 65% of the students and families. The main benefits would be obtained for PIE students, with a 6.83% of improvement, and prioritary students with a 3.16%. Nonetheless, this would impact negatively non-prioritary students, specially those in high school, devaluing their welfare assignment in 1.34%. This suggests that this policy would benefit the majority of the students and the improvement for the mean of the students would also be positive, so the analysis suggests that it is a policy that should be carried out, although it might be worth considering a hybrid policy, for example where at least PIE students do not pay for schools accounting for the extra costs families should carry for their children's disability. A more rigorous analysis could be done by simulating the complete assignment by considering the changes in preferences, and therefore in submitted rankings, although this would require more information at the individual level in order to be precise, and this information was not available for this work.

The main message that it is tried to be conveyed by this work is that public policies and aid in decision making of the different agents in the system should be accompanied by empirical analysis to measure the impact of them. The proposed methodology shows that this is not a simple task, as this must be dealt with great care from the underlying assumptions about the agents' behaviour to the manipulation of data, to then show if the models can generalize and reliable conclusions and insights can be obtained from them. In this manner, the best way to apply changes—and even if they should be made—can be studied, remarking that this should be done especially since any modifications in the system have the potential to impact positively millions of families.

Bibliography

[1] D. Gale and L.s. Shapley. “College Admissions and the Stability Of Marriage”. In: (1962). DOI: 10.21236/ad0251958.

[2] Lloyd Shapley and Herbert Scarf. “On Cores and Indivisibility”. In: *Journal of Mathematical Economics* 1.1 (1974), pp. 23–37. DOI: 10.1016/0304-4068(74)90033-0.

[3] L. E. Dubins and D. A. Freedman. “Machiavelli and the Gale-Shapley Algorithm”. In: *The American Mathematical Monthly* 88.7 (1981), p. 485. DOI: 10.2307/2321753.

[4] Alvin E. Roth. “Incentive Compatibility in a Market With Indivisible Goods”. In: *Economics Letters* 9.2 (1982), pp. 127–132. DOI: 10.1016/0165-1765(82)90003-9.

[5] Steven Berry, James Levinsohn, and Ariel Pakes. “Voluntary Export Restraints on Automobiles: Evaluating a Strategic TradePolicy”. In: *American Economic Review* (1999), pp. 400–430. DOI: 10.3386/w5235.

[6] Atila Abdulkadiroğlu and Tayfun Sönmez. “School Choice: A Mechanism Design Approach”. In: *American Economic Review* 93.3 (2003), pp. 729–747. DOI: 10.1257/000282803322157061.

[7] Peter E. Rossi, Greg M. Allenby, and Robert McCulloch. *Bayesian Statistics and Marketing*. Wiley, 2003.

[8] Sha Yang, Yuxin Chen, and Greg M. Allenby. “Bayesian Analysis of Simultaneous Demand and Supply”. In: *Quantitative Marketing and Economics* 1.3 (2003), pp. 251–275. DOI: 10.1023/b:qmec.0000003327.55605.26.

[9] Onur Kesten. “Student Placement to Public Schools in the US: Two New Solutions”. In: (2004).

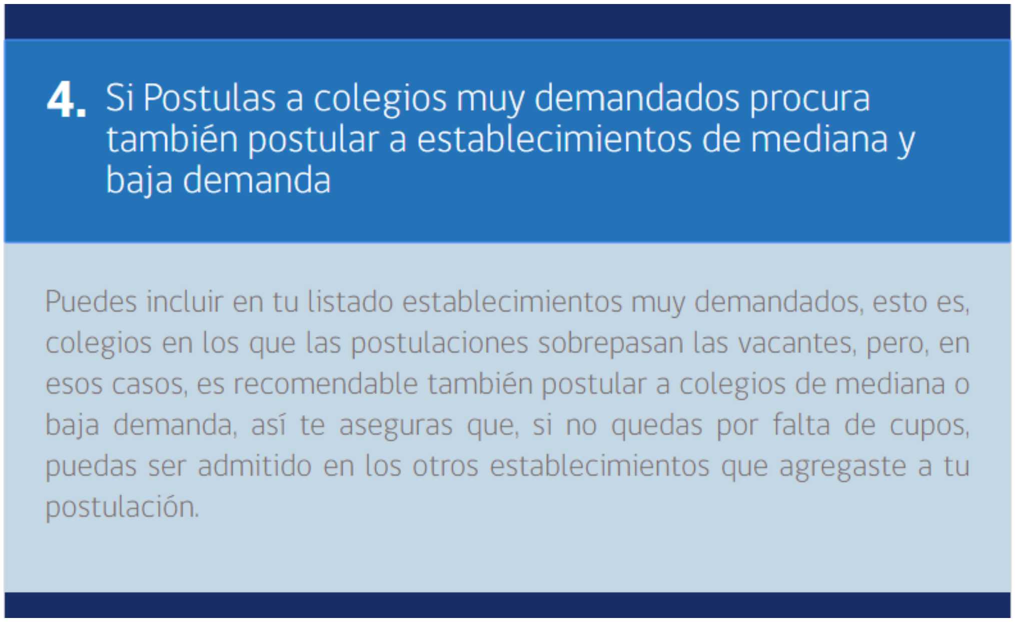
[10] Atila Abdulkadiroğlu, Parag Pathak, and Alvin Roth. “Strategy-proofness Versus Efficiency in Matching with Indifferences: Redesigning the New York City High School Match”. In: (2009). DOI: 10.3386/w14864.

[11] Andrew Gelman et al. *Bayesian Data Analysis*. CRC Press, 2009.

[12] Atila Abdulkadiroğlu, Nikhil Agarwal, and Parag Pathak. “The Welfare Effects of Coordinated Assignment: Evidence from the NYC High School Match”. In: (2017). DOI: 10.3386/w21046.

Annex

1. Recommendations from the school's admission system website



4. Si Postulas a colegios muy demandados procura también postular a establecimientos de mediana y baja demanda

Puedes incluir en tu listado establecimientos muy demandados, esto es, colegios en los que las postulaciones sobrepasan las vacantes, pero, en esos casos, es recomendable también postular a colegios de mediana o baja demanda, así te aseguras que, si no quedas por falta de cupos, puedas ser admitido en los otros establecimientos que agregaste a tu postulación.

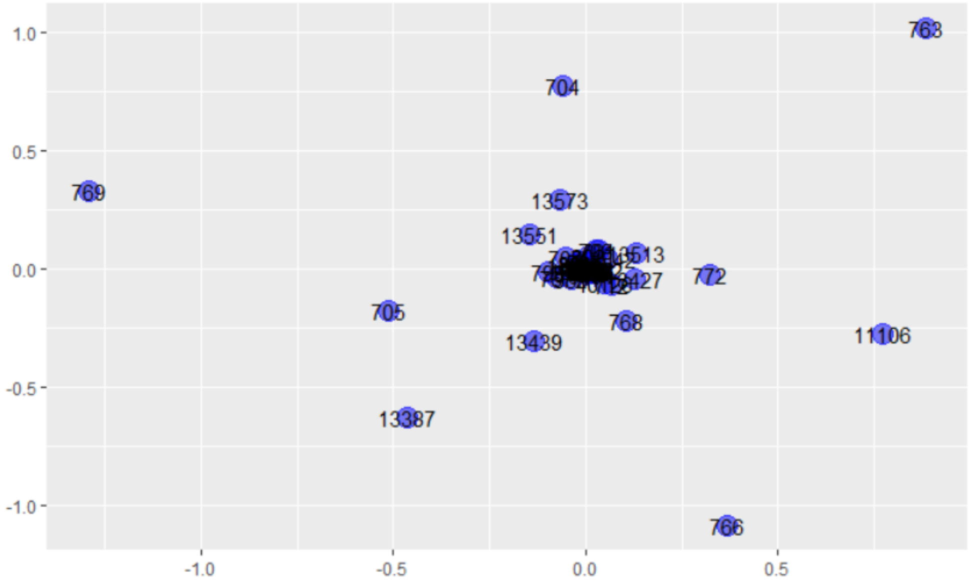
2. C++ implementation of the Gibbs sampler for the latent utilities by student

```
vec drawwi_mvop(vec const& w, vec const& mu, mat const& sigmai, int p,
ivec y){
  //function to draw w_i as in an ordered multivariate probit fashion

  int ny = y.size();

  vec outwi = w;

  for(int i = 0; i < ny; i++){

          if(i == 0 && y[i] != 0 && i+1 <ny && y[i+1] != 0){
                // if it's the first observed response, sample from a
                // truncated normal from below by the utility of the next
                // response (and it's not the last one, and the following
                //  is a ranked response)
                vec Cmout = condmom(outwi, mu, sigmai, p, i+1);
                outwi[i] = trunNorm(Cmout[0], Cmout[1], outwi[i+1], 0);
          }else if(i == 0 &&  y[i] != 0){
                // if it's the first observed response, sample from a
                // positive truncated normal (and there isn't a next
                // ranked response)
                vec Cmout = condmom(outwi, mu, sigmai, p, i+1);
                outwi[i] = trunNorm(Cmout[0], Cmout[1], 0, 0);

          }else if(y[i] != 0 && i+1 < ny && y[i+1] != 0){
                // if it's an observed response that it's not the first
                // one, and the following response it's a ranked
                // response, sample from a double-sided truncated normal
                vec Cmout = condmom(outwi, mu, sigmai, p, i+1);
                outwi[i] = rtrunSc(Cmout[0], Cmout[1], outwi[i+1],
                outwi[i-1]);

          }else if(y[i] != 0){
                // if it's an observed response that it's not the first
                // one and there isn't a next ranked response, sample
                // from a double-sided truncated normal truncated below
                // by 0
                vec Cmout = condmom(outwi, mu, sigmai, p, i+1);
                outwi[i] = rtrunSc(Cmout[0], Cmout[1], 0, outwi[i-1]);


          }else{
                // if it's not a ranked response sample from a truncated
                // normal, truncated above by 0
                vec Cmout = condmom(outwi, mu, sigmai, p, i+1);
                outwi[i] = trunNorm(Cmout[0], Cmout[1], 0, 1);
          }
  }
        return (outwi);
 }
}
```

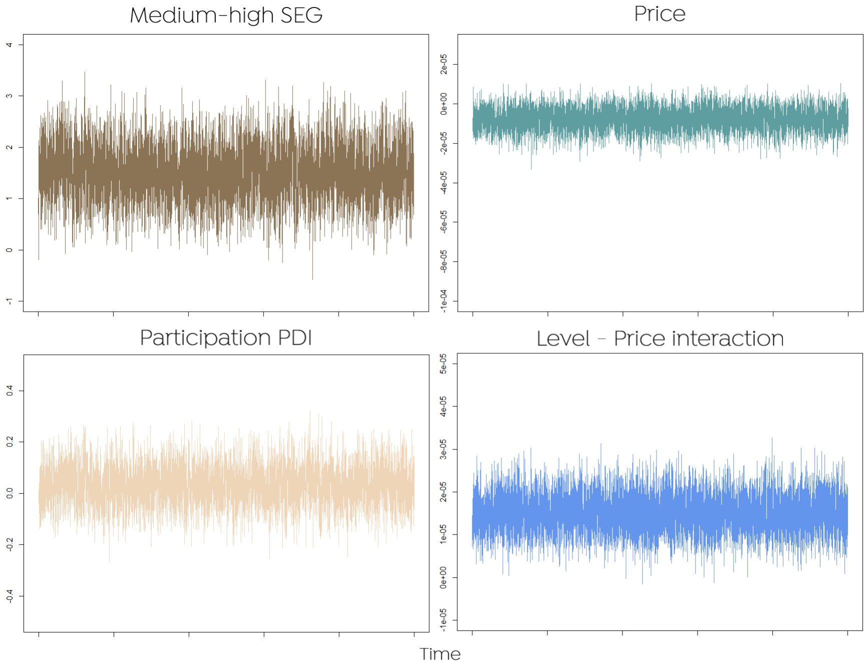
3. C++ implementation of the Gibbs sampler for the latent utilities for all students

```
vec draww_mvop(vec const& w, vec const& mu, mat const& sigmai,
ivec const& y){
  //function to draw all w vector for all n obs

  int p = sigmai.n_cols;
  int n = w.size()/p;
  int ind;
  vec outw = zeros<vec>(w.size());

 for(int i = 0; i < n; i++){

    ind = p*i;

    outw.subvec(ind,ind+p-1) =
    drawwi_mvop(w.subvec(ind,ind+p-1),mu.subvec(ind,ind+p-1),sigmai,p,y);
  }
```

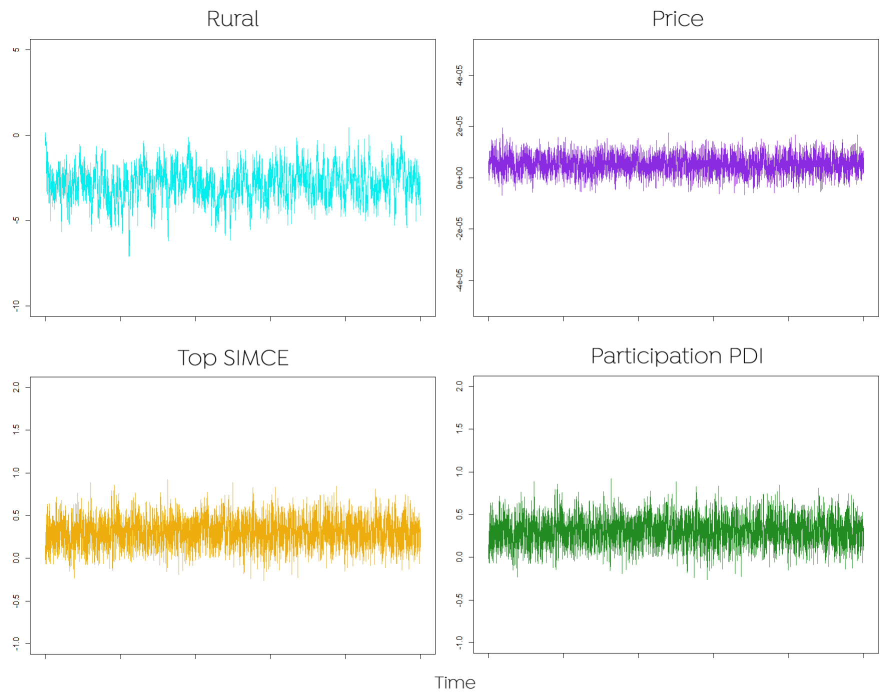
4. Multi-dimensional scaling plot of the co-occurrence matrix for the Limarí province schools identified by their RBD code



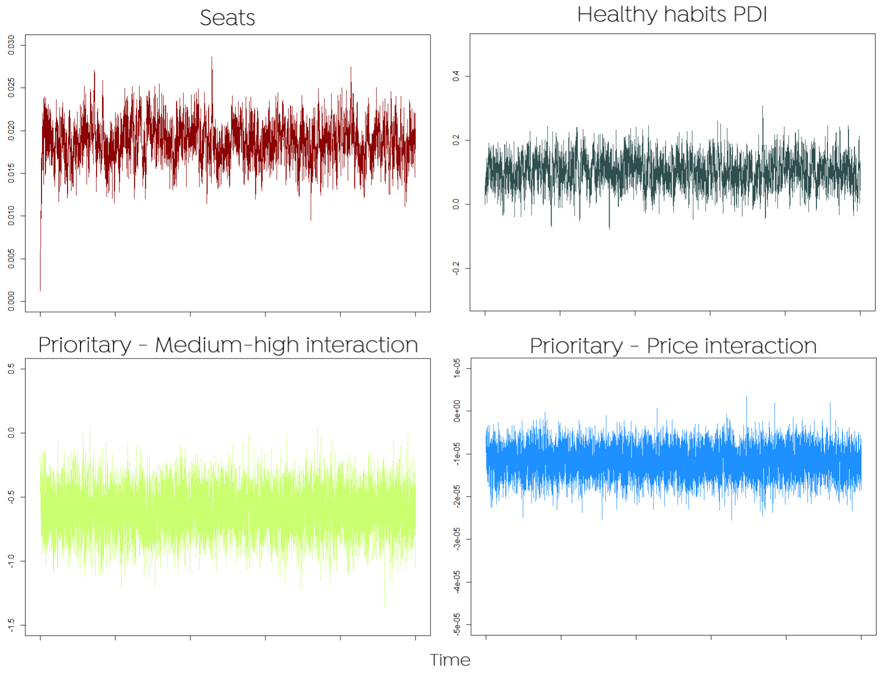
5. Convergence plot of the MVOP for the Limarí province, north-center cluster



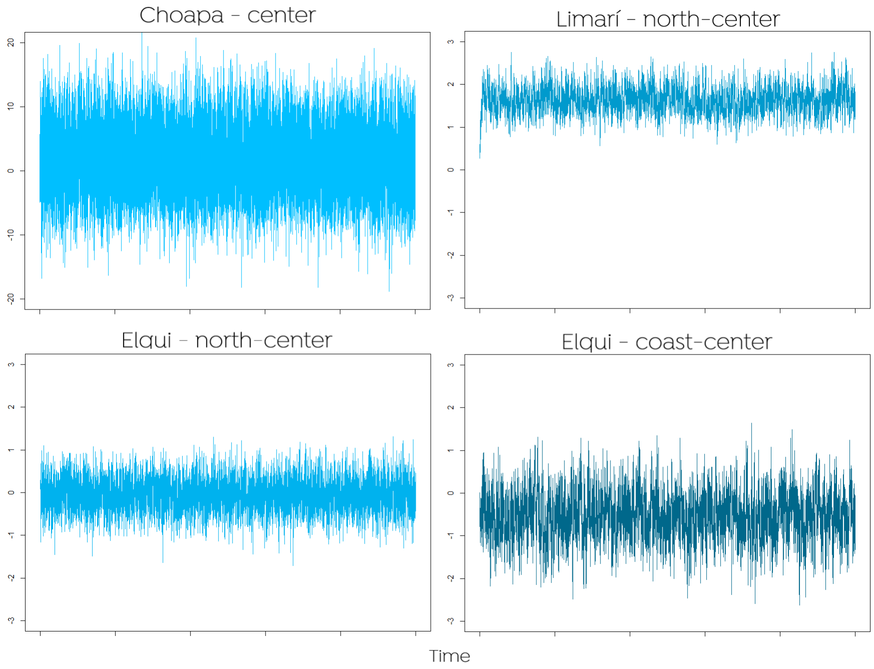
6. Convergence plot of the MVOP for the Elqui province, north-center cluster



7. Convergence plots of the MVOP for the Elqui province, coast-center cluster



8. Convergence plots of the MVOP for the gender covariate for multiple clusters


Choapa - center
Limarí - north-center
Elqui - north-center
Elqui - coast-center
Time